Universitat de Barcelona

Apunts

Tercer Semestre

Mètodes Numèrics I (MNI)

Autor:
Mario Vilar

Professor:
Dr. Joan Carles Tatjer

10 de desembre de 2021

Numerical Methods (NM) is a part of Mathematics mostly devoted to obtain (approximate) solutions of equations in a practical way. One of the concerns of NM is to obtain solutions efficiently, i.e., by using algorithms (procedures to compute the solution) which are as fast as possible. Another important concern is to estimate the errors, i.e., how much the numerical solution differs from the exact solution of the equation.

Universitat de Barcelona

Índex

Taula de continguts

Capítol 1

Capítol 2

I

Preliminars

Farem un petit repàs de les nocions de càlcul apreses a *Introducció al Càlcul Diferencial*.

1.1 Límits i continuïtat

Sigui $D \subseteq \mathbb{R}$ i $f : D \longrightarrow \mathbb{R}$. Donat $a \in D$ direm que el límit de f quan x tendeix a a és l si a mesura que x es va apropant a a el valor de $f(x)$ es va apropant a l.

Definició 1.1.1 (Límit). El límit de f quan x tendeix a a és $l \in \mathbb{R}$, i ho escriurem $\lim_{x\to a} f(x) = l$, si per tot $\varepsilon > 0$, $\exists\ \delta > 0$ tal que

$$\forall x \in D,\ 0 < |x - a| < \delta \implies |f(x) - l| < \varepsilon. \tag{1.1.1}$$

Definició 1.1.2 (Límit, alternativa). Per tot entorn de l existeix un entorn de $x = a$ si x viu a aquest entorn de a, sigui $x \neq a$, llavors $f(x)$ viu a l'entorn de l fixat al principi.

Definició 1.1.3 (Límit per la dreta). Direm que el límit de f quan x tendeix a a per la dreta és $l \in \mathbb{R}$ i ho escriurem $\lim_{x\to a^+} f(x) = l$, si $\forall \varepsilon > 0$, $\exists \delta > 0$ tal que

$$x \in D \cap (a, a + \delta) \implies f(x) \in (l - \varepsilon, l + \varepsilon) \implies |f(x) - l| < \varepsilon \tag{1.1.2}$$

De la mateixa manera, ho podríem definir per l'esquerra.

Definició 1.1.4 (Continuïtat). Sigui $f : D \longrightarrow \mathbb{R}$ i $a \in D$. Direm que f és contínua en el punt $a \in D$ si $\exists \lim_{x\in a} f(x)$ i aquest val $f(a)$. De manera equivalent, f és contínua al punt a si $\forall \varepsilon > 0,\ \exists \delta > 0$ tal que

$$x \in D \cap I(a, \delta) \implies f(x) \in I(f(a), \varepsilon) \tag{1.1.3}$$

Quan f diem que és contínua a D entendrem que és contínua a tots els $a \in D$.

Teorema 1.1.5 (Teorema de Weierstrass). *Sigui $f : [a, b] \longrightarrow \mathbb{R}$ contínua. Aleshores f és acotada i assoleix un màxim i un mínim absolut.*

Teorema 1.1.6 (Teorema de Bolzano). *Sigui $f : [a, b] \longrightarrow \mathbb{R}$ contínua tal que $f(a)f(b) < 0$. Aleshores, existeix $c \in (a, b)$ tal que $f(c) = 0$.*

Teorema 1.1.7 (Teorema del valor mig). *Sigui $f : [a, b] \longrightarrow \mathbb{R}$ contínua i sigui z un valor entre $f(a)$ i $f(b)$. Aleshores existeix $c \in [a, b]$ tal que $f(c) = z$.*

1.2 Derivades

Definició 1.2.1 (Derivada d'una funció en un punt)**.** Sigui f definida a un interval I i sigui $a \in I$. Direm que f és derivable al punt a si existeix el límit

$$\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h}. \tag{1.2.1}$$

Observem que

$$\lim_{x\to a} \frac{f(x)-f(a)-f'(a)(x-a)}{x-a} = 0. \tag{1.2.2}$$

Teorema 1.2.2 (Teorema de Rolle)**.** *Sigui $f \in \mathscr{C}([a,b])$, derivable a tot (a,b) tal que $f(a) = f(b)$. Aleshores existeix $c \in (a,b)$ tal que $f'(c) = 0$. Equivalentment,*

$$f \in \mathscr{C}([a,b]) \text{ derivable en } (a,b) \;\big|\; f(a) = f(b) \implies \exists c \in (a,b) \;\big|\; f'(c) = 0. \tag{1.2.3}$$

Teorema 1.2.3 (Teorema del valor mitjà de Lagrange)**.** *Sigui $f \in \mathscr{C}([a,b])$ i derivable a tot (a,b). Aleshores, $\exists c \in (a,b)$ tal que*

$$f(b) - f(a) = f'(c)(b-a). \tag{1.2.4}$$

Teorema 1.2.4. *Sigui f una funció derivable al punt $a \in I$. Aleshores, f és contínua al punt $a \in I$. En altres paraules, f derivable $\implies$ f contínua.*

1.3 Taylor

Teorema 1.3.1 (Polinomi de Taylor)**.** *Sigui f una funció $n-1$ cops derivable a un interval I, i sigui $a \in I$ on f^{n-1} és derivable. Aleshores, el polinomi*

$$P_n(x) = \sum_{j=0}^{n} \frac{f^{(j)}(a)}{j!}(x-a)^j = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n \tag{1.3.1}$$

*té ordre de contacte superior a n amb f al punt a. A $p_n(x)$ se l'anomena **polinomi de Taylor de grau n de f al punt a**, i depèn d'f, del grau de n i del punt a.*

Proposició 1.3.2. *El polinomi de Taylor $p_n(x)$ és l'únic polinomi de grau menor o igual a n que té ordre de contacte superior a n amb f en el punt a.*

Definició 1.3.3 (Fórmula de Taylor)**.** Sigui f derivable $n+1$ cops a un interval I. Per a $a, x \in I$ es té

$$f(x) = p_n(x) + R_n(x) = \sum_{j=0}^{n} \frac{f^{(j)}(a)}{j!}(x-a)^j + R_n(x), \tag{1.3.2}$$

i existeix c entre x i a tal que

$$R_n(x) = \frac{f^{n+1}(c)}{(n+1)!}(x-a)^{n+1}. \tag{1.3.3}$$

Aquesta darrera expressió $R_n(x)$ s'anomena **resta de Lagrange**.

II

Problemes numèrics i errors

2.1

Conceptes bàsics

Definició 2.1.1 (Problema numèric)**.** Un problema numèric és una descripció clara i no ambigua de la relació entre les dades d'entrada i les dades de sortida. Les dades tant d'entrada com de sortida són discretes, és a dir, consisteixen d'un nombre finit de registres.

Definició 2.1.2 (Algorisme)**.** Un algorisme per a un problema numèric és una descripció completa d'un nombre finit d'operacions ben definides, mitjançant el qual tot vector d'entrada és transformat a un de sortida. En efecte, les operacions són aritmètiques o lògiques.

Notació 2.1.3.

- $a \ll b$: a és molt més petit que b;
- $a \simeq b$ vol dir que a és aproximadament igual a b, equivalentment, $|a - b| \leq c$, on c el més petit possible.

2.1.1 Fonts d'error

Hi pot haver diversos tipus d'errors:

1. en les dades d'entrada, deguts a mesuraments incorrectes o a la finitud de la seva representació a l'ordinador;
2. errors d'arrodoniment en els càlculs;
3. la major part de mètodes que veurem durant el curs no produeixen la solució exacta del problema que aborden i es coneixen com a errors de truncament;
4. error de discretització, on passem d'un conjunt de dades infinit a un de finit (discret);
5. per la simplificació del model matemàtic i del model numèric o per errada humana.

2.1.2 Absolut i relatiu

Sigui $x \in \mathbb{R}$ el valor exacte d'una certa magnitud i $\overline{x} \in \mathbb{R}$ un valor aproximat d'x.

Definició 2.1.4 (Error absolut)**.** $\Delta a = \overline{a} - a$. Ho denotarem per $e_a(\overline{a})$.

Definició 2.1.5 (Error relatiu)**.** $\frac{\Delta a}{a}$, sempre que $a \neq 0$. Ho escriurem com $e_r(\overline{a})$.

Observació 2.1.6.

- $a \approx \overline{a} \implies e_r(\overline{a}) \approx \frac{\Delta a}{\overline{a}}$;
- Si $a \in \mathbb{R}^n$: $e_a(\overline{a}) = ||\overline{a} - a||$ i $e_r(\overline{a}) = \frac{||\overline{a}-a||}{||a||}$.

Definició 2.1.7 (Fita d'error)**.** Una fita d'error (e), relatiu (e_r) o absolut (e_a), és $\varepsilon(\overline{a})$ si, i només si, $|e(\overline{a})| \leq \varepsilon(\overline{a})$.

Fent un abús de notació, podem considerar $x = \overline{x} \pm \varepsilon \therefore \overline{x} - \varepsilon \leq x \leq \overline{x} + \varepsilon,\ \varepsilon > 0$. Posem $x = \overline{x}(1 \pm \varepsilon)$. Ens queda, operant, que $\varepsilon \geq \frac{|x-\overline{x}|}{|x|}$.

Exemple 2.1.8. Posem $0,1$ en base 10. Si ho passem a base binària, ens queda $0,000110011...$ i podem decidir on tallar. Si t és el nombre de xifres decimals que utilitzem per arrodonir/truncar. D'aquesta manera:

- $|\overline{x} - x| \leq 10^{-t}$ tallant;
- $|\overline{x} - x| \leq \frac{1}{2}10^{-t}$ arrodonint.

Suposem que volem arrodonir un nombre a t dígits. Sigui η la part del nombre que correspon a les posicions de la dreta del t-èsim decimal. Ja és ben sabut des dels nivells més elementals que si $\eta < 0,5 \cdot 10^{-t}$, aleshores el t-èsim decimal no varia; si $\eta > 0,5 \cdot 10^{-t}$, s'incrementa el valor del t-èsim decimal en una unitat. En el cas que $\eta = 0,5 \cdot 10^{-t}$, el t-èsim decimal augmentarà en 1 si, i només si, és senar. En canvi, si volguéssim truncar, simplement ignorem tots els decimals posteriors a la t-èsima posició.

En definitiva, quan un nombre s'aproxima per arrodoniment o truncament, l'error intrínsec es propaga i aquest fet s'ha de tenir en compte.

Definició 2.1.9 (Decimals correctes)**.** Una aproximació $\overline{x}$ de x té t decimals correctes si, i només si, $|\overline{x} - x| \leq \frac{1}{2}10^{-t}$.

Definició 2.1.10 (Dígits significatius)**.** En un valor aproximat $\overline{x}$ amb $t > 0$ decimals correctes, els digíts amb posicions amb unitat més gran o igual a 10^{-t} s'anomenen dígits significatius. Els zeros no es tenen en compte.

Exemple 2.1.11. Agafant posicions anteriors a t ignorant els zeros a l'esquerra. Això és, per exemple:

- $0.001234 \pm 0.5 \cdot 10^{-5}$ ens dona 5 decimals correctes i 3 dígits significatius.
- $56.895 \pm \frac{1}{2}10^{-3}$ ens dona 3 decimals correctes i 5 dígits significatius.

Exemple 2.1.12. $x_1 = 10,123456 \cdot 10^{-6}$ i $x_2 = 10,123788 \cdot 10^{-6}$ amb un error relatiu de $0,5 \cdot 10^{-7}$. L'error absolut és $x_1 - x_2 = -0,000332 \cdot 10^{-6}$.

Exemple 2.1.13. $x^2 - 18x + 1 = 0 \iff x_1 = 9 + \sqrt{80} = (9 + 8,9443) \pm 0,5 \cdot 10^{-4}, x_2 = 9 - \sqrt{80} = 0,0557 \pm 0,5 \cdot 10^{-4}$. Com veiem, perdem dígits. Fem un procés semblant a la racionalització però a la inversa:

$$(9 - \sqrt{80})\frac{9 + \sqrt{80}}{9 + \sqrt{80}} = \frac{1}{17,9443 \pm 0,5 \cdot 10^{-4}} = 0,00557280. \tag{2.1.1}$$

Notació 2.1.14 (Notació $\mathcal{O}$ gran)**.** Direm que $g(h) = \mathcal{O}(h^p)$ tal que $h \to 0$ quan existeixin nombres p, K i d tal que

$$|g(h)| \leq K \cdot h^p,\ \forall |h| \leq d. \tag{2.1.2}$$

Aquest concepte també serveix per esbrinar la càrrega de treball que suposa resoldre un cert problema numèric utilitzant un algorisme concret.

2.2

Propagació de l'error

Quan usem valors aproximats en processos computacionals el seu error, evidentment, al seu torn, augmentarà l'error del resultat final. De fet, la idea bàsica de l'anàlisi de l'error cap al darrera consisteix a estudiar les modificacions que hauríem de fer sobre les dades d'entrada suposant que no hi haguessin errors en les operacions s'obtingués el mateix error en el resultat. Així doncs, ens cal desenvolupar alguns mètodes simples per estimar com els errors en les dades es propaguen.

Imaginem que volem computar $f(x)$, f derivable. Assumim que coneixem una fita d'error tal que $x = \overline{x} \pm \varepsilon$. Si f és monòtona (creixent o decreixent, indistintament) podem estimar l'error propagat simplement calculant $f(\overline{x} - \varepsilon)$ i $f(\overline{x} + \varepsilon)$:

$$|R_X| = |\Delta f| = |f(\overline{x}) - f(x)| \leq \max\{|f(\overline{x} - \varepsilon) - f(\overline{x})|, |f(\overline{x} + \varepsilon) - f(\overline{x})|\}. \tag{2.2.1}$$

Teorema 2.2.1 (Teorema del valor mitjà)**.** *Si la funció f és derivable, existeix un punt $\xi \mid \xi \in [x, \overline{x}]$ tal que*

$$\Delta f = f(\overline{x}) - f(x) = f'(\xi)(\overline{x} - x). \tag{2.2.2}$$

Observació 2.2.2. Quan aquest teorema s'utilitza per a una estimació d'error pràctica, la derivada es computa a $\overline{x}$ i l'entorn d'error s'ajusta afegint una correcció adequada. Això té un cert sentit geomètric.

Proposició 2.2.3. *Si $|f'(y)| \leq M, \forall y \in \langle x, \overline{x} \rangle$, aleshores $|f(\overline{x}) - f(x)| \leq M|x - \overline{x}|$. Així doncs, $f(\overline{x}) - f(x) \approx f'(\overline{x})(\overline{x} - x)$ i*

$$|e_a(f(\overline{x}))| \lesssim |f'(\overline{x})||e_a(\overline{x})|. \tag{2.2.3}$$

Exemple 2.2.4. Sigui $f(x) = \sqrt{x}, x = 2.05 \pm 0.01$ i $\overline{x} = 2.05$, així que

$$|e_a(f(\overline{x}))| \lesssim \frac{1}{2\sqrt{x}}|e_a(\overline{x})| \leq 0.0035. \tag{2.2.4}$$

Podem, encara, abstraure més aquest teorema per tal de fer-lo útil per a l'examinació de la propagació d'error en l'avaluació d'una funció f en n variables:

Teorema 2.2.5. *Si la funció de variable real f és derivable en un entorn d'$x = (x_1, \ldots, x_n)$ i $\overline{x} = x + \Delta x$ és un punt que hi viu, aleshores existeix un nombre $\theta, 0 < \theta < 1$, tal que*

$$\Delta f = f(\overline{x}) - f(x) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x + \theta \Delta x)\Delta x_k. \tag{2.2.5}$$

Demostració. Definim la funció $F(t) = f(x + t\Delta x)$. 2.2.1 i la regla de la cadena ens dona:

$$\Delta f = F(1) - F(0) = F'(\theta) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x + \theta \Delta x_k)\Delta x_k. \tag{2.2.6}$$

■

Quan volem una estimació pràctica de l'error, les derivades parcials s'avaluen a $x = \overline{x}$. Quan sols hi ha fites d'error a $\overline{x}$, la fita per Δf es determina amb desigualtat triangular. En definitiva:

$$\Delta f = f(\overline{x}) - f(x) \approx \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\overline{x}) \Delta x_k \tag{2.2.7}$$

$$|\Delta f| \lesssim \sum_{k=1}^{n} \left| \frac{\partial f}{\partial x_k}(\overline{x}) \right| \Delta x_k \tag{2.2.8}$$

Figura 2.1: Fórmula general de la propagació de l'error (2.2.7) i la fita d'error màxima (2.2.8).

Exemple 2.2.6. Sigui $y = \sin(x_1^2 x_2)$, on $x_1 = 0.75 \pm 10^{-2}$ i $x_2 = 0.413 \pm 3 \cdot 10^{-3}$. La cota d'error màxima dona:

$$|\Delta y| \lesssim \cos(x_1^2 x_2) \cdot 2x_1 x_2 \Delta x_1| + |\cos(x_1^2 x_2) \cdot x_1^2 \Delta x_2| \leq 0.0077. \tag{2.2.9}$$

El valor aproximat és

$$\overline{y} = \sin(0.75^2 \cdot 0.413) = 0.230229 \pm 0,5 \cdot 10^{-6}. \tag{2.2.10}$$

Evidentment, la fita d'error màxima és força imprecisa quan es té un alt nombre de variables.

2.2.1 Comparació d'errors

2.2.1.1 Suma

Per a l'error absolut, podem determinar els valors o un entorn d'aquestes operacions aritmètiques: $e_a(x+y), e_a(x-y), |e_a(x+y)|, |e_a(x-y)|$.

$$\begin{aligned} e_a(x+y) &= (\overline{x}+\overline{y}) - (x+y) = (\overline{x}-x) + (\overline{y}-y) = e_a(x) + e_a(y), \\ e_a(x-y) &= (\overline{x}-\overline{y}) - (x-y) = (\overline{x}-x) - (\overline{y}-y) = e_a(x) - e_a(y), \\ |e_a(x+y)| &= |e_a(x) + e_a(y)| \leq |e_a(x)| + |e_a(y)| \leq \varepsilon_a(x) + \varepsilon_a(y) \\ |e_a(x-y)| &= |e_a(x) - e_a(y)| \leq |e_a(x)| + |e_a(y)| \leq \varepsilon_a(x) + \varepsilon_a(y) \end{aligned} \tag{2.2.11}$$

I, per tant, $\varepsilon(x \pm y) = \varepsilon_a(x) + \varepsilon_a(y)$. Per a l'error relatiu:

$$\begin{aligned} \varepsilon_r(x+y) &= \left| \frac{x}{x+y} \right| \varepsilon_r(x) + \left| \frac{y}{x+y} \right| \varepsilon_r(y) \\ \varepsilon_r(x-y) &= \left| \frac{x}{x-y} \right| \varepsilon_r(x) + \left| \frac{y}{x-y} \right| \varepsilon_r(y) \end{aligned} \tag{2.2.12}$$

2.2.1.2 Producte

Per al producte, tenim el següent error absolut i relatiu:

$$\begin{aligned} e_a(xy) &= \overline{xy} - xy = (x + e_a(x))(y + e_a(y)) - xy = y e_a(x) + x e_a(y) + e_a(x) e_a(y) \\ e_r(xy) &= \frac{e_a(xy)}{xy} = \frac{e_a(x)}{x} + \frac{e_a(y)}{y} + \frac{e_a(x)}{x}\frac{e_a(y)}{y} = e_r(x) + e_r(y) + e_r(x)e_r(y) \approx e_r(x) + e_r(y). \end{aligned} \tag{2.2.13}$$

Observació 2.2.7. En aquest curs valorem i operem sobre l'error de primer ordre. Com els termes de la forma $e_a(x)e_a(y)$ són de segon ordre, els menyspreem.

Per calcular una fita de l'error relatiu podem fer el següent:

$$|e_r(xy)| \approx |e_r(x) + e_r(y)| \leq |e_r(x)| + |e_r(y)|, \tag{2.2.14}$$

on $\varepsilon_r(xy) = \varepsilon_r(x) + \varepsilon_r(y)$.

2.2.1.3 Divisió

Finalment, l'absolut i el relatiu aplicats a la divisió:

$$\begin{aligned} e_a\left(\frac{x}{y}\right) &= \frac{\overline{x}}{\overline{y}} - \frac{x}{y} = \frac{\overline{x}y - x\overline{y}}{y\overline{y}} = \frac{1}{1+e_r(y)}\frac{ye_a(x) - xe_a(y)}{y^2} \\ e_r\left(\frac{x}{y}\right) = e_r(x) - e_r(y) &\implies \left|e_r(\frac{x}{y})\right| \leq |e_r(x)| + |e_r(y)|. \end{aligned} \tag{2.2.15}$$

2.2.1.4 Resum

A continuació ho plantegem d'una altra manera, a partir d'aquesta espècie de quadre resum. Hi ha un punt en què fem servir el polinomi de Taylor, específicament en $\frac{1}{1+e_r(\overline{x_2})}$.

$$\left(\begin{array}{c|c} y = x_1 + x_2 & e_a(\overline{y}) = \overline{y} - y = \overline{x_1} - \overline{x_2} - x_1 + x_2 = e_a(\overline{x_1}) - e_a(\overline{x_2}) \\ \overline{y} = \overline{x_1} + \overline{x_2} & |e_a(\overline{y})| = |e_a(\overline{x_1}) + e_a(\overline{x_2})| \leq |e_a(\overline{x_1})| + |e_a(\overline{x_2})| \\ y = x_1x_2 & \overline{x_1x_2} = x_1x_2 + x_1e_a(\overline{x_2}) + x_2e_a(\overline{x_1}) + e_a(\overline{x_1})e_a(\overline{x_2}) \\ \overline{y} = \overline{x_1x_2} & e_r(\overline{x_1x_2}) = e_r(\overline{x_1}) + e_r(\overline{x_2}) + e_r(\overline{x_1})e_r(\overline{x_2}) \\ \overline{x_1} = x_1 + e_a(\overline{x_1}) & e_r(\overline{y}) = \dfrac{e_r(\overline{x_1}) - e_r(\overline{x_2})}{1 + e_r(\overline{x_2})} \\ \overline{x_2} = x_2 + e_a(\overline{x_2}) & \dfrac{\overline{x_1}}{\overline{x_2}} = \dfrac{x_1(1 + e_r(\overline{x_1}))}{x_2(1 + e_r(\overline{x_2}))} \\ \dfrac{1}{1 + e_r(\overline{x_2})} = 1 - e_r(\overline{x_2}) + O(e_r(\overline{x_2}))^2 & e_r(\overline{y}) = (e_r(\overline{x_1}) - e_r(\overline{x_2}))(1 - e_r(\overline{x_2}) + O(e_r(\overline{x_2}))^2) \\ |e_r(\overline{y})| \lesssim |e_r(\overline{x_1})| + |e_r(\overline{x_2})| & e_r(\overline{y}) \approx e_r(\overline{x_1}) - e_r(\overline{x_2}). \end{array}\right) \tag{2.2.16}$$

2.3

Representació d'un nombre

2.3.1 Representació en punt flotant

Sigui $\beta \in \mathbb{N} \setminus \{1\}$, que anomenarem *base*. El següent teorema garanteix l'existència teòrica de representació (infinita) en un punt flotant de qualsevol nombre real.

Teorema 2.3.1. *Fixem $\beta \in \mathbb{N}, b \geq 2$. Tot $x \in \mathbb{R} \setminus \{0\}$ pot ser representat de la forma*

$$x = s(\sum_{i=1}^{\infty} \alpha_i b^{-i})b^q, \tag{2.3.1}$$

amb $s \in \{-1, 1\}, q \in \mathbb{Z}$ i $\alpha_i \in \{0, 1, \ldots, \beta - 1\}$. A més, la representació anterior és única si $\alpha_1 \neq 0$ i els α_i no són tots $b-1$ d'una posició en endavant: $\forall i_o \in \mathbb{N} \exists i \geq i_0 \mid \alpha_i \neq \beta - 1$.

Notació 2.3.2. Escriurem l'expressió (2.3.1) abreujadament de la següent manera: $x = s(0.\alpha_1\alpha_2\dots)_b b^q$. El subíndex b dels parèntesis indica que els dígits $\alpha_1, \dots$ es troben en base b i s'anomenen *mantissa*. L'exponent q està limitat a un rang prefixat $q_{\min} \leq q \leq q_{\max}$.

2.3.2 Representació en punt fixe

Definició 2.3.3 (Nombres màquina)**.** Denotarem el conjunt de nombres representables exactament en aritmètica de punt flotant que *ocupa* t dígits, en base β i amb exponent entre $q_{\min}$ i $q_{\max}$ per $\mathcal{F}(\beta, t, q_{\min}, q_{\max})$, és a dir:

$$\mathcal{F}(\beta, t, q_{\min}, q_{\max}) = \{0\} \cup \{\pm(0.\delta_1 \cdots \delta_t)_b b^q, \delta_i \in \{0, 1, \dots, b-1\}, \delta_1 \neq 0, q_{\min} \leq q \leq q_{\max}\}. \tag{2.3.2}$$

és a dir, la seva longitud està fixada: a la seva representació també se l'anomena de *punt fixe*.

2.3.3 Arrodoniment i truncament en punt flotant

Donat $x \in \mathbb{R}$, si $x \notin \mathcal{F}(b, t, q_{\min}, q_{\max})$ no el podrem representar exactament, sinó que haurem d'operar amb una aproximació seva $\overline{x} \in \mathcal{F}(b, t, q_{\min}, q_{\max})$. Aquesta aproximació es pot trobar per truncament o arrodoniment:

Definició 2.3.4 (Truncament i arrodoniment)**.** Sigui $x \in \mathbb{R}$ amb representació en base b donada per

$$x = s(\sum_{i=1}^{\infty} \alpha_i b^{-i}) b^q, \ q_{\min} \leq q \leq q_{\max}. \tag{2.3.3}$$

Anomenem representació en punt flotant per truncament d'x a $fl_T(x) \in \mathcal{F}(b, t, q_{\min}, q_{\max})$ definit per

$$fl_T(x) = s(0.\alpha_1\alpha_2 \dots \alpha_t) b^q, \tag{2.3.4}$$

és a dir, $\delta_i = \alpha_i$ per a $i = 1, 2, \dots, t$. Anomenarem representació en punt flotant per arrodoniment de x a $fl_A(x) \in \mathcal{F}(b, t, q_{\min}, q_{\max})$ definit per:

$$fl_A(x) = \begin{cases} s(0.\alpha_1 \dots \alpha_t) b^q, & \text{si } 0 \leq \alpha_{t+1} < \frac{b}{2}, \\ s(0.\alpha_1 \dots \alpha_{t-1}(\alpha_t + 1)) b^q & \text{si } \frac{b}{2} \leq \alpha_{t+1} \leq b-1. \end{cases} \tag{2.3.5}$$

2.4 Aritmètica en punt flotant

Definició 2.4.1 (Sistema de punt flotant)**.** El sistema de punt flotant (β, t, L, U) és el conjunt de punts flotants normalitzats en base β i $t+1$ dígits significatius, és a dir, tots els nombres de la forma $x = m\beta^e$, on

$$\begin{gathered} m = \pm d_0.d_1 \dots d_t, 1 \leq d_0 \leq \beta - 1, \\ 0 \leq d_i \leq \beta - 1, \ i = 1, 2, \dots, t, \\ 1 \leq |m| < \beta, \end{gathered} \tag{2.4.1}$$

on tots els exponents satisfan $L \leq e \leq U$.

Quan els nombres són representats en el sistema de punt flotant (β, t, L, U) obtenim un error d'arrodoniment a causa d'una precisió limitada. Hem de delimitar una fita per a l'error relatiu.

Teorema 2.4.2 (Error relatiu d'aproximació). *L'error relatiu d'aproximació en una representació amb punt flotant es pot estimar com:*

$$\frac{|\overline{x} - x|}{|\overline{x}|} \leq \frac{1}{2}\beta^{-t} \tag{2.4.2}$$

on t és el nombre de dígits en base β, $\overline{x}$ el nombre aproximat i x el valor real.

Demostració. Sigui $x \neq 0$ i $\overline{x}$ el seu valor arrodonit a $t+1$ dígits. Podem representar el primer com un punt flotant normalitzat assignant-li $x = M\beta^e, 1 \leq |M| < \beta$ i el segon com $\overline{x} = m\beta^e$, on m és M arrodonit a $t+1$ dígits. Aleshores,

$$|m - M| \leq \frac{1}{2}\beta^{-t}, \tag{2.4.3}$$

i tenim una fita de l'error absolut:

$$|\overline{x} - x| \leq \frac{1}{2}\beta^{-t}\beta^e. \tag{2.4.4}$$

Això ens porta a considerar, per a la fita de l'error relatiu, el següent:

$$\left|\frac{\overline{x} - x}{x}\right| \leq \frac{\frac{1}{2}\beta^{-t}\beta^e}{|M|\beta^e} = \frac{\frac{1}{2}\beta^{-t}}{|M|} \leq \frac{1}{2}\beta^{-t}. \tag{2.4.5}$$

De la última inequació extraiem que $|M| \geq 1$. ■

Corol·lari 2.4.3. *Equivalentment, existeix un ε tal que $\overline{x} = x(1+\varepsilon)$.*

Observació 2.4.4. La fita per a l'error relatiu és independent de la magnitud d'x. Això vol dir que tant nombres grans com nombres petits es representen amb la mateixa precisió relativa.

En el producte de dos nombres amb $t+1$ dígits, s'obtenen $2t+1$ o $2t+2$ dígits. A part d'operacions aritmètiques i lògiques, hom ha de ser capaç de operar amb rotacions (*shifts*), que s'utilitzen principalment per igualar els exponents dels dos punts flotants que vulguem operar. Discutim breument la suma (i resta, per analogia) de punts flotants. Siguin $x = m_x\beta^{e_x}$ i $y = m_y\beta^{e_y}$ i $z = fl[x+y]$ la suma resultant. Assumint $e_x \geq e_y$ i, en particular, si $e_x > e_y$, y es trasllada $e_x - e_y$ posicions cap a la dreta abans de la suma:

$$x + y = \beta^{e_x}(m_x + m_y e^{y-x}) \tag{2.4.6}$$

Teorema 2.4.5 (Arrodoniment d'operacions aritmètiques). *Sigui $\odot$ qualsevol operador aritmètic del conjunt $\{+, -, \cdot, /\}$, assumint que $x \odot y \neq 0$ i x, y no tenen error en la representació:*

$$\left|\frac{fl[x \odot y] - x \odot y}{x \odot y}\right| \leq \mu \equiv fl[x \odot y] = (x \odot y)(1 \pm \varepsilon), \tag{2.4.7}$$

per a algun ε que satisfà $|\varepsilon| \leq \mu$ on $\mu = \frac{1}{2}\beta^{-t}$.

Corol·lari 2.4.6. *Amb error en la representació d'x, y considerem $fl(x), fl(y)$ i ens queda:*

$$fl(x \odot y) = fl(fl(x) \odot fl(y)) = [x(1 \pm \varepsilon_r(x)) \odot y(1 \pm \varepsilon_r(y))](1 \pm \varepsilon_r(x \odot y)) \tag{2.4.8}$$

Observació 2.4.7. Una conseqüència dels errors en operacions aritmètiques de punt flotant és que algunes propietats matemàtiques bàsiques deixen de funcionar. Estem parlant, per exemple, de l'associativitat (es perden sempre i quan haguem treballat sobre un cos que les tingués en primer lloc, com ara $\mathbb{R}$). Aleshores, pot passar que:

$$fl[fl[a+b]+c] \neq fl[a+fl[b+c]]. \tag{2.4.9}$$

2.5 Errors acumulats

Per treballar l'error acumulat en operacions iteratives de punts flotants ho farem sobre un exemple, la computació d'un simple sumatori.

Exemple 2.5.1. Sigui $S_n = 1 + \cdots + x_k$ i sigui $\hat{S}_i$ el sumatori parcial fins a i. Tenim:

$$\hat{S}_1 := x_1, \hat{S}_i := fl[\hat{S}_{i-1} + x_i], i = 2, 3, \ldots, n. \tag{2.5.1}$$

Si utilitzem 2.4.5 ens queda que:

$$fl[a+b] = (a+b)(1+\varepsilon) \implies \hat{S}_i = (\hat{S}_{i-1} + x_i)(1+\varepsilon_i), |\varepsilon| \leq \mu, \forall i. \tag{2.5.2}$$

Per inducció simple sobre n podem reescriure $\hat{S}_n$ com

$$\hat{S}_n = \hat{x}_1 + \cdots + \hat{x}_n, \quad \begin{array}{c} \hat{x_1} = x_1(1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n), \\ \hat{x_2} = x_2(1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n), \\ \vdots \\ \hat{x_n} = x_n(1+\varepsilon_n). \end{array} \tag{2.5.3}$$

Lema 2.5.2. *Siguin nombres $\varepsilon_1, \ldots, \varepsilon_r$ que satisfan $|\varepsilon_i| \leq \mu$, $i = 1, \ldots, r$ i assumim que $r\mu \leq 0.1$. Aleshores, existeix δ_r tal que*

$$(1+\varepsilon_1)(1+\varepsilon_2)\cdots(1+\varepsilon_r) = 1+\delta_r,\ |\delta_r| \leq 1.06r\mu. \tag{2.5.4}$$

Per tenir una fita aproximada dels errors relatius, seria convenient posar una condició de l'estil $n\mu < \Omega$, on Ω és tal valor. Ω serà 0.1.

Teorema 2.5.3 (Forward analysis)**.** *Si $n\mu \leq 0.1$, aleshores l'error en la suma es pot estimar de la següent manera:*

$$|\hat{S}_n - S_n| \leq |x_1||\delta_{n-1}| + |x_2||\delta_{n-1}| + |x_3||\delta_{n-2}| + \cdots + |x_n||\delta_1|,\ |\delta_i| \leq i \cdot 1.06\mu, \forall i. \tag{2.5.5}$$

Demostració. Podem escriure $\hat{S} = x_1(1+\delta_{n-1}) + x_2(1+\delta_{n-1}) + x_3(1+\delta_{n-2}) + \cdots + x_n(1+\delta_1)$, on δ_i satisfà la inequació del teorema. Restem S_n i utilitzem la desigualtat triangular. ■

En *forward analysis* es diu que la solució aproximada $\hat{S}$ que s'ha calculat per al problema $\mathcal{P}$ és la *solució exacta* d'un problema *pertorbat* $\hat{\mathcal{P}}$ i es determina la distància entre $\hat{\mathcal{P}}$ i $\mathcal{P}$. Mitjançant l'estudi de la pertorbació del problema aleshores és possible estudiar la diferència entre $\hat{S}$ i S.

> The objective of backward error analysis is to stop worrying about whether one has the "exact" answer, because this is not a welldefined thing in most real-world situations. What one wants is to find an answer which is the true mathematical solution to a problem which is within the domain of uncertainty of the original problem. Any result that does this must be acceptable as an answer to the problem, at least with the philosophy of backward error analysis. [Ric81]

En el cas de 2.5.1 podem formular el següent teorema:

Teorema 2.5.4. *La suma pot expressar-se com:*

$$\hat{S}_n = \hat{x}_1 + \cdots + \hat{x}_n, \quad \begin{array}{c} \hat{x}_1 = x_1(1+\delta_{n-1}), \\ \hat{x}_2 = x_2(1+\delta_{n-1}), \\ \hat{x}_3 = x_3(1+\delta_{n-2}), \\ \hat{x}_i = x_i(1+\delta_{n+1-i}, \forall i) \\ \hat{x}_n = x_n(1+\delta_1). \end{array} \tag{2.5.6}$$

Si $n\mu \leq 0.1$*, aleshores*

$$|\delta_k| \leq k \cdot 1.06\mu, 1 \leq k \leq n-1. \tag{2.5.7}$$

L'estimació d'error en aquests dos teoremes ens porta cap a una conclusió força important: podem reescriure els estimats com:

$$|\hat{S}_n - S_n| \leq ((n-1)|x_1| + (n-1)|x_2| + (n-2)|x_3| + \cdots + |x_n|)1.06\mu. \tag{2.5.8}$$

Això ens diu que, per acotar al màxim la fita d'error, ens caldria afegir els termes en un ordre creixent de magnitud; els primers termes tenen els factors més grans. De fet, es pot donar un error força gran en sèries convergents que se sumen en ordre decreixent.

2.6

Inestabilitat numèrica

$f(x+\Delta x) = \tilde{f}(x) = \tilde{y}$. Posem $f(x+\Delta x) = y + \Delta y$ i $|\Delta y| \leq \varepsilon|y|$ i $|\Delta x| \leq \varepsilon|x|$.

Definició 2.6.1 (Estable numèricament)**.** Diem que Δx és estable numèricament si ε és petit. En altres paraules, un algorisme és numèricament estable quan petites pertorbacions dels resultats inicials *no* produeixen una gran diferència en el resultat final.

Exemple 2.6.2. $x_0 = 1, x_1 = \frac{1}{3}, x_{n+1} = \frac{13}{3}x_n - \frac{4}{3}x_{n-1},\ n \geq 1$. La solució exacta és $x_n = \left(\frac{1}{3}\right)^n$. Siguin x_0 i x_1 qualsevol,

$$x_n = \frac{3(4x_0 - x_1)}{11}\left(\frac{1}{3}\right)^n + \frac{3x_1 - x_0}{11}4^n. \tag{2.6.1}$$

Si x_1 té error, posem un $\varepsilon = A$ tal que $x_n = A4^n$, amb $|A|$ petita (i serà estable). Suposem ara $x_0 = 1$ i $x_1 = 4$; d'aquesta manera, $x_n = 4^n$ i és estable.

2.6.1 Condicionament

Definició 2.6.3 (Ben condicionat)**.** Sigui ara $\tilde{y} = f(x + \Delta x)$ i anomenem $y = f(x)$.

$$\frac{\tilde{y} - y}{y} = \frac{f(x + \Delta x) - f(x)}{f(x)} = \frac{f'(x)\Delta x + \mathcal{O}(\Delta(x)^2)}{f(x)} \approx \frac{f'(x)}{f(x)}\frac{\Delta x}{x}x = \frac{xf'(x)}{f(x)} = e_r(x + \Delta x),$$

nombre de condició definida: $\left|\frac{xf'(x)}{f(x)}\right|$. (2.6.2)

Està ben condicionat si el nombre de condició definida és petit i mal condicionat si és gran. Anàlogament amb la distinció que hem fet amb l'estabilitat numèrica, direm que un problema està ben condicionat si petites variacions en les dades d'entrada *no* produeixen grans variacions en els resultats finals, independentment de l'algorisme emprat en la seva resolució.

Exercici 2.6.1. *Calculeu la distància focal d'una lent usant la fórmula*

$$\frac{1}{f} = \frac{1}{a} + \frac{1}{b}, \tag{2.6.3}$$

on $a = 32 \pm 1mm$ *i* $b = 46 \pm 1mm$*. Doneu una estimació de l'error.*

Resolució. Tenim que els valors aproximats d'a i b són $\overline{a} = 32$mm i $\overline{b} = 46$mm i que la fita dels seus errors absoluts és la mateixa $\varepsilon_a(a) = \varepsilon_a(b) = \varepsilon_a = 1$mm. Aïllant f obtenim que:

$$f = \frac{ab}{a + b} = h(a, b). \tag{2.6.4}$$

Així, $\overline{f} = h(\overline{a}, \overline{b}) \approx 18.87$. Ens faltaria trobar, doncs, la fita. Usarem la fórmula de propagació de l'error: ■

2.6.2 Avaluació de funcions

La computació numèrica involucra normalment funcions matemàtiques elementals. Quan les sèries numèriques de funcions no elementals s'utilitzen en ordinadors, la seva suma s'aproxima per a una suma parcial i després s'estima l'error per truncament.

Sigui

$$S = \sum_{n=1}^{\infty} a_n \tag{2.6.5}$$

una sèrie convergent. Assumim que no podem donar un resultat analític, però sí una aproximació numèrica.

Definició 2.6.4 (Suma parcial)**.** La suma parcial S_N és definida com

$$S_N = \sum_{n=1}^{N} a_n. \tag{2.6.6}$$

Com ja hem dit, usarem la suma parcial S_n com una aproximació de S.

Definició 2.6.5 (Residu). El corresponent error de truncament, el residu, és el següent:

$$R_N = S - S_N = \sum_{n=N+1}^{\infty} a_n. \tag{2.6.7}$$

Estimem l'error de truncament mitjançant l'acotació d'R_N. Primerament, considerem el cas que la sèrie és alternant i el valor absolut dels termes tendeix monòtonament cap a 0:

$$a_n a_{n+1} < 0,\ |a_n| > |a_{n+1}|,\ n = N, N+1, \ldots,\ \lim_{x\to\infty} a_n = 0. \tag{2.6.8}$$

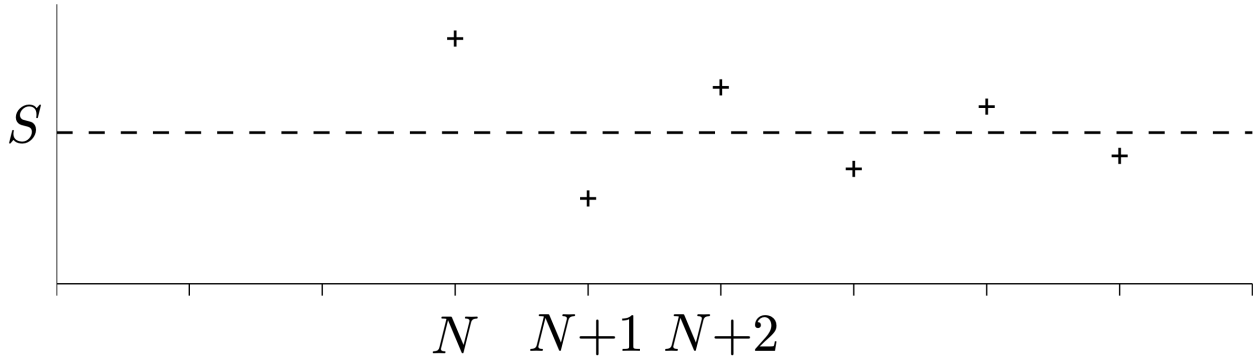

Figura 2.2: Un exemple de successió convergent en valors absoluts.

Fixem-nos que en 2.2 les sumes parcials $S_N, S_{N+2}, S_{N+4}, \ldots$ formen una successió monòtona; en conseqüència, la sèrie és convergent i el residu es pot aproximar a partir del primer terme negligit.

Teorema 2.6.6. *El terme residu d'una successió alternada convergent es pot estimar a partir del primer terme negligit,*

$$|R_N| \leq |a_{N+1}|. \tag{2.6.9}$$

2.7 Estàndard IEEE

El desenvolupament de microcomputadors ha fet necessari estandarditzar l'aritmètica de punt flotant. El principal objectiu és facilitar la *portabilitat*: que dos programes s'executin de la mateixa manera i tinguin resultat aproximadament idèntic.

Els components d'un nombre de punt flotant x són el signe s (un bit), l'exponent E (8 bits) i la fracció f (23 bits). El valor v de x és:

1. $v = (-1)^s(1.f)2^{E-127}$ si $0 < E < 255$.
2. $v = (-1)^s(0.f)2^{-126}$ si $E = 0$ i $f \neq 0$.
3. $v = (-1)^s 0$ si $E = 0$ i $f = 0$.
4. $v =$ `NaN` si $E = 255$ i $f \neq 0$.
5. $v = (-1)^s$`infty` si $E = 255$ i $f = 0$.

III

Àlgebra lineal numèrica

3.1

Conceptes bàsics

3.1.1 Introducció: sistemes d'equacions

Els sistemes d'equacions són molt comuns. Ara ens tocarà estudiar el nombre de solucions dels sistemes lineals amb n incògnites. Suposem que volem trobar $x_1, \dots, x_n \in \mathbb{R}$ tal que

$$\begin{aligned} a_1^1 x_1 + \cdots + a_n^1 x_n &= b_1 \\ &\vdots \\ a_1^n x_1 + \cdots + a_n^n x_n &= b_n \end{aligned} \tag{3.1.1}$$

que podem escriure de forma matricial, recordem, $Ax = b$, on

$$A = \begin{pmatrix} a_1^1 & \cdots & a_n^1 \\ \vdots & \ddots & \vdots \\ a_1^n & \cdots & a_n^n \end{pmatrix},\ b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}. \tag{3.1.2}$$

Seria convenient interpretar la matriu A del sistema com un conjunt d'n vectors columna tals que

$$A = [a_{:1} \cdots a_{:n}],\ a_{:j} = \begin{pmatrix} a_j^1 \\ \vdots \\ a_j^n \end{pmatrix}. \tag{3.1.3}$$

En altres paraules, volem que $x_1 a_1^{\vdots} + \cdots + x_n a_n^{\vdots} = b$ i el problema de solucionar aquest sistema es converteix en trobar una combinació lineal de vectors $a_{:1}, \dots, a_{:n}$ tal que sigui igual al vector b. A més, necessitem que $\det A \neq 0$, ja que el sistema té solució per a b si, i només si, els vectors columna són linealment independents si, i només si, els coeficients de la diagonal no són nuls.

Definició 3.1.1 (Matriu regular)**.** Es diu que la matriu A és regular (o no singular) si els vectors columna són linealment independents, amb la qual cosa la solució del sistema $Ax = b$ és única.

Definició 3.1.2 (Mètode directe)**.** Obtenim la solució amb un nombre finit d'operacions aritmètiques simples i, en l'absència d'errors d'arrodoniment, l'x de sortida seria la solució exacta del sistema.

Definició 3.1.3 (Mètode iteratiu)**.** Comencem per $x^{(0)} \in \mathbb{R}^n$ i generen una successió de vectors $x_n^{|n|} \implies \overline{x}$, $A\overline{x} = b$. És a dir, convergeix cap a la solució.

3.1.2 Tipus de matrius

Denotarem per $\mathcal{M}_{m,n}$ l'espai vectorial de les matrius complexes $m \times n$ de m files i n columnes.

Definició 3.1.4.

1. Els vectors de dimensió m són elements d'$\mathcal{M}_{m,1}$.
2. Els elements d'una matriu $M = (m_{kj}) \in \mathcal{M}_{m,n}$ seran denotats per $m_{kj}, 1 \leq k \leq m, 1 \leq j \leq n$.
3. Els elements d'un vector $y = (y_k)$ es denotaran per $y_k, 1 \leq k \leq m$.
4. Per a una matriu $M \in \mathcal{M}_{m,n}$, denotarem la seva transposada per M^T, la seva conjugada per $\overline{M}$ i la seva adjunta per M^*. $M^T \in \mathcal{M}_{n,m}$ es forma canviant files per columnes, $\overline{M} \in \mathcal{M}_{m,n}$ es forma conjugant els elements d'M i $M^* \in \mathcal{M}_{m,n}$ és la transposada de la matriu conjugada d'M.
5. Les matrius quadrades són aquelles tals que $m = n$.
6. Denotarem la matriu identitat amb Id, la inversa per M^{-1} i el determinant per $\det M$.
7. Les matrius $k \times k$ formades per les k primeres files de les k primeres columnes s'anomenen submatrius principals d'ordre k d'M i denotades per $(M)_k$. Anomenarem determinants principals d'M als determinants de les seves submatrius principals i els denotarem per $\det A_k$.

Ara veurem diferents tipus de matrius:

1. A singular: $\det A = 0$;
2. A regular: $\det A \neq 0$;
3. A hermítica: $A^* = A$;
4. A simètrica: $A^T = A$;
5. A unitària: $A^{-1} = A^*$;
6. A ortogonal: $A^{-1} = A^T$;
7. A definida positiva: A hermítica i $x^*Ax > 0, \forall x \neq 0$.
8. A estrictament diagonal dominant:

$$|a_{ii}| > \sum_{i \neq j, j=1} |a_{ij}|, 1 \leq i \leq n. \tag{3.1.4}$$

Notem que les definicions de matrius unitàries i hermítiques corresponen a matrius complexes i les de matrius ortogonals i simètriques corresponen a matrius reals. Noteu que les matrius unitàries i hermítiques reals són, respectivament, matrius ortogonals i simètriques.

Definició 3.1.5 (Matrius semblants)**.** Dues matrius A, B s'anomenen semblants si existeix una matriu invertible C, anomenada transformació de semblança d'A a B, tal que es compleix $B = C^{-1}AC$.

Definició 3.1.6 (Diagonalitzable)**.** Una matriu A és diagonalitzable quan és semblant a una matriu diagonal.

3.1.3 Valors i vectors propis

Definició 3.1.7 (Matriu diagonal)**.** Donats escalars $\lambda_1, \ldots, \lambda_n$, on $n \geq 1$ és un enter, denotarem per $D(\lambda_1, \ldots, \lambda_n)$ la matriu diagonal de mida $n \times n$ donada per

$$D(\lambda_1, \ldots, \lambda_n) := \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}. \tag{3.1.5}$$

Definició 3.1.8 (Vector propi)**.** Diem que un vector no nul $u \in E$ és un *vector propi de* f si existeix un escalar $\lambda \in \mathbb{K}$ tal que $f(u) = \lambda u$.

Definició 3.1.9 (Valor propi)**.** Si u és un vector propi de f, l'escalar λ s'anomena *valor propi* de u.

Observació 3.1.10.

1. La definició de vector propi i de valor propi d'un endomorfisme no utilitza bases.
2. En la definició que hem donat de vector propi s'exigeix que el vector sigui diferent de zero. El vector $0 \in E$ satisfà l'equació $f(0) = \lambda \cdot 0$ per a tot $\lambda \in \mathbb{K}$, però 0 no se sol considerar com a valor propi de f.
3. Si u és un vector propi λ aleshores, per a tot $a \in \mathbb{K}, a \neq 0$, el vector au també és un vector propi de f del mateix valor propi (unicitat del valor propi).
4. Els vectors no nuls del nucli de f són vectors propis de valor propi $\lambda = 0$.

Exemple 3.1.11. Si A és una matriu diagonal,

$$A = D(\lambda_1, \lambda_2) = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \tag{3.1.6}$$

aleshores,

$$A \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \lambda_1 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad A \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \lambda_2 \end{pmatrix} = \lambda_2 \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{3.1.7}$$

Definició 3.1.12 (Subespai propi de f de valor propi λ)**.** Si $\lambda \in \mathbb{K}$, denotarem per E_λ el subespai vectorial d'E donat per

$$E_\lambda := \ker(f - \lambda I). \tag{3.1.8}$$

$u \neq 0$ és vector propi de f si, i només si, $f(u) - \lambda u = 0$. Això equival a demanar que $u \in E_\lambda$. D'altra banda, un escalar $\lambda \in \mathbb{K}$ és valor propi si, i només si, $E_\lambda \neq \{0\}$.

Definició 3.1.13 (Multiplicitat geomètrica i algebraica)**.** Sigui λ un valor propi de f, Definim la *multiplicitat geomètrica* de λ com la dimensió de l'espai vectorial E_λ. La *multiplicitat algebraica* és la multiplicitat de λ com a arrel del polinomi característic. $MG \leq MA$.

Proposició 3.1.14. *Si r és la multiplicitat del valor propi k, és a dir, si es té $r = \dim(\ker(f - k\mathbb{I}))$ i s és la multiplicitat del zero k del polinomi característic, aleshores $r \leq s$.*

Proposició 3.1.15. *Un escalar $\lambda \in \mathbb{K}$ és valor propi de f si, i només si,* $\det(f - \lambda I) = 0$.

Demostració. Sabem que el subespai $E_\lambda := \ker(f - \lambda I)$ és trivial si, i només si, la imatge de $f - \lambda I$ té rang màxim. Això passa si, i només si, $\det(f - \lambda I) \neq 0$, tal i com volíem provar. ■

Definició 3.1.16 (Polinomi característic)**.** El polinomi característic d'una matriu quadrada A és $p_A(\lambda) := \det(A - \lambda I)$, on

$$\begin{pmatrix} a_1^1 - \lambda & a_2^1 & \dots & a_n^1 \\ a_1^2 & a_2^2 - \lambda & \dots & a_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^n & a_2^n & \dots & a_n^n - \lambda \end{pmatrix} \tag{3.1.9}$$

correspon a la matriu $A - \lambda I$. Aquest polinomi no depèn de la base escollida, sinó de l'endomorfisme en si: parlem del polinomi característic d'un *endomorfisme*, $p_f(\lambda)$.

Procés 3.1.17 (Resum dels passos per al càlcul de valors i vectors propis)**.**

1. *Buscar les arrels* $\{\lambda_1, \ldots, \lambda_n\}$ *del polinomi característic* $p_A(\lambda) := \det(A - \lambda I)$*. Notem que poden haver-hi diverses arrels amb el mateix valor.*
2. *Per a cada arrel* λ_i*: buscar les solucions del sistema homogeni*

$$(A - \lambda_i I) \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.1.10}$$

3. *Escollir vectors representant d'aquestes solucions per a descriure els subespais propis i la multiplicitat de cada valor propi* λ_i*.*

Propietat 3.1.18 (Propietats dels valors propis)**.**

1. *Les matrius* A *i* A^T *tenen els mateixos valors propis. Els vectors propis per la dreta són ortogonals als vectors propis per l'esquerra de valors propis diferents.*
2. *Una matriu* A *és regular si, i només si, tots els seus valors propis són diferents de zero; en tal cas, els valors propis d'*A^{-1} *són els recíprocs dels valors propis d'*A *i* v *és un vector propi de valor propi* λ *d'*A *si, i només si,* v *és un vector propi de valor propi* $\frac{1}{\lambda}$ *d'*A^{-1}*.*
3. *Dues matrius semblants* A *i* B *tenen els mateixos valors propis si, i* v *és un vector propi de valor propi* λ *d'*A *si, i només si,* $C^{-1}v$ *és un vector propi de valor propi* λ *de* B*, essent* C *la transformació de semblança d'*A *a* B*.*
4. *Sigui* A *diagonalitzable per una transformació de semblança* V*. Aleshores,* $D = V^{-1}AV$ *és diagonal, els elements de la diagonal de* D *són els valors propis d'*A*, les columnes de* V *formen una base de vectors propis (per la dreta) d'*A *i les fileres de* V^{-1}*, una base de vectors propis per l'esquerra d'*A*.*

3.1.4 Normes vectorials i matricials

Definició 3.1.19 (Norma)**.** Sigui E un espai vectorial sobre $\mathbb{R}$ o $\mathbb{C}$. Una norma a E és una aplicació

$$\begin{array}{rccl} ||\ \ || : & E & \longrightarrow & \mathbb{R} \text{ (sempre a } \mathbb{R}!) \\ & v & \longmapsto & ||v|| \end{array} \tag{3.1.11}$$

que compleix

1. $||v|| = 0 \iff v = \vec{0}$,
2. $||kv|| = |k| \cdot ||v||$,
3. $||u + v|| \leq ||u|| + ||v||$ (desigualtat triangular).

$|k|$ indica el valor absolut si $k \in \mathbb{R}$ o bé indica el mòdul si $k \in \mathbb{C}$.

Definició 3.1.20 (Normes vectorials)**.** Les normes vectorials són aquelles que estan definides en espais vectorials de la forma $E = \mathbb{R}^n$ o $E = \mathbb{C}^n$.

Definició 3.1.21 (Norma matricial)**.** Una norma matricial és una norma en l'espai vectorial $\mathcal{M}_{n,n}$ que sigui multiplicativa; això és, que compleixi

$$||AB|| \leq ||A|| \cdot ||B||,\ \forall A, B \in \mathcal{M}_{n,n}. \tag{3.1.12}$$

Propietat 3.1.22 (Propietats de les normes matricials)**.**

1. *Les normes matricials subordinades a les normes vectorials* $||\quad||_1, ||\quad||_2, ||\quad||_\infty$ *resulten ser:*

$$\begin{aligned} ||A||_1 &= \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|, \\ ||A||_2 &= \sqrt{\rho(A^*A)}, \\ ||A||_\infty &= \max_{1\leq i\leq n} \sum_{i=1}^{n} |a_{ij}|, \end{aligned} \tag{3.1.13}$$

2. *Donada una matriu* A *qualsevol i per a qualsevol* $\epsilon > 0$*, es pot trobar una norma matricial, subordinada a una certa norma vectorial, tal que*

$$||A|| \leq \rho(A) + \epsilon. \tag{3.1.14}$$

3.2

Sistemes triangulars

Dintre dels sistemes triangular, en tenim de dos tipus: els inferiors i els superiors. Essencialment, són molt similars i estan estretament lligats a l'estructura de la matriu sobre la qual es construeixen.

Definició 3.2.1 (Matriu triangular superior)**.** Una matriu $n \times n$, $U = (u^i_j)$, és triangular superior (*upper triangular*) si $u^i_j = 0, i > j$.

Definició 3.2.2 (Matriu triangular inferior)**.** Una matriu $n \times n$, $L = (l^i_j)$, és triangular inferior (*lower triangular*) si $l^i_j = 0, i < j$.

$$U = \begin{pmatrix} u^1_1 & u^1_2 & u^1_3 & \cdots & u^1_n \\ 0 & u^2_2 & u^2_3 & \cdots & u^2_n \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & u^n_n \end{pmatrix}, \quad L = \begin{pmatrix} l^1_1 & 0 & 0 & \cdots & 0 \\ l^2_1 & l^2_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ l^n_1 & l^n_2 & l^n_3 & \dots & l^n_n \end{pmatrix} \tag{3.2.1}$$

Figura 3.1: Una matriu triangular superior U i una inferior L.

Ja hem comentat a la primera secció que U és regular si, i només si, tots els elements de la diagonal u_{ii} són diferents de zero. En tal cas, el sistema

$$\begin{aligned} u_{11}x_1 + u_{12}x_2 + \cdots + u_{1n}x_n &= c_1 \\ u_{22}x_2 + \cdots + u_{2n}x_n &= c_2 \\ &\vdots \\ u_{nn}x_n &= c_n \end{aligned} \tag{3.2.2}$$

es pot resoldre amb el següent algorisme de substitució enrere.

Algorisme 3.2.3 (Algorisme de substitució enrere)**.** *Usualment utilitzat per trobar la solució única d'un sistema amb una matriu triangular superior regular, és a dir, aquella en què tots els elements de la diagonal són diferents a 0. S'aplica iterativament el següent:*

$$\begin{aligned} x_n &= \frac{c_n}{u_n^n}, \\ x_i &= \frac{c_i - \sum_{j=i+1}^{n} u_j^i x_j}{u_i^i}, i = n-1, \dots, 1. \end{aligned} \tag{3.2.3}$$

De manera totalment anàloga, la matriu L és regular si $l_{ii} \neq 0,\ \forall i$. En tal cas, el sistema

$$\begin{aligned} l_{11}y_1 &= d_1 \\ l_{21}y_1 + l_{22}y_2 &= d_2 \\ &\vdots \\ l_{n1}y_1 + l_{n2}y_2 + \cdots + l_{nn}y_n &= d_n \end{aligned} \tag{3.2.4}$$

es pot resoldre amb l'algorisme de substitució endavant.

Algorisme 3.2.4 (Algorisme de substitució endavant)**.** *Utilitzat per trobar la solució única d'un sistema amb una matriu triangular inferior regular. S'aplica iterativament el següent:*

$$\begin{aligned} y_1 &= \frac{d_1}{l_1^1}, \\ y_i &= \frac{d_i - \sum_{j=1}^{i-1} l_j^i y_j}{l_i^i}, i = 2, \dots, n. \end{aligned} \tag{3.2.5}$$

Operacions	Total	
$+, -$	$\sum_{i=1}^{n-1}(n-i) = \frac{n(n-1)}{2}$	$\sum_{i=1}^{n-1}(n-i) = \frac{n(n-1)}{2}$
$\cdot$	$\sum_{i=1}^{n-1}(n-i) = \frac{n(n-1)}{2}$	$\sum_{i=1}^{n-1}(n-i) = \frac{n(n-1)}{2}$
$/$	$\sum_{i=1}^{n} 1 = n$	$\sum_{i=1}^{n} 1 = n$
TOTAL	$n + 2\frac{n(n-1)}{2} = n^2$	$n + 2\frac{n(n-1)}{2} = n^2$
	forward	*backward*

Figura 3.2: Nombre d'operacions en els dos algorismes de substitucions que hem vist

Observació 3.2.5. $\sum_{j=i+1}^{n} 1 = n - i \implies \sum_{i=1}^{n-1}(n-i) = \frac{n(n-1)}{2}$.

Definició 3.2.6 (Matriu unitària triangular)**.** És aquella matriu tal que és triangular inferior i els coeficients de la diagonal, l_i^i, són tots iguals a 1.

Definició 3.2.7 (Flops)**.** Definim un *flop* com una operació aritmètica simple amb punts flotants. Estem parlant de suma, multiplicació i divisió. Anteriorment, es considerava com a *flop* una suma i una multiplicació o bé una divisió.

Corol·lari 3.2.8. *La solució d'un sistema amb una matriu triangular amb n incògnites necessita n^2 flops, excepte en cas que la matriu també és unitària, quan en necessita $n(n-1)$.*

3.3

Eliminació gaussiana

L'objectiu que persegueix l'eliminació gaussiana és transformar el sistema lineal $Ax = b$ en un sistema $Ux = c$, triangular superior, al qual aplicarem l'algorisme de substitució enrere que hem vist a la secció anterior. El mètode d'eliminació gaussiana consta de $n-1$ passos i, en cada pas, s'intenta eliminar x_k de les últimes $n-k$ equacions.

Notació 3.3.1. A partir d'ara, ens referirem als coeficients de les matrius amb els dos índexs a la part inferior, tal que les files seran el primer i les columnes el segon: a_{ij}, i files, j columnes.

Definició 3.3.2 (Eliminació gaussiana)**.** Una sèrie de transformacions simples que resulten en una matriu triangular. Les mateixes transformacions s'apliquen a la banda dreta:

$$(A|b) \longrightarrow (U|c), \tag{3.3.1}$$

on U és triangular superior.

Algorisme 3.3.3 (Eliminació gaussiana)**.** *Solucionarem un sistema mitjançant eliminació gaussiana per al cas general n. Notem que comencem amb una matriu ampliada de la forma*

$$(A|b) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right) \tag{3.3.2}$$

Ens cal assumir que els elements de la diagonal no són nuls i reduïm la matriu:

$$\left(\begin{array}{ccccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1r}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & \overline{a}_2^{(2)} & \overline{a}_{23}^{(2)} & \cdots & \overline{a}_{2r}^{(2)} & \cdots & \overline{a}_{rn}^{(2)} & \overline{b}_2^{(2)} \\ \vdots & \vdots & \ddots & & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \overline{a}_{rr}^{(r)} & \ldots & \overline{a}_{rn}^{(r)} & \overline{b}_r^{(r)} \\ 0 & 0 & 0 & \ldots & \overline{a}_{r+1,r}^{(r)} & \ldots & \overline{a}_{r+1,n}^{(r)} & \overline{b}_{r+1}^{(r)} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \overline{a}_{nr}^{(r)} & \ldots & \overline{a}_{nn}^{(r)} & \overline{b}_n^{(r)} \end{array}\right) \tag{3.3.3}$$

on els elements que han variat els identificarem amb $\overline{a}_j^i$ i l'exponent identifica el pas i-èsim d'eliminació. Ara ho farem amb més detall: ens toca suposar novament que $a_{rr}^{(r)} \neq 0$. Per reduir la columna r els elements $\overline{a}_{ij}^{(r)} \;\big|\; r < i \leq n, r < j \leq n$ es transformen en:

$$\left.\begin{array}{c} m_{ir} = \dfrac{\overline{a}_{ir}^{(r)}}{\overline{a}_{rr}^{(r)}}; \\ \overline{a}_{ij}^{(r+1)} = \overline{a}_{ij}^{(r)} - m_{ir}\overline{a}_{rj}^{(r)}, j = r+1, \ldots, n; \\ \overline{b}_i^{(r+1)} = \overline{b}_i^{(r)} - m_{ir}b_r^{(r)}; \end{array}\right\} i = k+1, \ldots, n. \tag{3.3.4}$$

Ens adonem que, en efecte, els elements $a_{ir}^{(r+1)}$ amb $r < i \leq n$, és a dir, els elements de la columna r que queden per sota la diagonal, s'anul·len.

$$\overline{a}_{ir}^{(r+1)} = \overline{a}_{ir}^{(r)} - m_{ir}\overline{a}_{rr}^{(r)} = \overline{a}_{ir}^{(r)} - \frac{\overline{a}_{ir}^{(r)}}{\overline{a}_{rr}^{(r)}}\overline{a}_{rr}^{(r)} = 0. \tag{3.3.5}$$

La fila que s'utilitza per fer zeros a la k-èsima columna s'anomena k-èsima fila de pivot i $\overline{a}_{rr}^{(r)}$ és l'element de pivot. L'algorisme d'eliminació gaussiana es basa en el procés de reducció exposat en (3.3.5) per a $k = 1, 2, \ldots, n-1$. El sistema $Ax = b$ deriva, finalment, en $Ux = c$.

Proposició 3.3.4. *La transformació d'una matriu $n \times n$ a la seva forma triangular per a un sistema d'n equacions utilitzant l'eliminació gaussiana necessita, aproximadament, de $\frac{2}{3}n^3$ flops.*

Demostració. Volem saber la complexitat del procés de transformació del sistema $Ax = b$ a $Ux = c$ amb n incògnites. El r-èsim pas de transformació involucra $n - k$ files i per cada fila necessitem d'una divisió (per obtenir m_{ir}) i $n - r + 1$ multiplicacions i sumes; aproximadament, això correspon a uns $2(n-r)^2$ *flops* al k-èsim pas, per a un total de

$$2\sum_{k=1}^{n-1}(n-k)^2 = 2\sum_{\nu=1}^{n-1}\nu^2. \tag{3.3.6}$$

Lema 3.3.5.

$$S_n \equiv \sum_{\nu=1}^{n-1}\nu^2 = \frac{n(n-1)(2n-1)}{6}. \tag{3.3.7}$$

Demostració. Per inducció, la fórmula és evident per a $n = 1$. Suposem que és cert per a $n = N$ i ho provem per a $N + 1$:

$$\begin{aligned} S_{N+1} &= S_N + N^2 = \frac{N}{6}(2N^2 - 3N + 1 + 6N) \\ &= \frac{N}{6}(2N+1)(N+1) = \frac{(N+1)N(2(N+1)-1)}{6}. \end{aligned} \tag{3.3.8}$$

■

Hem de tenir en compte que, normalment, en l'aplicació d'aquest algorisme per a matrius grans, la majoria d'elements són zeros i es poden utilitzar altres mecanismes per a reduir la quantitat de càlculs. Això queda fora de l'abast d'aquest curs. En definitiva, en notació de Landau, podem dir que aquest algorisme té una complexitat de l'ordre d'$\mathcal{O}(x^3)$. ■

Observació 3.3.6. A més, en el pas r-èsim d'eliminació gaussiana hem de dividir per l'element de la diagonal $\overline{a}_{rr}^{(r)}$. Hem de fer diverses observacions:

- En aquest algorisme hem assumit que tots els pivots eren diferents a zero. De totes maneres, en el cas que $\overline{a}_{rr}^{(r)} = 0$, llavors podem cercar $\overline{a}_{ir}^{(r)} \neq 0, i > k$ (que existeix sempre, en ser $\det A \neq 0$) i permutar l'equació r-èsima amb la i-èsima.
- En cas que $\overline{a}_{rr}^{(r)} \neq 0$ però és petit, el mètode de Gauss, tot i que teòricament és aplicable, és molt inestable numèricament en el sentit que els errors de la solució, propagats a partir dels errors de les dades, poden ser augmentats notablement.

3.4

Estratègies de pivotatge

Definició 3.4.1 (Pivotatge)**.** En el procés d'eliminació gaussiana anomenarem a l'intercanvi de files *pivotatge*.

Per a reduir una possible falta de precisió (revisar l'exemple de [EWN04, pàg. 206 - 207]) es recomana que els pivotatges es facin amb un pivot el més gran possible. No cal fer les operacions que sabem que donen com a resultat 0. Puc resoldre amb la mateixa matriu diversos sistemes simultàniament.

Algorisme 3.4.2 (Pivotatge parcial)**.** *Considerem el pas r-èsim de l'eliminació gaussiana, recollit a* (3.3.3)*. Agafem la r-èsima columna i busquem, des d'$\overline{a}_{rr}^{(r)}$ fins al final, l'element de major magnitud. Trobem l'índex de fila ν tal que*

$$|a_{\nu r}| = \max_{r \leq i \leq n} |\overline{a}_{ir}^{(r)}|. \tag{3.4.1}$$

Si $\nu > r$, aleshores les files r i ν s'intercanvien i es procedeix amb l'eliminació gaussiana. Efectivament, amb el pivotatge parcial els m_{ir} satisfan

$$|m_{ir}| = \left| \frac{\overline{a}_{ir}^{(r)}}{\overline{a}_{rr}^{(r)}} \right| \leq 1. \tag{3.4.2}$$

Algorisme 3.4.3 (Pivotatge total)**.** *És una extensió del pivotatge parcial: el r-èsim pivot s'obtindria buscant l'element més gran en la submatriu*

$$\begin{pmatrix} \overline{a}_{rr}^{(r)} & \cdots & \overline{a}_{rn}^{(r)} \\ \vdots & \ddots & \vdots \\ \overline{a}_{nr}^{(r)} & \cdots & \overline{a}_{nn}^{(r)} \end{pmatrix} \tag{3.4.3}$$

Una vegada trobat, es procedeix a intercanviar files i columnes fins que tal element ocupi la posició (r, r). És un mètode que pràcticament no s'utilitza avui en dia.

L'objectiu del pivotatge és evitar que els elements de la matriu esdevinguin massa grans durant el procés d'eliminació gaussiana i aquest fet derivi en pèrdua de precisió.

Definició 3.4.4 (Definida positiva)**.** Una matriu simètrica A és definida positiva si $x^T Ax \geq 0$ i és igual a 0 si, i només si, $x = \vec{0}$. La definició és força anàloga a la que vam donar el seu moment de *forma lineal definida positiva* a *Àlgebra Lineal*.

Definició 3.4.5 (Diagonalment dominant)**.** Una matriu $A = (a_{ij})$ és diagonalment dominant si

$$|a_{ii}| \geq \sum_{j=1, i \neq j}^{n} |a_{ij}|, i = 1, 2, \ldots, n \tag{3.4.4}$$

Aleshores,

1. Es recomana usar aquest mètode quan es treballa amb l'eliminació gaussiana per a un sistema $Ax = b$.

2. No es recomana el pivotatge en matrius diagonalment dominants ni en aquelles simètriques i definides positives.

Observació 3.4.6. L'eliminació gaussiana sense pivotatges és estable per a aquests dos tipus de matrius: l'element més gran contingut en A és més gran que qualsevol generat durant l'eliminació. De fet, els multiplicadors m seran més petits o iguals que 1, com hem dit a (3.4.2).

Recordem que al principi del capítol dèiem que un sistema $Ax = b$ tenia solució única si els vectors columna de la matriu A eren linealment independents.

Exemple 3.4.7.

$$\begin{pmatrix} -10^5 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ x_1 = -\frac{1}{2.00001} \approx -0.4999975 = -\frac{1}{2}x_2$$

$$(10, 3, -9, 9) \text{ (base, fracció, exponents, mínim, màxim)} \tag{3.4.5}$$

$$\begin{matrix} a'_{22} = 2 \cdot 10^5 \\ b'_2 = 2 \cdot 10^5 \end{matrix}, \quad \begin{pmatrix} -10^5 & 1 \\ 0 & 2 \cdot 10^5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \cdot 10^5 \end{pmatrix}, x_2 = 1, x_1 = 0.$$

Proposició 3.4.8. *Si els vectors columna de la matriu A són linealment independents, l'eliminació gaussiana amb pivotatge parcial té tots els pivots diferents a zero.*

Demostració. Per la descripció de 3.4.2 se segueix que el procés s'acaba al pas r-èsim si $\overline{a}_{ir} = 0, \forall i \geq r$. Això pot ocórrer solament en cas que una columna $\overline{a}_{:r}$ és una combinació lineal de $\overline{a}_{:1}, \ldots, \overline{a}_{:r-1}$. Considerem el sistema:

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1,r-1} & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2,r-1} & a_{2r} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{n,r-1} & a_{nr} \end{array}\right) \tag{3.4.6}$$

Sabem que aquesta matriu es transforma mitjançant l'eliminació gaussiana amb pivotatge parcial en una matriu de la següent forma

$$\left(\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1,r-1}^{(1)} & a_{1r}^{(1)} \\ 0 & \overline{a}_{22}^{(2)} & \overline{a}_{23}^{(2)} & \cdots & \overline{a}_{2,r-1}^{(2)} & \overline{a}_{2r}^{(2)} \\ 0 & 0 & \overline{a}_{33}^{(3)} & \cdots & \overline{a}_{3,r-1}^{(3)} & \overline{a}_{3r}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \overline{a}_{r-1,r-1}^{(r-1)} & \overline{a}_{r-1,r}^{(r-1)} \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{array}\right), \tag{3.4.7}$$

on tot $a_{11}^{(1)}, \overline{a}_{ii}^{(i)} \neq 0, i = 2, \ldots, r-1$. Això implica que podem trobar $\{\alpha_i\}_{i=1}^{k-1}$ tal que

$$\alpha_1 a_{:1} + \cdots + \alpha_{k-1} a_{:k-1} - a_{:k} = 0 \tag{3.4.8}$$

i, per tant, les r primeres columnes no són linealment independents. Així doncs, sota l'assumpció que les columnes d'A són linealment independents, l'eliminació gaussiana amb pivotatge parcial utilitza pivots tots diferents a zero. Cap esmentar que aquest resultat negligeix els efectes dels errors d'arrodoniment. ■

Corol·lari 3.4.9. *Si* $\det A \neq 0$*, aleshores sempre es pot fer eliminació gaussiana amb pivotatge.*

3.5

Aplicacions

3.5.1 Permutacions

L'eliminació gaussiana és equivalent a una factorització de la matriu. Un intercanvi de files és equivalent a una multiplicació per l'esquerra per una matriu de permutació.

Definició 3.5.1 (Matriu de permutació simple o transposició)**.** La matriu més simple que ho compleix, que anomenarem P_{rs}, és obtinguda intercanviant les files r i s a la matriu unitària I i difereix d'I en quatre posicions: té zeros en les posicions $(r,r),(s,s),(r,s),(s,r)$.

Si multipliquem un vector x per P_{rs}, aleshores els elements x_r i x_s s'intercanvien. Anàlogament:

$$P_{rs}A = P_{rs}[a_{:1}a_{:2}\cdots a_{:n}] = [P_{rs}a_{:1}P_{rs}a_{:2}\cdots P_{rs}a_{:n}] \tag{3.5.1}$$

mostra que les files r i s són intercanviades. Amb $r<s$, tenim:

$$P_{rs}A = \begin{pmatrix} \vdots & \vdots & & \vdots \\ a_{r-1,1} & a_{r-1,2} & \cdots & a_{r-1,n} \\ a_{s,1} & a_{s,2} & \cdots & a_{s,n} \\ a_{r+1,1} & a_{r+1,2} & \cdots & a_{r+1,n} \\ \vdots & \vdots & & \vdots \\ a_{s-1,1} & a_{s-1,2} & \cdots & a_{s-1,n} \\ a_{r,1} & a_{r,2} & \cdots & a_{r,n} \\ a_{s+1,1} & a_{s+1,2} & \cdots & a_{s+1,n} \\ \vdots & \vdots & & \vdots \end{pmatrix} \tag{3.5.2}$$

Observació 3.5.2. Notem que una matriu de permutació simple és simètrica i és la seva pròpia inversa:

$$P_{rs} = P_{rs}^T = P_{rs}^{-1}. \tag{3.5.3}$$

La definició general d'una matriu de permutació és una matriu P tal que Px és solament una reordenació dels components en x. Així doncs, un producte de transposicions és una matriu de permutació i, de fet, totes les matriu de permutacions es poden construir d'aquesta manera.

Proposició 3.5.3. *Qualsevol matriu de permutació es pot escriure com un producte de transposicions.*

Demostració. Px defineix un ordre dels elements en x. Sigui $(Px)_1 = x_{k_1}$; $P_{1k_1}x$ posa x_{k_1} a la posició desitjada, on $k_i \neq k_j, \forall i,j \mid i \neq j$. Tot seguit, sigui $(Px)_2 = (P_{1k_1}x)_{k_2}$; $P_{2k_2}P_{1k_1}x$ té els elements desitjats a les primeres dues posicions. Aplicant aquest procediment iterativament, arribem a la solució. ■

Proposició 3.5.4. *Les matrius de permutació són ortogonals:*

$$P^TP = PP^T = I. \tag{3.5.4}$$

Demostració. Sigui $P = P_a P_b$ el producte de dues matrius de permutació elementals. Sabent que $P_{rs} = P_{rs}^T = P_{rs}^{-1}$, obtenim:

$$P^T = P_b^T P_a^T P_a P_b = I. \tag{3.5.5}$$

La demostració és similar per PP^T i per més de dos factors. ■

La multiplicació PA dona una permutació de les files en A. La corresponent permutació de les columnes d'A s'obté multiplicant P^T per la dreta, AP^T. Si dividim (*partition*) la matriu en submatrius de la mateixa manera, podem redefinir el producte de dues matrius $AB, n \times n$:

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N1} & A_{N2} & \cdots & A_{NN} \end{pmatrix}, \quad B = \begin{pmatrix} B_{11} & B_{12} & \dots & B_{1N} \\ B_{21} & B_{22} & \dots & B_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ B_{N1} & B_{N2} & \cdots & B_{NN} \end{pmatrix} \tag{3.5.6}$$

$$C = AB,\ C_{ij} = \sum_{k=1}^{N} A_{ik} B_{kj}$$

sempre que aquests productes siguin compatibles (recordem que per poder multiplicar dues matrius cal que el nombre de columnes en A_{ik} ha d'equivaldre al nombre de files en B_{kj}, per això abans hem imposat que havíem de dividir la matriu de la mateixa manera).

3.5.2 Matrius inverses

Definició 3.5.5 (Transformada de Gauss)**.** Definim la transformació de Gauss L_k com una matriu triangular inferior que difereix de la matriu unitària solament en la r-èsima columna:

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & m_{k+1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_n & 0 & \cdots & 1 \end{pmatrix}, \tag{3.5.7}$$

on l'element m_i correspon al multiplicador m_{ir} a l'eliminació gaussiana. En efecte, l'eliminació gaussiana no és sinó multiplicar A amb matrius astutes per l'esquerra.

Si multipliquem un vector x per L_k obtenim:

$$y = L_k x = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} + m_{k+1} x_k \\ \vdots \\ x_n + m_n x_k \end{pmatrix} \equiv y_i = \begin{cases} x_i, & i = 1, \dots, r; \\ x_i + m_i x_k, & i = r+1, \dots, n. \end{cases} \tag{3.5.8}$$

Aleshores, $x = L_k^{-1} y$ ha de satisfer

$$x_i = \begin{cases} y_i, & i = 1, \dots, k; \\ y_i - m_i y_k, & i = k+1, \dots, n. \end{cases} \tag{3.5.9}$$

La transformada de Gauss es pot utilitzar per fer zeros en un vector i és útil per reduir-lo: sigui x un vector amb $x_k \neq 0$ i sigui L_k donat per

$$m_{jk} = -\frac{x_j}{x_k},\ j = k+1, \ldots, n. \tag{3.5.10}$$

Aleshores,

$$L_k^{-1} x = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} - m_{k+1} x_k \\ \vdots \\ x_n - m_n x_k \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.5.11}$$

En efecte, és força simple invertir la transformada de Gauss:

$$L_k^{-1} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & -m_{k+1,k} & 1 & \\ & & & \vdots & & \ddots \\ & & & -m_{n,k} & & 1 \end{pmatrix}, \tag{3.5.12}$$

Observació 3.5.6. En l'eliminació gaussiana aplicada a la matriu A, $L_1 A$ ens faria zeros en la primera columna i així successivament.

Exemple 3.5.7. La transformada de Gauss L_1 es pot partir en

$$L_1 = \begin{pmatrix} 1 & 0 \\ m & I \end{pmatrix}, \tag{3.5.13}$$

on el vector columna m conté $\{m_i\}_{i=2}^n$. Sigui

$$P = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P} \end{pmatrix}, \tag{3.5.14}$$

on $\tilde{P}$ és una permutació d'ordre $n-1$. Aleshores,

$$P L_1 P^T = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ m & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P}^T \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \tilde{P} m & \tilde{P}\tilde{P}^T \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \tilde{P} m & I \end{pmatrix}. \tag{3.5.15}$$

Doncs, l'única diferència entre L_1 i $P L_1 P^T$ és que els elements en la posició $(2,1), \ldots, (n,1)$ són permutats segons la definició de $\tilde{P}$. Tot seguit, considerem el producte $L_1 L_k$ de dues transformades de Gauss amb $k > 1$:

$$L_1 L_k = \begin{pmatrix} 1 & 0 \\ m & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{L}_k \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ m & \tilde{L}_k \end{pmatrix}. \tag{3.5.16}$$

Aquí, $\tilde{L}_k$ és la submatriu inferior dreta de dimensió $(n-1) \times (n-1)$. El càlcul mostra que la diferència entre L_k i $L_1 L_k$ és que m apareix sota la diagonal principal a la primera columna.

3.5.3 Determinants

Les transformacions que experimenta una matriu A al llarg del procés d'eliminació gaussiana amb pivotatge consisteix a fer combinacions lineals de files de manera que, llevat del signe, que pot variar en permutar files i permutar columnes, es preserva el determinant.

Proposició 3.5.8. *Si σ, τ són permutacions de n elements, es compleix $(\sigma\tau)^{-1} = \tau^{-1}\sigma^{-1}$*

Proposició 3.5.9. *Tota permutació és producte de transposicions.*

Definició 3.5.10 (Permutacions)**.** Considerem un enter $n \geq 2$ i el conjunt $C = \{1, 2, \ldots, n-1, n\}$. Una permutació d'$n$ elements és una bijecció σ de C en C. Per tant, cada enter $i \in C$ té una imatge $\sigma(i)$ per σ i tenim $\sigma(i) = \sigma(j) \implies i = j$ (és injectiva). Al seu torn, tot enter de C és $\sigma(i)$ per a un únic $i \in C$ (és exhaustiva). Denotarem per S_n el conjunt de les permutacions d'n elements.

Definició 3.5.11. Per a $A = (a^i_j)_{\substack{1\leqslant i\leqslant m\\ 1\leqslant j\leqslant n}}$, definim:

$$\det(A) = \sum_{\sigma\in S_n} \epsilon(\sigma) a_1^{\sigma(1)} a_2^{\sigma(2)} \ldots a_n^{\sigma(n)}. \tag{3.5.17}$$

Així doncs, el determinant d'una matriu $n \times n$ és una suma on cada sumand correspon a una permutació σ de n elements. El sumand correspon a una permutació σ de n elements. El sumand corresponent a la permutació σ és el producte d'un element de cada fila i columna de forma que l'índex de fila és la imatge per σ de l'índex de columna i el sigme del sumand és la signatura de σ.

Demostració. Sigui $A = (a^i_j)_{\substack{1\leqslant i\leqslant n\\ 1\leqslant j\leqslant n}}$ i suposem columnes k i ℓ, amb $k < \ell$ són iguals. Posem $t = (k, \ell)$. Com en el conjunt S_n de permutacions de n elements hi ha exactament el mateix nombre de permutacions parelles que de permutacions senars, aleshores $\sigma \longmapsto \sigma t$ defineix una bijecció d'A_n en $S_n \setminus A_n$. Per a una permutació parella fixada, el sumand corresponent és

$$a_1^{\sigma(1)} \ldots a_k^{\sigma(k)} \ldots a_\ell^{\sigma(\ell)} \ldots a_n^{\sigma(n)}, \tag{3.5.18}$$

i la signatura, $\epsilon(\sigma) = +1$. El corresponent a σt:

$$-a_1^{\sigma(t(1))} \ldots a_k^{\sigma(t(k))} \ldots a_\ell^{\sigma(t(\ell))} \ldots a_n^{\sigma(t(n))} = -a_1^{\sigma(1)} \ldots a_k^{\sigma(\ell)} \ldots a_\ell^{\sigma(k)} \ldots a_n^{\sigma(n)}, \tag{3.5.19}$$

ja que $\epsilon(\sigma t) = -1$. Ara, com les columnes k i ℓ d'A són iguals, tenim que $a_k^{\sigma(\ell)} = a_\ell^{\sigma(\ell)}$ i $a_\ell^{\sigma(k)} = a_k^{\sigma(k)}$. Per tant, el sumand corresponent a σt és igual al corresponent a σ amb el signe canviat. Tenim, doncs, que cada sumand corresponent a una permutació parella σ s'anul·la amb el sumand corresponent a la permutació senar σt i el determinant, per tant, és igual a 0. ■

Corol·lari 3.5.12. *Si una columna d'A és combinació lineal de les altres columnes, $A_j = \sum_{k\neq j} b_k A_k$, aleshores* $\det A = 0$.

Corol·lari 3.5.13. *Si a una columna d'A li sumem una combinació lineal de les altres columnes, el determinant d'A no varia.*

Corol·lari 3.5.14. *Si intercanviem les posicions de dues columnes d'A, el determinant d'A canvia de signe.*

Corol·lari 3.5.15. *Si permutem les columnes d'A, el determinant no varia si la permutació és parella i canvia de signe si la permutació és senar. En altres paraules,*

$$\det(A_{\sigma(1)}, \ldots, A_{\sigma(n)}) = \epsilon(\sigma) \det(A_1, \ldots, A_n). \tag{3.5.20}$$

Proposició 3.5.16. *Si A, B matrius quadrades $n \times n$, aleshores es compleix* $\det(AB) = \det A \det B$

Lema 3.5.17. *Sigui A una matriu $n \times n$ i siguin $A_1, \ldots, A_n$ les seves columnes. Si A_k és combinació lineal de vectors columna $C_1, \ldots, C_r$, $A_k = \sum_{\ell=1}^{r} b_\ell C_\ell$ es compleix:*

$$\begin{array}{c} \det(A_1, \ldots, A_{k-1}, \sum_{\ell=1}^{r} b_\ell C_\ell, A_{k+1}, \ldots, A_n) = \\ = \sum_{\ell=1}^{r} b_\ell \det(A_1, \ldots, A_{k-1}, C_\ell, A_{k+1}, \ldots, A_n). \end{array} \tag{3.5.21}$$

Proposició 3.5.18. *Per a qualsevol matriu quadrada A, es compleix* $\det(A^T) = \det A$.

Observació 3.5.19. Amb l'anterior proposició obtenim que fins a 3.5.15 els enunciats són igualment vàlids si canviem columnes per files.

Proposició 3.5.20. *Si A és matriu $n \times n$, es compleix:*

$$\det A = \sum_{j=1}^{n} a_j^i A_j^i, \tag{3.5.22}$$

per a qualsevol columna j d'A. En altres paraules, obtenim el determinant d'A com la suma de cada coeficient de la columna j multiplicat pel seu adjunt.

Teorema 3.5.21. *Una matriu $n \times n$ té rang n si, i només si,* $\det A \neq 0$.

Demostració. Si $\operatorname{rg} A < n$, una de les columnes d'A és combinació lineal de les altres. Així, els vectors columna d'A són linealment dependents i, a més, per 3.5.12, si una columna d'A és combinació lineal de les altres columnes, $A_j = \sum_{k \neq j} b_k A_k$, aleshores $\det A = 0$. Si $\operatorname{rg} A = n$, A és invertible i A^{-1} és la matriu inversa d'A. Al seu torn, tenim $AA^{-1} = Id$, que implica $(\det A)(\det A^{-1}) = \det Id = 1$, i per tant, $\det A \neq 0$. ■

3.5.4 Factorització LU

Sigui A una matriu $n \times n$. Suposem que es pot fer la següent descomposició $PA = LU$ on L és una matriu triangular inferior i U una matriu triangular superior i P una matriu de permutació. Alguna de les seves utilitats són les següents:

1. El sistema $Ax = b$ és equivalent a resoldre els següents sistemes en aquest ordre: $Ly = b$ i $Ux = y$.
2. Càlcul del determinant d'A: $\det A = \det L \det U = (\prod_{i=1}^{n} l_{ii})(\prod_{i=1}^{n} u_{ii})$.
3. Si A és regular el càlcul de la inversa és equivalent a resoldre n sistemes tal que $Ax = Id$. Però podem usar $A^{-1} = (LU)^{-1} = U^{-1}L^{-1}$.

Hem de demostrar que l'eliminació gaussiana amb pivotatge parcial aplicada a una matriu regular A és equivalent a la factorització $PA = LU$, on P és una matriu de permutació.

Observació 3.5.22. Cal preguntar-se per l'existència i la unicitat d'$A = LU$. Si comptem incògnites d'L i U obtenim que cada matriu presenta $\frac{n(n+1)}{2}$ incògnites.

Teorema 3.5.23 (Factorització LU)**.** *Tota matriu regular A, $n \times n$, es pot factoritzar en $PA = LU$, on P és una matriu de permutació, L una matriu triangular inferior unitària i U una matriu triangular superior.*

Demostració. Per inducció. La demostració és trivial per al cas $n = 1$. Assumim que és certa per al cas $n = N - 1$ i considerem la matriu A de mida $N \times N$. Hem de demostrar que el teorema es compleix per N. El primer pas de l'eliminació gaussiana amb pivotatge parcial es pot formular com

$$A_1 = L_1^{-1} P_1 A, \tag{3.5.23}$$

on P_1 és una matriu de permutació i L_1 és una transformació de Gauss,

$$L_1 = \begin{pmatrix} 1 & 0 \\ m & I \end{pmatrix}, \; m = \begin{pmatrix} m_{21} \\ \vdots \\ m_{n1} \end{pmatrix}. \tag{3.5.24}$$

El resultat és

$$A_1 = \begin{pmatrix} u_{11} & u_1^T \\ 0 & \tilde{A}_2 \end{pmatrix}, \tag{3.5.25}$$

on $u_1^T = (u_{12}, \ldots, u_{1n}) \in \mathbb{R}^n$ i $\tilde{A}_2$ té ordre $N-1$. La hipòtesi d'inducció ens diu que $\tilde{P}_2 \tilde{A}_2 = \tilde{L}_2 \tilde{U}_2$, on totes les matrius tenen ordre $N - 1$. Ara definim les matrius de mida $N \times N$:

$$P_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P}_2 \end{pmatrix}, \; L_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{L}_2 \end{pmatrix}, U = \begin{pmatrix} u_{11} & u_1^T \\ 0 & \tilde{U}_2 \end{pmatrix}. \tag{3.5.26}$$

Aleshores,

$$\begin{aligned} P_2 A_1 &= \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P}_2 \end{pmatrix} \begin{pmatrix} u_{11} & u_1^T \\ 0 & \tilde{U}_2 \end{pmatrix} = \begin{pmatrix} u_{11} & u_1^T \\ 0 & \tilde{P}_2 \tilde{A}_2 \end{pmatrix} \\ &= \begin{pmatrix} u_{11} & u_1^T \\ 0 & \tilde{L}_2 \tilde{U}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{L}_2 \end{pmatrix} \begin{pmatrix} u_{11} & u_1^T \\ 0 & \tilde{U}_2 \end{pmatrix} = L_2 U. \end{aligned} \tag{3.5.27}$$

Si ho combinem amb el fet que $A_1 = L_1^{-1} P_1 A$ i $P_2^T P_2 = I$ obtenim que

$$P_2 P_1 A = P_2 L_1 A_1 = P_2 L_1 P_2^T P_2 A_1 = P_2 L_1 P_2^T L_2 U. \tag{3.5.28}$$

Ara fixem que $P = P_2 P_1$. Es veu clarament que $L = P_2 L_1 P_2^T L_2$ és una matriu triangular inferior unitària amb

$$l_{:1} = \begin{pmatrix} 1 \\ \tilde{P}_2 m \end{pmatrix}, \tag{3.5.29}$$

i les columnes $l_{:2}, \ldots, l_{:n}$ equivalen a les mateixes columnes en L_2. Hem demostrat també que la matriu A de mida $N \times N$ satisfà $PA = LU$ i, per tant, el teorema queda provat. ■

Corol·lari 3.5.24. *Si fixem la seqüència de pivotatge (fixem la matriu P), aleshores els factors L, U són únics.*

Demostració. El resultat és una conseqüència directa del fet que la descomposició LU és equivalent a eliminació gaussiana aplicada a la matriu PA: si $A = L_1U_1 = L_2U_2$, U_1, U_2 no singulars:

$$L_1 = L_2U_2U_1^{-1},\ L_2^{-1}L_1 = U_2U_1^{-1} = Id. \tag{3.5.30}$$

■

Una vegada hem trobat la factorització LU, podem resoldre fàcilment el sistema $Ax = b$; és equivalent a $PAx = LUx = Pb$. Definint $y = Ux$, el sistema es pot resoldre en dos passos:

1. resoldre $Ly = Pb$,
2. resoldre $Ux = y$.

Observació 3.5.25. Com L i U són ambdós triangulars, la complexitat algorísmica és de l'ordre de $\mathcal{O}(n^2)$ (en particular, $2n^2$ flops).

Exemple 3.5.26. Volem aplicar eliminació gaussiana amb pivotatge parcial, on usarem com a pivot l'element a_{21}.

$$A = \begin{pmatrix} 0.6 & 1.52 & 3.5 \\ 2 & 4 & 1 \\ 1 & 2.8 & 1 \end{pmatrix} \tag{3.5.31}$$

Resolem:

$$\begin{pmatrix} 0.6 & 1.52 & 3.5 \\ 2 & 4 & 1 \\ 1 & 2.8 & 1 \end{pmatrix}, A_1 = \begin{pmatrix} 2 & 4 & 1 \\ 0.6 & 1.52 & 3.5 \\ 1 & 2.8 & 1 \end{pmatrix}, A_2 = \begin{pmatrix} 2 & 4 & 1 \\ 0 & 0.32 & 3.2 \\ 0 & 0.8 & 0.5 \end{pmatrix}, A_3 = \begin{pmatrix} 2 & 4 & 1 \\ 0.5 & 0.32 & 3.2 \\ 0.3 & 0.8 & 0.5 \end{pmatrix}$$

$$m_{21} = 0.5, m_{31} = 0.3. \qquad \begin{matrix} A_1 = P_{12}A \\ A_2 = L_1^{-1}P_{12}A \\ A_3 = P_{23}L_1^{-1}P_{12}A \end{matrix} \qquad \begin{matrix} P = P_{23}P_{12} \\ U = A_4 = L_2^{-1}P_{23}L_1^{-1}P_{12}A \\ A = P_{12}L_1P_{23}L_2U \end{matrix} \tag{3.5.32}$$

3.5.5 | Matrius simètriques i definides positives

En una secció anterior ja hem fet una petita introducció dels diferents tipus de matrius i ara ens toca estudiar-ne algun. Veurem que aquells sistemes d'equacions determinats per una matriu simètrica i definida positiva poden ser resolts sense pivotatge i sense pèrdua de precisió. A mode d'introducció, la següent observació:

Observació 3.5.27 (Procediment sense pivotatge)**.** Suposem que no ens cal pivotar i, per tant, no ens cal la matriu P ($A = LU$). Aleshores, el procés per obtenir U se'ns simplifica a:

$$L_{n-1}^{-1}\cdots L_1^{-1}A = U,\ m_{j1} = \frac{a_{j1}}{a_{11}} \implies L_1L_2\cdots L_{n-1}U$$

$$L_1L_2 = \begin{pmatrix} 1 & & & & & \\ m_{21} & 1 & & 0 & & \\ m_{31} & m_{32} & 1 & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ \vdots & \vdots & \vdots & 0 & \ddots & \\ m_{n1} & m_{n2} & \cdots & \cdots & \cdots & 1 \end{pmatrix} \tag{3.5.33}$$

En aquesta subsecció veurem que la factorització LU sempre es pot resoldre sense pivotatge, així com aprendre com treure partit de la simetria de la matriu de tal manera que la descomposició esdevingui simètrica i necessiti de la meitat de *flops* en el pitjor cas. Al llarg d'aquesta també considerarem A, una matriu simètrica i definida positiva.

Lema 3.5.28. *L'element més gran en magnitud d'A és positiu i està a la diagonal principal. Tots els elements que sorgeixen del procés de la descomposició LU d'A sense pivotatge són o bé més petits en magnitud o bé iguals a l'element més gran d'A. En altres paraules:*

$$a_{kk} > 0, k = 1, \ldots, n \quad i \quad \max_{i,j} |a_{ij}| = \max_{k} \{a_{kk}\}. \tag{3.5.34}$$

Demostració. Una matriu simètrica i definida positiva compleix $x^T Ax > 0, \forall x \neq 0$. Usarem e_k per denotar el vector donat com el k-èsim vector columna a la matriu unitària Id:

$$(e_k)_i = \begin{cases} 0, & i \neq k, \\ 1, & i = k. \end{cases} \tag{3.5.35}$$

Si utilitzem aquest vector per x en la condició de matriu simètrica i definida positiva, ens queda que

$$0 < e_k^T A e_k = e_k^T a_{:k} = a_{kk} > 0, \tag{3.5.36}$$

i hem provat la primera part del lema. A continuació, fem:

$$0 < (e_i + e_j)^T A (e_i + e_j) = a_{ii} + a_{ij} + a_{ji} + a_{jj} = a_{ii} + a_{jj} + 2a_{ij}. \tag{3.5.37}$$

Hem usat la simetria $a_{ij} = a_{ji}$. De manera anàloga, obtenim:

$$0 < (e_i - e_j)^T A (e_i - e_j) = a_{ii} + a_{jj} - 2a_{ij}. \tag{3.5.38}$$

Combinant les dues últimes inequacions i que $i \neq j$ veiem que

$$|a_{ij}| < \frac{1}{2}(a_{ii} + a_{jj}) \leq \max\{a_{ii}, a_{jj}\} \leq \max_{1 \leq k \leq n} a_{kk}. \tag{3.5.39}$$

■

Teorema 3.5.29. *La factorització LU, $A = LU$, de qualsevol matriu A simètrica i definida positiva es pot resoldre sense pivotatge. En particular, $A = LDL^T$, on L és una matriu triangular inferior unitària i D és una matriu diagonal amb elements positius a la diagonal.*

Demostració. Per inducció. El teorema és trivialment cert per $n = 1$: $A = (a)$ i, per tant, $x^T Ax = xax = ax^2 > 0$ si $x \in \mathbb{R} \setminus \{0\} \implies a > 0$. Per al pas d'inducció, assumim que és cert per a $n = N - 1$ i fragmentem la matriu A de mida $N \times N$:

$$A = \begin{pmatrix} a_{11} & a_1^T \\ a_1 & A_2 \end{pmatrix}, \quad a_1 = \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}. \tag{3.5.40}$$

Donat que A és simètrica i definida positiva, el lema anterior ens garanteix que $a_{11} > 0$ de manera que el farem servir com el primer pivot. La corresponent transformació de Gauss és la següent:

$$L_1 = \begin{pmatrix} 1 & 0 \\ m_1 & Id \end{pmatrix}, \quad m_1 = \frac{1}{a_{11}} a_1 \tag{3.5.41}$$

i la matriu transformada és la següent:

$$A_1 = L_1^{-1}A = \begin{pmatrix} 1 & 0 \\ -m_1 & Id \end{pmatrix}\begin{pmatrix} a_{11} & a_1^T \\ a_1 & A_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_1^T \\ 0 & \tilde{A}_2 \end{pmatrix}, \tilde{A}_2 = A_2 - m_1 a_1^T = A_2 - \frac{1}{a_{11}} a_1 a_1^T. \tag{3.5.42}$$

Podem reescriure A_1 de la següent manera:

$$A_1 = \begin{pmatrix} a_{11} & a_1^T \\ a_1 & \tilde{A}_2 \end{pmatrix} = \begin{pmatrix} a_{11} & 0 \\ 0 & \tilde{A}_2 \end{pmatrix}\begin{pmatrix} 1 & m_1^T \\ 0 & Id \end{pmatrix} \xrightarrow{(3.5.42)} A = L_1 \begin{pmatrix} a_{11} & 0 \\ 0 & \tilde{A}_2 \end{pmatrix} L_1^T. \tag{3.5.43}$$

Evidentment la matriu $\tilde{A}_2$ de mida $(n-1)\times(n-1)$ és simètrica. Sigui x un vector tal que

$$L_1^T x = \begin{pmatrix} 0 \\ y \end{pmatrix} \iff x = (L_1^T)^{-1}\begin{pmatrix} 0 \\ y \end{pmatrix} \tag{3.5.44}$$

per algun vector no nul $y \in \mathbb{R}^{n-1}$. Tal x existeix i és no nul ja que la matriu triangular superior unitària L_1^T és regular. Ens queda que

$$0 < x^T A x = y^T \tilde{A}_2 y. \tag{3.5.45}$$

Això mostra que $\tilde{A}_2$ és simètrica i definida positiva. A part, segons la hipòtesi d'inducció, té una factorització $\tilde{A}_2 = \tilde{L}_2\tilde{D}_2\tilde{L}_2^T$, on $\tilde{L}_2$ és una matriu triangular inferior unitària i $\tilde{D}_2$ és diagonal amb elements positius a la diagonal. Obtenim:

$$A = L_1\begin{pmatrix} 1 & 0 \\ 0 & \tilde{L}_2 \end{pmatrix}\begin{pmatrix} a_{11} & 0 \\ 0 & \tilde{D}_2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & \tilde{L}_2^T \end{pmatrix} L_1^T = \begin{pmatrix} 1 & 0 \\ m_1 & \tilde{L}_2 \end{pmatrix}\begin{pmatrix} a_{11} & 0 \\ 0 & \tilde{D}_2 \end{pmatrix}\begin{pmatrix} 1 & m_1^T \\ 0 & \tilde{L}_2^T \end{pmatrix}. \tag{3.5.46}$$

Aquesta és, en efecte, la factorització LDL^T. ■

Ja vam veure que la pèrdua de precisió pot ocórrer a l'eliminació gaussiana; alguns elements de la matriu esdevenien molt grans durant el procés. Demostrarem que tal creixement no es pot donar en les matrius simètriques i definides positives.

Teorema 3.5.30. *Tots els elements generats durant la factorització LDL^T estan fitats en magnitud per l'element més gran d'A.*

Demostració. Solament ens cal mostrar que l'element més gran en la matriu transformada $\tilde{A}_2$ donada per (3.5.42) és fitada per l'element més gran d'A. Ja hem vist que la matriu $\tilde{A}_2$ és simètrica i definida positiva. Aleshores,

$$\max_{ij}|\tilde{a}_{ij}| = \max_i\{\tilde{a}_{ii}\}, \tag{3.5.47}$$

i per (3.5.42) un altre cop i la definició d'a_1 obtenim que

$$\tilde{a}_{ii} = a_{ii} - \frac{1}{a_{11}}a_{i,1}^2 \leq a_{ii}, \tag{3.5.48}$$

ja que $a_{11} > 0$. Al seu torn:

$$a_{ii} \leq \max_{1\leq k\leq n}\{\tilde{a}_{kk}\} = \max_{ij}|a_{ij}|, \tag{3.5.49}$$

i la nostra demostració ha finalitzat. ■

Exemple 3.5.31. La matriu simètrica i definida positiva A mostrada a continuació té la factorització LU:

$$A = \begin{pmatrix} 4 & 6 & 2 \\ 6 & 18 & -1.5 \\ 2 & -1.5 & 4.25 \end{pmatrix} = LU = \begin{pmatrix} 1 & 0 & 0 \\ 1.5 & 1 & 0 \\ 0.5 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} 4 & 6 & 2 \\ 0 & 9 & -4.5 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.5.50}$$

i la factorització LDL^T:

$$A = LDL^T,\ D = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.5.51}$$

La simetria d'A de la factorització LDL^T implica que solament cal guardar i modificar elements a la diagonal principal i en els triangles estrictament inferior o superior d'A. Com a conseqüència, el següent corol·lari:

Corol·lari 3.5.32. *El nombre de flops involucrats en resoldre la factorització LDL^T d'una matriu $n \times n$, simètrica i definida positiva és d'aproximadament $\frac{1}{3}n^3$. Ens cal solament emmagatzemar uns $\frac{1}{2}n^2$ elements de la matriu.*

Procés 3.5.33 (Factorització de Cholesky)**.** *Els elements a la diagonal a D són positius. Per tant, la matriu*

$$D^{1/2} = \begin{pmatrix} \sqrt{d_{11}} & & \\ & \ddots & \\ & & \sqrt{d_{nn}} \end{pmatrix} \tag{3.5.52}$$

també té elements reals. Obtenim que:

$$A = LDL^T = (LD^{1/2})(D^{1/2}L^T) = C^TC,\ D = D^{1/2}D^{1/2}. \tag{3.5.53}$$

La matriu $C = D^{1/2}L^T$ és una matriu triangular superior. Aquesta versió de la factorització LDL^T s'anomena factorització de Cholesky d'A. Si la coneixem, el sistema $Ax = C^TCx = b$ es resol en els següents dos passos:

1. *resoldre $C^Ty = b$,*
2. *resoldre $Cx = y$.*

Observació 3.5.34. A tall d'observació, notar que la factorització de Cholesky necessita uns $\frac{1}{3}n^3$ flops i la complexitat en trobar la solució de cadascun dels sistemes triangulars $C^Ty = b$ i $Cx = y$ és de n^2 flops.

Observació 3.5.35 (Descomposició LU i matrius simètriques definides positives)**.** En el cas que A sigui simètrica i definida positiva, tenim que $A = LDL^T = LU$ i $U = DL^T$.

Proposició 3.5.36 (Criteri de Sylvester)**.** *Per a matrius definides positives, els elements diagonals $d_{k,k}$ són positius.*

3.6

Matrius banda

Definició 3.6.1 (Banda i amplada)**.** Una matriu és *banda* si existeixen $p, q \in \mathbb{N}$ tals que $a_{ij} = 0,\ j < i - q,\ j > i + p$. L'amplitud de la banda és el màxim nombre de no nuls en una mateixa fila de la matriu i val $w = p + q + 1$.

Amb una matriu banda solament ens cal guardar els elements que es trobin dins la banda. Això implica que per a resoldre un sistema d'equacions lineals amb tal matriu els $\mathcal{O}(n^3)$ *flops* teòricament necessaris es vegin notablement reduïts.

Definició 3.6.2 (Matriu tridiagonal)**.** És aquella matriu banda tal que $p = q = 1$ i pot ser guardada en tres vectors diagonals a, b, c tals que

$$A = \begin{pmatrix} a_1 & b_1 & & & & \\ c_2 & a_2 & b_2 & & & \\ & c_3 & a_3 & b_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & & c_n & a_n \end{pmatrix} \tag{3.6.1}$$

Observació 3.6.3. Suposem que no necessitem pivotatge i que el sistema tridiagonal és de la forma $Ax = f$, donat que A és diagonalment dominant. En tal cas, solament ens fa falta anul·lar un únic element per columna durant l'eliminació:

$$\begin{aligned} m_{k+1,k} &= c_{k+1}/a'_k, \\ a'_{k+1} &= a_{k+1} - m_{k+1,k} b_k, \\ f'_{k+1} &= f_{k+1} - m_{k+1,k} f_k, \end{aligned} \tag{3.6.2}$$

on $a'_1 = a_1$ per a $k = 1, 2, \ldots, n-1$. Sense pivotatge, necessitem de $3(n-1)$ *flops* per a la factorització, $5(n-1)$ per a les substitucions endavant i endarrere, sumant un total de $8(n-1)$ flops.

Observació 3.6.4. El pivotatge parcial destrueix fins a cert punt l'estructura de banda de la matriu. Considerem, per exemple, una matriu tridiagonal d'ordre 5:

$$\begin{pmatrix} \times & \times & & & \\ \times & \times & \times & & \\ & \times & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix}, \tag{3.6.3}$$

tal que $\times$ és un element no nul. Suposem que cal intercanviar les dues primeres files, de tal manera que ens queda:

$$\begin{pmatrix} \times & \times & \times & & \\ \times & \times & & & \\ & \times & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix} \tag{3.6.4}$$

aplicant el procés d'eliminació solament s'anul·la un element per columna, com hem comentat a 3.6.3, i obtenim:

$$\begin{pmatrix} \times & \times & \times & & \\ 0 & \times & * & & \\ & \times & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix}, \tag{3.6.5}$$

on $*$ denota un nou element no nul generat en l'eliminació quan restem a la segona fila un múltiple de la primera. En el pitjor cas, les files k i $k+1$ s'han de restar a cada pas, la qual cosa ens porta que la matriu triangular superior U és de la forma:

$$\begin{pmatrix} \times & \times & \times & & \\ & \times & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix}. \tag{3.6.6}$$

Tals elements no existiran en una matriu triangular inferior L. Generalitzant, l'eliminació gaussiana amb pivotatge en aquest tipus de matrius l'amplada de banda d'U pot créixer fins $p+q+1$. Igualment, l'amplada de banda d'L es manté en $q+1$.

Observació 3.6.5. En canvi, si fem servir eliminació gaussiana sense pivotatge els elements $*$ no es donaran. Així doncs, si A té q elements no nuls per sobre la diagonal i p per sota, les matrius L i U tindran bandes d'amplada $w_L = q+1$ i $w_U = p+1$.

Observació 3.6.6. Si la matriu de banda és simètrica i definida positiva, podem utilitzar la factorització de Cholesky; la matriu C serà una matriu de banda.

Com ja hem vist, a la factorització LU d'una matriu de banda tant L com U són bandades (tot i que les columnes en L es poden permutar).

3.7 Inversa d'una matriu

Per una matriu invertible A la solució al sistema $Ax = b$ es pot expressar com $x = A^{-1}b$, on A^{-1} és la matriu inversa. Notem que això és equivalent a resoldre els j sistemes

$$Ax_{:j} = b_{:j},\ j = 1, \dots, p; \tag{3.7.1}$$

on $x_{:j}, b_{:j}$ són vectors columna en X i B, respectivament. Ja vam veure que trobar $A = LU$ amb L, U conegudes tenia un cost de $2pn^2$ *flops*, de tal manera que computar A^{-1} ens pot portar uns $2n^3$ *flops* i no és recomanable fer-ho a no ser que sigui estrictament necessari. Tot i així, mostrem el procés necessari per fer-ho:

Procés 3.7.1 (Com calcular la matriu inversa)**.** *Podem resoldre el sistema $AX = Id$, on definim I com la matriu unitària amb vectors columna tals que*

$$(e_k)_i = \begin{cases} 0, & i \neq k, \\ 1, & i = k. \end{cases} \tag{3.7.2}$$

Suposant que hem trobat $A = LU$ sense pivotatge, aleshores $x_{:k}$, el k-èsim vector columna d'A^{-1} es troba en els dos següents passos:

1. *resoldre* $Ly_{:k} = e_{:k}$,
2. *resoldre* $Ux_{:k} = y_{:k}$.

De forma general, la solució de cadascun d'aquests dos sistemes necessita de $2n^2$ *flops, però ens estalviem operacions ja que el vector* $e_{:k}$ *sols té un element no nul: com que els primers* $k-1$ *elements en* e_k *són zero, els* k *primers elements de* $y_{:k}$ *són zero. Essencialment, els elements no nuls a* $y_{:k}$ *es troben resolent les últimes* $n-k+1$ *equacions de les últimes* $n-k+1$ *incògnites: necessitem* $(n-k+1)^2$ *flops. Així, el treball per a conèixer totes les columnes* $y_{:k}$ *és de*

$$\sum_{i=1}^{n}(n-k+1)^2 = \sum_{\nu=1}\nu^2 \approx \frac{1}{3}n^3. \tag{3.7.3}$$

Malauradament, en el sistema $Ux_{:k} = y_{:k}$ *no ens podem estalviar càlculs. Així doncs, el càlcul de la inversa d'una matriu* $n \times n$ *necessita d'uns* $2n^3$ *flops.*

3.8

Normes del vector i de la matriu

En aquesta secció veurem com la solució d'un sistema d'equacions lineals és afectada pels errors dels elements de la matrius implicades. Ens cal recordar 3.8 per com mesurar la magnitud del vector i de la matriu. Ara introduirem una definició complementària de la comentada.

Definició 3.8.1 (Normes ℓ_p)**.** Definim una norma ℓ_p com

$$||x||_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}},\ p \geq 1. \tag{3.8.1}$$

essent aquestes les més utilitzades:

$$\begin{aligned} ||x||_1 &= \sum_{1\leq i\leq n}|x_i|, \\ ||x||_2 &= (x_1^2 + \cdots + x_n^2)^{\frac{1}{2}} = \sqrt{x^T x}, \\ ||x||_\infty &= \max_{1\leq i\leq n}|x_i|, \end{aligned} \tag{3.8.2}$$

on la $||\cdot||_2$ s'anomena *norma euclidiana* i $||\cdot||_\infty$ és la *norma màxima.*

Definició 3.8.2 (Error absolut i relatiu en un vector)**.** Sigui $\overline{x}$ una aproximació a un vector x. Donada una norma $||\cdot||$ definim:

1. l'error absolut, $||\delta x|| = ||\overline{x} - x||, \delta x \in \mathbb{R}^n$,
2. l'error relatiu, $\frac{||\delta x||}{||x||} = \frac{||\overline{x}-x||}{||x||}$.

Definició 3.8.3 (Norma induïda de la matriu)**.** Sigui $||\cdot||$ una norma d'un vector. La norma induïda de la matriu és

$$||A|| = \sup_{x\neq 0}\frac{||Ax||}{||x||}, \tag{3.8.3}$$

Lema 3.8.4.

$$||A|| = \max_{||z||=1}||Az|| \iff \max_{||z||=1} = ||Ay||,\ ||y|| = 1. \tag{3.8.4}$$

Demostració. Usant $y = \frac{x}{||x||}$:

$$x \neq 0,\ y = \frac{x}{||x||} \implies \frac{||Ax||}{||x||} = ||\frac{1}{||x||}Ax|| = ||Ay|| \leq \max_{||z||=1} ||Az||. \tag{3.8.5}$$

Per tant,

$$||A|| \leq \max_{||z||=1} ||Az|| \tag{3.8.6}$$

i si $\max_{||z||=1} ||Az|| = ||Ay||,\ ||y|| = 1$, ens queda que

$$\frac{||Ay||}{||y||} = \max_{||z||=1} ||Az|| \leq ||A|| \tag{3.8.7}$$

$$||A|| = \max_{||x||=1} ||Ax||. \tag{3.8.8}$$

■

Per la definició anterior es dedueix directament que la matriu identitat compleix $||Id|| = 1$. De la mateixa manera, compleix totes les propietats de però aplicades a matrius:

Lema 3.8.5. *La norma de 3.8.3 és, efectivament, una norma matricial.*

Propietat 3.8.6 (Propietats de les normes matricials)**.**

1. $||A|| \geq 0,\ \forall A$,
2. $||A|| = 0 \iff A = 0$,
3. $||\alpha A|| = |\alpha| \cdot ||A||,\ \alpha \in \mathbb{R}$,
4. $||A + B|| \leq ||A|| + ||B||$.

Lema 3.8.7. *Sigui $||\cdot||$ una norma vectorial i la seva norma matricial induïda. Es compleix que*

$$\begin{aligned} ||Ax|| &\leq ||A|| \cdot ||x||, \\ ||AB|| &\leq ||A|| \cdot ||B||. \end{aligned} \tag{3.8.9}$$

Demostració. Per 3.8.3 sabem que $\frac{||Ax||}{||x||} \leq ||A||$ per tot $x \neq 0$. Donat que $x \neq 0 \implies ||x|| > 0$, obtenim la primera desigualtat, $||Ax|| \leq ||A|| \cdot ||x||$. La segona s'obté aplicant la primera desigualtat dues vegades sobre $||ABx||$:

$$||A(Bx)|| \leq ||A|| \cdot ||Bx|| \leq ||A|| \cdot ||B|| \cdot ||x|| \implies \frac{||Abx||}{||z||} \leq ||A|| \cdot ||B|| \implies ||AB|| \leq ||A|| \cdot ||B||. \tag{3.8.10}$$

■

Lema 3.8.8.

$$||A||_2 = \sqrt{\max_{1 \leq i \leq n} \lambda_j(A^T A)}, \tag{3.8.11}$$

on $\lambda_j(A^T A)$ és el valor propi j-èsim d'$A^T A$.

Lema 3.8.9.

$$||A||_\infty = \max_{1 \leq i \leq n} \left\{ \sum_{i=1}^{n} |a_{ij}| \right\}. \tag{3.8.12}$$

Demostració. Usant (3.8.8), considerem el producte Ax per

$$||x||_\infty = \max_{1\leq i\leq n} |x_i| = 1. \tag{3.8.13}$$

Aleshores, la i-èsima component del vector Ax es pot estimar de la següent manera:

$$|(Ax)_i| = \left|\sum_{j=1}^{n} a_{ij}x_j\right| \leq \sum_{j=1}^{n} |a_{ij}|\cdot|x_j| \leq \sum_{j=1}^{n} |a_{ij}| \leq \max_{1\leq i\leq n}\sum_{j=1}^{n} |a_{ij}|. \tag{3.8.14}$$

Això demostra que la banda dreta, $r = \max_{1\leq i\leq n}\left\{\sum_{j=1}^{n}|a_{ij}|\right\}$ és una fita superior per a $||Ax||_\infty$. Ens cal mostrar que existeix un vector $\hat{x}$ amb $||\hat{x}||_\infty = 1$ tal que $||A\hat{x}||_\infty = r$. Sigui ν un nombre de fila tal que

$$\sum_{j=1}^{n} |a_{\nu j}| = \max_{1\leq i\leq n}\left\{\sum_{j=1}^{n}|a_{ij}|\right\} = r, \tag{3.8.15}$$

i sigui $\hat{x}_j = \operatorname{sgn}(a_{\nu j}), j = 1,\ldots,n$, on

$$\operatorname{sgn} x := \begin{cases} -1 & \text{si } x < 0, \\ 0 & \text{si } x = 0, \\ 1 & \text{si } x > 0. \end{cases} \tag{3.8.16}$$

Aquest vector és unitari a la norma màxima i

$$|(A\hat{x})_\nu| = \left|\sum_{j=1}^{n} a_{\nu j}\hat{x}_j\right| = \sum_{j=1}^{n}|a_{\nu j}| = r. \tag{3.8.17}$$

Sabent que $|(A\hat{x})_\nu| \leq r, \nu = 1,\ldots,n$, ens queda que $||Ax||_\infty = r$. ■

Corol·lari 3.8.10.

$$||A||_1 = \max_{1\leq j\leq n}\left\{\sum_{i=1}^{n}|a_{ij}|\right\}. \tag{3.8.18}$$

Lema 3.8.11. *Si $||F|| < 1$, la matriu $I + F$ és regular.*

Demostració. Raonarem per reducció a l'absurd suposant que $I + F$ és no regular. Aleshores, $(I + F)x = 0$ per algun $x \neq 0$. Això implica que $||x|| = ||-Fx|| \leq ||F||\cdot||x||$, mostrant així que $||F|| \geq 1$, la qual cosa és una contradicció. ■

3.9

Cost operacional i tractament dels errors

3.9.1 Anàlisi i tractament dels errors

En aquesta secció veurem com la solució d'un sistema d'equacions lineals $Ax = b$ és afectada per les pertorbacions en elements de la matriu invertible del sistema A i la de termes independents b. En aquest cas, diem que la solució *exacta* del sistema és x. Si b és pertorbada, la solució s'expressa com

$$A(x + \delta x) = b + \delta b. \tag{3.9.1}$$

Recordem que l'error absolut $||\delta x||$ s'expressa com $||\delta x|| = ||\overline{x} - x||, \delta x \in \mathbb{R}^n$ i l'error relatiu $\frac{||\delta x||}{||x||}$. Volem estimar aquest últim i tenim en compte $Ax = b$ a ambdues bandes de la igualtat de tal manera que obtenim:

$$A(x + \delta x) = b + \delta b \iff Ax + A\delta x = b + \delta b \iff A\delta x = \delta b \iff \delta x = A^{-1}\delta b. \tag{3.9.2}$$

Traiem la norma i fem un estimat de l'error absolut fent ús de $||Ax|| \leq ||A|| \cdot ||x||$:

$$||\delta x|| = ||A^{-1}\delta b|| \leq ||A^{-1}|| \cdot ||\delta b||. \tag{3.9.3}$$

De manera anàloga per b podem fer:

$$||b|| = ||Ax|| \leq ||A|| \cdot ||x||, \tag{3.9.4}$$

que es pot reescriure, tenint en compte que les normes són positives, com

$$\frac{1}{||x||} \leq ||A|| \frac{1}{||b||}, \tag{3.9.5}$$

i veiem que

$$\frac{||\delta x||}{||x||} \leq \frac{||A^{-1}|| \cdot ||\delta b||}{||x||} \leq ||A|| \cdot ||A^{-1}|| \cdot \frac{||\delta b||}{||b||}. \tag{3.9.6}$$

Això ens mostra que l'error relatiu de la banda dreta de la igualtat es veu multiplicat per un factor $||A|| \cdot ||A^{-1}||$.

Definició 3.9.1 (Nombre de condició)**.** Per a una matriu regular A el nombre de condició és $\kappa(A) = ||A|| \cdot ||A^{-1}||$. Evidentment, depèn de la norma associada: posem $\kappa_\infty(A) = ||A||_\infty \cdot ||A^{-1}||_\infty$.

Amb la derivació veiem que el nombre de condició és una mesura de com de sensible és la solució a pertorbacions a la banda dreta de la igualtat. El teorema següent mostra que tal fet també s'aplica a matrius.

Teorema 3.9.2. *Sigui $Ax = b$ i $(A + \delta A)(x + \delta x) = b + \delta b$. Si A és regular i*

$$||A^{-1}|| \cdot ||\delta A|| = \kappa(A) \frac{||\delta A||}{||A||} = \tau < 1, \tag{3.9.7}$$

aleshores la matriu $A + \delta A$ també és regular i

$$\frac{||\overline{x} - x||}{||x||} \leq \frac{\kappa(A)}{1 - \tau} \left(\frac{||\delta b||}{||b||} + \frac{||\delta A||}{||A||} \right). \tag{3.9.8}$$

Demostració. Reescrivim la matriu pertorbada:

$$A + \delta A = A(I + F), \text{ per } F = A^{-1}\delta A, \tag{3.9.9}$$

utilitzant 3.8.7 i la hipòtesi sobre τ ens queda que:

$$||F|| = ||A^{-1}\delta A|| \leq ||A^{-1}|| \cdot ||\delta A|| = \tau < 1. \tag{3.9.10}$$

Aleshores, segons 3.8.11 la matriu $I+F$ és regular. Operant:

$$I+F = Id + A^{-1}\delta A = A^{-1}(A+\delta A) \xrightarrow{I+F \text{ regular}} A+\delta A \text{ regular} \tag{3.9.11}$$

Per tant, $(A+\delta A)^{-1} = (I+F)^{-1}A^{-1}$ existeix, demostrant que $A+\delta A$ és regular. Aplicant la hipòtesi ens queda que

$$\begin{aligned}(A+\delta A)(x+\delta x) = d+\delta b &\iff A(x+\delta x)+\delta A(x+\delta x) = d+\delta b \xrightarrow{Ax=b} A\delta x+\delta A(x+\delta x) = \delta b \\ &\iff A(\delta x) = \delta b - \delta A(x+\delta x) \iff \delta x = A^{-1}\delta b - A^{-1}\delta A(x+\delta x).\end{aligned} \tag{3.9.12}$$

Ara ens quedem amb les normes i usem les propietats amb desigualtats i la primera expressió que hem donat per τ:

$$\begin{aligned}||\delta x|| &\leq ||A^{-1}|| \cdot ||\delta b|| + ||A^{-1}|| \cdot ||\delta A|| \cdot ||x+\delta x|| \\ &\leq ||A^{-1}|| \cdot ||\delta b|| + \tau(||x|| + ||\delta x||),\end{aligned} \tag{3.9.13}$$

de tal manera que $(1-\tau)||\delta x|| \leq ||A^{-1}|| \cdot ||\delta b|| + \tau||x||$ o bé

$$\frac{||\delta x||}{||x||} \leq \frac{1}{1-\tau}\left(\frac{||A^{-1} \cdot ||\delta b||}{||x||} + \tau\right). \tag{3.9.14}$$

El primer terme en el parèntesi s'estima de manera anàloga a (3.9.3) i el teorema se segueix quan inserim la segona expressió per a τ. ■

Observació 3.9.3. Per la definició de norma d'una matriu induïda, la identitat $I = AA^{-1}$, i de les desigualtats de la norma se segueix que:

$$1 = ||I|| = \max_{||x||=1} ||Ix|| = \max_{||x|=1} ||x|| = ||AA^{-1}|| \leq ||A|| \cdot ||A^{-1}||, \tag{3.9.15}$$

mostrant clarament que $\kappa(A) \geq 1$.

Definició 3.9.4 (Matriu ben condicionada). Anàloga a la definició que vam donar de condicionament anteriorment, diem que una matriu amb un nombre de condició petit és *ben condicionada*. En canvi, una matriu mal condicionada és aquella que té un nombre de condició gran.

Notació 3.9.5.

$$\begin{aligned}\kappa_\infty(A) &= ||A||_\infty \cdot ||A^{-1}||_\infty \\ \kappa_p(A) &= ||A||_p \cdot ||A^{-1}||_p, p \geq 1\end{aligned} \tag{3.9.16}$$

Observació 3.9.6. Una matriu ortogonal ($A^TA = I$) té $\kappa_2(A) = 1$, així que és ben condicionada. El nombre de condició, doncs, és usat per obtenir un estimat de la precisió que es pot arribar a obtenir al resoldre un sistema del tipus $Ax = b$.

Corol·lari 3.9.7. *Sigui el sistema $Ax = b$ i un vector $\tilde{x}$ donat. Aleshores,*

$$\frac{||\tilde{x}-x||}{||x||} \leq \kappa(A)\frac{||r||}{||b||},\ r = b - A\tilde{x}, \tag{3.9.17}$$

on r s'anomena residual.

Demostració. Evidentment, $\tilde{x}$ és la solució al sistema pertorbat $A\tilde{x} = b - r$, i el corol·lari se segueix del teorema anterior amb $\delta A = 0, \delta b = -r$. ■

Observació 3.9.8. És fals que $||A^{-1}|| \leq \frac{1}{||A||}$, el que és cert és que

$$||A^{-1}|| \geq \frac{1}{||A||}. \tag{3.9.18}$$

Exemple 3.9.9. Considerem el sistema $Ax = b$:

$$\begin{pmatrix} 57.5 & 43.75 \\ 77 & 47 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 13.75 \\ 30 \end{pmatrix}, \tag{3.9.19}$$

i els dos vectors

$$x^{(1)} = \begin{pmatrix} 1.02 \\ -1.03 \end{pmatrix}, \; x^{(2)} = \begin{pmatrix} 1.0006 \\ -1.0004 \end{pmatrix}. \tag{3.9.20}$$

Per determinar quin de tots és més proper a x, calculem els residus. En efecte, els residus són

$$r^{(1)} = b - Ax^{(1)} = \begin{pmatrix} 0.1625 \\ -0.1300 \end{pmatrix}, \; r^{(2)} = b - Ax^{(2)} = \begin{pmatrix} -0.0170 \\ -0.0274 \end{pmatrix}. \tag{3.9.21}$$

Tots dos són petits comparat amb b i, com els elements a $r^{(2)}$ són molt més petits que els de $r^{(1)}$, esperem que $x^{(2)}$ és molt més proper a x que $x^{(1)}$. Ara ens falta mesurar en quina magnitud: el nombre de condició d'A és $\kappa_\infty(A) \approx 25.0$ i l'estimació de l'error és la següent:

$$\frac{||x^{(1)} - x||_\infty}{||x||_\infty} \leq 0.14, \; \frac{||x^{(2)} - x||_\infty}{||x||_\infty} \leq 0.023. \tag{3.9.22}$$

La solució exacta és $x = \begin{pmatrix} 1 & -1 \end{pmatrix}^T$ i els errors relatius són:

$$\frac{||x^{(1)} - x||_\infty}{||x||_\infty} = 0.03, \; \frac{||x^{(2)} - x||_\infty}{||x||_\infty} = 0.0006. \tag{3.9.23}$$

Així doncs, la conclusió que $x^{(2)}$ és la millor aproximació és certa, però l'estimació de l'error és molt poc precisa. Considerem el sistema $Bx = c$ amb

$$B = \begin{pmatrix} 95.75 & 64.375 \\ 120.2 & 79.7 \end{pmatrix}, \; c = \begin{pmatrix} 31.375 \\ 40.5 \end{pmatrix} \tag{3.9.24}$$

Aquest sistema té la mateixa solució que $Ax = b$, però un nombre de condició considerablement més gran, $\kappa_\infty(B) \approx 405$. Amb les mateixes solucions aproximades obtenim que els residuals

$$r^{(1)} = c - Bx^{(1)} = \begin{pmatrix} 0.0162 \\ -0.0130 \end{pmatrix}, \; r^{(2)} = c - Bx^{(2)} = \begin{pmatrix} -0.0317 \\ -0.0402 \end{pmatrix}, \tag{3.9.25}$$

i els errors estimats

$$\frac{||x^{(1)} - x||_\infty}{||x||_\infty} \leq 0.17, \; \frac{||x^{(2)} - x||_\infty}{||x||_\infty} \leq 0.41. \tag{3.9.26}$$

Els residuals segueixen essent petits comparats amb la banda dreta, però ara no és del tot clar que $x^{(2)}$ és la millor de les dues aproximacions.

La matriu A és no regular si, i només si, $\det(A) = 0$. En tal cas, és fàcil creure que podem mesurar si la matriu és ben condicionada o no usant el seu determinant. Res més lluny de la realitat, aquesta afirmació no és certa. A la pràctica, el nombre de condició no és computat paral·lelament a la resolució d'un sistema d'equacions lineals $Ax = b$, ja que necessitaríem de l'obtenció d'A^{-1} a un cost de $2n^3$ *flops*.

3.9.2 Errors d'arrodoniment a l'eliminació gaussiana

Sabem que qualsevol nombre real que es representa en un sistema de punt flotant ho fa amb un error relatiu intrínsec que està fitat per l'arrodoniment de la unitat μ. Tal fet es pot expressar com:

$$fl[x] = x(1+\epsilon),\ |\varepsilon| \leq \mu. \tag{3.9.27}$$

Així doncs, quan representem una matriu A i un vector b es donen errors

$$fl[a_{ij}] = a_{ij}(1+\varepsilon) = a_{ij} + \varepsilon a_{ij},\ |\varepsilon| \leq \mu, \tag{3.9.28}$$

i de manera anàloga per b. Se segueix que

$$\begin{array}{cc} fl[A] = A + \delta A & ||\delta A||_\infty \leq \varepsilon ||A||_\infty, \\ fl[b] = b + \delta b & ||\delta b||_\infty \leq \varepsilon ||b||_\infty. \end{array} \tag{3.9.29}$$

Doncs, la solució exacta del sistema $Ax = b$ es representa pel sistema pertorbat

$$(A + \delta A)\hat{x} = b + \delta b, \tag{3.9.30}$$

on $\hat{x}$ és la solució pertorbada. Suposem de moment que no hi ha errors d'arrodoniment en la solució del sistema pertorbat. En tal cas, $\hat{x}$ és la solució i per 3.9.2 tenim que:

$$\frac{||\hat{x} - x||_\infty}{||x||_\infty} \leq \frac{\kappa_\infty(A)}{1-\tau} \cdot 2\mu,\ \tau = \mu\kappa_\infty(A), \tag{3.9.31}$$

sempre que $\tau < 1$. L'error es pot estudiar amb mètodes de la subsecció anterior, 3.9.1.

Teorema 3.9.10. *Suposem que es troba la solució del sistema $(A+\delta A)\hat{x} = b + \delta b$. Siguin $\hat{L}, \hat{U}$ els factors pertorbats de la descomposició LU aplicada a PA. Així, sigui $\hat{x}$ la solució obtinguda per substitució endavant i enrere, respectivament, en els sistemes*

$$\hat{L}\hat{y} = Pb,\ \hat{U}\hat{x} = \hat{y}. \tag{3.9.32}$$

Aleshores, $\hat{x}$ és la solució del sistema pertorbat $(A + \delta A)\hat{x} = b$, on

$$||\delta A||_\infty \leq \mu(n^3 + 3n^2) g_n ||A||_\infty,\ g_n = \frac{\max_{i,j,k} |\hat{a}_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}. \tag{3.9.33}$$

Els $\hat{a}_{ij}^{(k+1)}$ són els elements que resulten del k-èsim pas de l'eliminació gaussiana. Si $\tau = \mu\kappa_\infty(A)(n^3 + 3n^2)g_n < 1$, essent μ la unitat d'arrodoniment, aleshores:

$$\frac{||\hat{x} - x||_\infty}{||x||_\infty} \leq \frac{\tau}{1-\tau}. \tag{3.9.34}$$

Els elements $\hat{a}_{ij}^{(k)}$ venen donats per (): $\hat{a}_{ij}^{(k+1)} = fl[\hat{a}_{ij}^{(k)} - \hat{m}_{ik}\hat{a}_{kj}^{(k)}]$, amb $\hat{a}_{ij}^{(1)} = a_{ij}$. El factor de creixement g_n depèn del creixement dels elements de la matriu i no pas de la mida dels multiplicadors $\hat{m}_{ik}$. Amb pivotatge parcial tenim $|\hat{m}_{ik}| \leq 1$ i, per tant, una fita superior per al creixement es veu minimitzada en cada pas.

Es pot demostrar que, a priori, si utilitzem pivotatge parcial, $g_n \leq 2^{n-1}$. És possible que hi hagi matrius on aquest creixement tingui lloc ($g_{31} = 10^9$) i sigui necessari utilitzar pivotatge complet. En tal cas, es pot demostrar que $g_n \leq 1.8n^{0.25\ln n}$ i $g_{31} \leq 34.4$). A la pràctica, però, g_n és usualment més petit o igual que 10.

Observació 3.9.11. Anteriorment havíem demostrat que si A és simètrica i definida positiva i utilitzem eliminació gaussiana sense pivotatge, aleshores $\max_{i,j} |\hat{a}_{ij}^{(k)}|$ no creix amb k i, en conseqüència, $g_n = 1$.

Exemple 3.9.12. Considerem la matriu A amb la seva respectiva inversa:

$$A = \begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix},\ A^{-1} = \frac{1}{\varepsilon - 1} \begin{pmatrix} 1 & -1 \\ -1 & \varepsilon \end{pmatrix}. \tag{3.9.35}$$

Per a un ε petit tenim que $\kappa_\infty(A) \approx 4$. Això mostra que resoldre $Ax = b$ és un problema ben condicionat. En canvi, resoldre eliminació gaussiana sense pivotatge correspon a:

$$A = LU,\ L = \begin{pmatrix} 1 & 0 \\ \frac{1}{\varepsilon} & 1 \end{pmatrix},\ U = \begin{pmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{1}{\varepsilon} \end{pmatrix}, \tag{3.9.36}$$

i el factor de creixement $g_n = |1 - \frac{1}{\varepsilon}|$, el qual és molt gran per a $|\varepsilon|$ petit, 3.9.10 ens diu que l'error de la solució serà alt. Cal notar que els factors de la descomposició LU estaran mal condicionats:

$$\kappa_\infty(L) \approx \kappa_\infty(U) \approx \frac{1}{\varepsilon^2}. \tag{3.9.37}$$

L'eliminació gaussiana sense pivotatge és un algorisme inestable. Ara ho provem amb pivotatge parcial i obtenim que

$$L = \begin{pmatrix} 1 & 0 \\ \varepsilon & 1 \end{pmatrix},\ U = \begin{pmatrix} 1 & 1 \\ 0 & 1 - \varepsilon \end{pmatrix}. \tag{3.9.38}$$

D'aquesta manera, no hi ha creixement i $g_n = 1, \kappa_\infty(L) \approx 1, \kappa_\infty(U) \approx 4$. L'eliminació gaussiana amb pivotatge parcial, per tant, és un algorisme estable.

Per tant, els algorismes estables per a l'eliminació gaussiana són:

1. si la matriu és simètrica i definida positiva, l'eliminació sense pivotatge;
2. si no ho és, amb pivotatge parcial.

A la pràctica, l'error estimat a 3.9.10 és força pessimista; suposa que en cada *flop* l'error d'arrodoniment que es produeix és maximal: $fl[a \odot b] = (a \oplus b)(1 + \varepsilon),\ |\varepsilon| = \mu,\ \mu$ la unitat d'arrodoniment. Utilitzant un algorisme estable per a l'eliminació gaussiana es dona una estimació de la fita d'error en $\hat{x}$ més acurada:

$$(A + \delta A)\hat{x} = b,\ ||\delta A||_\infty \leq \mu ||A||_\infty. \tag{3.9.39}$$

Sota aquesta assumpció, el corresponent residual és molt inferior:

$$r = b - A\hat{x},\ ||r||_\infty \lesssim \mu ||A||_\infty ||\hat{x}||_\infty. \tag{3.9.40}$$

A més, si tenim en compte que $b - A\hat{x} = A(x - \hat{x}) = r$, obtenim que $||\hat{x} - x||_\infty \leq ||A^{-1}||_\infty ||r||_\infty$. Operant i suposant que $||\hat{x}||_\infty \approx ||x||_\infty$:

$$\frac{||\hat{x} - x||_\infty}{||x||_\infty} \leq \frac{||A^{-1}||_\infty ||r||_\infty}{||x||_\infty} \lesssim \frac{\mu ||A||_\infty ||A^{-1}||_\infty ||\hat{x}||_\infty}{||x||_\infty} = \mu \cdot \kappa_\infty(A). \tag{3.9.41}$$

Proposició 3.9.13. *Si la unitat d'arrodoniment i el nombre de condició $\kappa_\infty(A)$ satisfan que $\mu \approx 10^{-d}$ i $\kappa_\infty(A) \approx 10^q$, un algorisme estable d'eliminació gaussiana produirà una solució $\hat{x}$ amb $d - q$ decimals correctes.*

IV

Interpolació polinòmica i aplicacions

4.1 Introducció

Per a una funció f, suposem que coneixem el seu valor $f_i = f(x_i)$ en $n+1$ punts diferents $x_0, \ldots, x_n$.

Definició 4.1.1 (Interpolació polinòmica)**.** És el procés de determinació d'un polinomi $P(x)$ de grau més petit o igual a n tal que

$$P(x_i) = f_i,\ i = 0, \ldots, n. \tag{4.1.1}$$

Aquest polinomi es pot fer servir per estimar el valor d'f en un punt x, on x és en l'interval format per $x_0, \ldots, x_n$. Quan f_k sigui el valor d'una funció f en x_i, parlarem d'interpolació polinòmica de la funció f en les abscisses d'interpolació x_i.

Definició 4.1.2 (Extrapolació)**.** Si x és fora de l'interval format per $x_0, \ldots, x_n$, parlem d'*extrapolació*.

Observació 4.1.3. Si solament es coneixen aproximacions dels valors f_i, no és recomanable construir una funció aproximada a través d'interpolació, sinó per mètodes d'aproximacions que ja hem vist en capítols anteriors.

Notació 4.1.4. Direm que l'interval format per $x_0, \ldots, x_n$, és a dir, l'interval format pel mínim valor i el màxim valor del conjunt és $\langle x_0, \ldots, x_n \rangle$.

La funció P buscada formarà part del conjunt de polinomis de grau més petit o igual a n; serà de la forma:

$$P(x) = \sum_{i=0}^{n} c_i x^i. \tag{4.1.2}$$

Per tal de determinar-lo, ens caldrà conèixer el valor dels $n+1$ coeficients $c_0, \ldots, c_n$. En cas que c_n sigui no nul, direm que $P(x)$ té exactament grau n.

4.2 Interpolació polinomial

Exemple 4.2.1 (Per al cas $n = 2$)**.** Tenim una funció P tal que $P = \{f_0, \ldots, f_2\}$, on cada subíndex correspon al grau del polinomi, és

$$\begin{array}{l} p(x_0) = f_0 = c_0 + c_1 x_0 + c_2 x_0^2 \\ p(x_1) = f_1 = c_0 + c_1 x_1 + c_2 x_1^2 \\ p(x_2) = f_2 = c_0 + c_1 x_2 + c_2 x_2^2 \end{array} \iff \begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \end{pmatrix}. \tag{4.2.1}$$

Notem que aquesta expressió no és final: és molt fàcil veure que el sistema té solució única (el determinant associat és el de Vandermonde) i la matriu és mal condicionada. Utilitzant $p(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1)$:

$$\begin{gathered} p(x_0) = f_0 = c_0 \\ p(x_1) = f_1 = c_0 + c_1(x_1 - x_0) \qquad \Longleftrightarrow \\ p(x_2) = f_2 = c_0 + c_1(x_1 - x_0) + c_2(x_1 - x_0)(x_2 - x_1) \\ \begin{pmatrix} 1 & 0 & 0 \\ 1 & x_1 - x_0 & 0 \\ 1 & x_2 - x_0 & (x_2 - x_0)(x_2 - x_1) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \end{pmatrix}. \end{gathered} \tag{4.2.2}$$

Ho podem generalitzar per a un conjunt de funcions $P = \{f_0, \ldots, f_n\}$:

Teorema 4.2.2 (Existència i unicitat del polinomi interpolador)**.** *Siguin $x_0, \ldots, x_n$ coeficients aleatoris i diferents. Per a valors arbitraris $f_0, \ldots, f_n$ existeix un únic polinomi P de grau més petit o igual a n que interpola.*

Demostració. Per inducció provem existència. Per a $n = 0$ podem extreure el polinomi $P_0(x) = f_0$, de grau zero i amb $P_0(x_0) = f_0$. Ara, suposem que P_k és un polinomi de grau més petit o igual a k, tal que

$$P_k(x_i) = f_i,\ i = 0, 1, \ldots, k. \tag{4.2.3}$$

Hem de demostrar que podem construir P_{k+1} de grau màxim $k + 1$ que interpola f en els punts $x_i,\ i = 0, 1, \ldots, k, k + 1$. Posem:

$$P_{k+1}(x) = P_k(x) + c(x - x_0) \cdots (x - x_k),\ P_k(x_i) = f_i. \tag{4.2.4}$$

Per a $c \neq 0$, P_{k+1} és un polinomi de grau $k + 1$, i per a qualsevol c satisfà que $P_{k+1}(x_i) = P_k(x_i) = f_i$, per a $i = 0, \ldots, k$. Com els punts $x_0, \ldots, x_{k+1}$ són diferents, podem imposar com a condició per a trobar c que se satisfà $P_{k+1}(x_{k+1}) = f_{k+1}$:

$$c = \frac{f_{k+1} - P_k(x_{k+1})}{(x_{k+1} - x_0) \cdots (x_{k+1} - x_k)}. \tag{4.2.5}$$

Així, P_{k+1} és un polinomi de grau $\leq k + 1$, que interpola f en punts $x_0, \ldots, x_{k+1}$.

La unicitat es demostra per reducció a l'absurd: siguin P i Q dos polinomis de grau màxim n, on els dos interpolen els punts $x_0, \ldots, x_n$:

$$P(x_i) = Q(x_i) = f_i,\ i = 0, 1, \ldots, n. \tag{4.2.6}$$

Això significa que $P - Q$ és un polinomi de grau $\leq n$ amb $n + 1$ zeros diferents. Pel teorema fonamental de l'àlgebra, tenim que tal polinomi és exactament zero. En conseqüència, $P = Q$. ■

Teorema 4.2.3. *Sigui f una funció amb $n + 1$ derivades contínues en l'interval format pels punts $x, x_0, \ldots, x_n$. Si P és el polinomi únic de grau $\leq n$, satisfà que:*

$$\begin{gathered} P(x_i) = f_i = f(x_i),\ i = 0, 1, \ldots, n \\ f(x) - P(x) = \frac{f^{(n+1)}(\zeta(x))}{(n+1)!}(x - x_0) \cdots (x - x_n), \end{gathered} \tag{4.2.7}$$

per a $\zeta(x)$ en l'interval $\langle x, x_0, \ldots, x_n \rangle$.

Demostració. Usem el teorema de Rolle: si una funció g és contínua a l'interval $[a, b]$, derivable a l'interval (a, b) i $g(a) = g(b)$, aleshores existeix un punt p en (a, b) tal que $g'(p) = 0$. Escollim un punt arbitrari $\hat{x} \neq x_i,\ i = 0, 1, \ldots, n$ i escrivim l'error $f(\hat{x}) - P(\hat{x})$ de la forma

$$f(\hat{x}) - P(\hat{x}) = A(\hat{x} - x_0) \cdots (\hat{x} - x_n). \tag{4.2.8}$$

Per a determinar la constant A introduïm la funció auxiliar ψ:

$$\psi(x) = f(\hat{x}) - P(\hat{x}) - A(x - x_0) \cdots (x - x_n). \tag{4.2.9}$$

Observem que $\psi(\hat{x}) = 0$ i

$$\psi(x_i) = f(x_i) - P(x_i) - A \cdot 0 = 0,\ i = 0, 1, \ldots, n. \tag{4.2.10}$$

Així, ψ té com a mínim $n + 2$ arrels diferents, $\hat{x}, x_0, \ldots, x_n$. D'acord amb les suposicions que hem fet, ψ és $n + 1$ vegades derivable contínua, i el teorema de Rolle mostra que ψ' té un zero en cada subinterval d'arrels de ψ. Així doncs, ψ' té $n+1$ arrels diferents. Anàlogament, en cada subinterval entre dos arrels consecutives de ψ', es troba una arrel de ψ'', de tal manera que ψ'' té n arrels diferents. Aplicant aquest raonament iterativament, arribem a la conclusió que $\psi^{(n+1)}$ té un zero en l'interval $\langle \hat{x}, x_0, \ldots, x_n \rangle$:

$$\psi^{(n+1)}(\zeta) = 0. \tag{4.2.11}$$

Derivant l'expressió respecte ψ $n - 1$ vegades, obtenim:

$$\psi^{(n+1)}(x) = f^{(n+1)}(x) - A(n + 1)! \implies A = \frac{f^{(n+1)}(\zeta)}{(n + 1)!}, \tag{4.2.12}$$

on ζ depèn d'$\hat{x}$. Donat que $\hat{x}$ és arbitrari, el teorema queda provat. ■

L'expressió de l'error en el teorema és fàcil de provar: el factor $(x - x_0) \cdots (x - x_n)$ assegura que l'error és zero a tots els punts d'interpolació, i que si $\forall i\ x_i$ tendeix a x_0, el terme d'error tendeix cap el residu d'ordre n en el polinomi de Taylor d'f prop d'x_0:

$$|f(x) - P(x)| \leq \frac{1}{(n + 1)!}. \tag{4.2.13}$$

4.3 Fórmula d'interpolació de Newton

Fixats els punts $x_0, \ldots, x_n$, tots diferents, el polinomi interpolador d'una funció està unívocament determinat, com hem vist a 4.2.3. Newton va proposar una formulació en què la derivació es fa usant tècniques anàlogues a la demostració de 4.2.3, on generem una seqüència de polinomis $P_0, \ldots, P_n$ amb la recursió

$$\begin{aligned} P_0(x) &= c_0 \\ P_k(x) &= P_{k-1}(x) + c_k(x - x_0) \cdots (x - x_{k-1}),\ k = 1, \ldots, n. \end{aligned} \tag{4.3.1}$$

El polinomi P_n interpola f en $x_0, \ldots, x_n$. Els coeficients de tal polinomi són donats per:

$$\begin{aligned} c_0 &= f_0, \\ c_k &= \frac{f_k - P_{k-1}(x_k)}{(x_k - x_0)\cdots(x_k - x_{k-1})},\ k = 1, \ldots, n, \end{aligned} \tag{4.3.2}$$

on utilitzarem $f_i = f(x_i)$ per alleugerir notació. Aquesta expressió recursiva esdevé més complicada a n més gran; ens interessa redefinir-la i ho fem de la següent manera:

$$\left.\begin{array}{l} P_{k-1} \text{ interpola } f \text{ en } x_0, \ldots, x_{k-1}, \\ \overset{\circ\rightarrow}{P}_{k-1} \text{ interpola } f \text{ en } x_1, \ldots, x_k, \end{array}\right\} \quad P_k(x) = P_{k-1}(x) + \frac{x - x_0}{x_k - x_0}(\overset{\circ\rightarrow}{P}_{k-1}(x) - P_{k-1}(x)). \tag{4.3.3}$$

Com P_{k-1} i $\overset{\circ\rightarrow}{P}_{k-1}$ són polinomis de grau $\leq k-1$, P_k és un polinomi de grau $\leq k$; aquests dos estan multiplicats pel terme $x - x_0$. Així:

$$\begin{aligned} &P_k(x_0) = f_0 = P_{k-1}(x_0) + 0; \\ &P_k(x_i) = f_i = P_{k-1}(x_i) + \frac{x_i - x_0}{x_k - x_0}(f_i - f_i),\ i = 1, \ldots, k-1; \\ &P_k(x_k) = f_k = P_{k-1}(x_k) + \overset{\circ\rightarrow}{P}_{k-1}(x_k) - P_{k-1}(x_k) \end{aligned} \tag{4.3.4}$$

En conseqüència, P_k és l'únic polinomi de grau $\leq k$, que interpola f en $x_0, \ldots, x_k$. El coeficient c_k depèn d'$x_0, \ldots, x_k$ i els corresponents valors d'f.

Notació 4.3.1. Anomenarem al coeficient d'x^k a $P_k(x)$: $c_k = f[x_0, \ldots, x_k]$.

Observació 4.3.2. Els coeficients en x^{k-1} en $P_{k-1}(x)$ i $\overset{\circ\rightarrow}{P}_{k-1}(x)$ són $f[x_0, \ldots, x_{k-1}]$ i $f[x_1, \ldots, x_k]$, respectivament. Obtenim:

$$c_k = f[x_0, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}. \tag{4.3.5}$$

Definició 4.3.3 (Diferència dividida)**.** El k-èsim quocient de diferències (diferència dividida) d'f respecte els punts $x_0, \ldots, x_k$ és donat per:

$$\begin{aligned} f[x_i] &= f(x_i), \\ f[x_0, \ldots, x_k] &= \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}. \end{aligned} \tag{4.3.6}$$

Observació 4.3.4. Podem construir una taula de diferències dividides de la següent manera:

x	$f(x)$	$f[\cdot,\cdot]$	$f[\cdot,\cdot,\cdot]$	$f[\cdot,\cdot,\cdot,\cdot]$	$f[\cdot,\cdot,\ldots,\cdot,\cdot]$
x_0	f_0				
		$f[x_0,x_1]$			
x_1	f_1		$f[x_0,x_1,x_2]$		
		$f[x_1,x_2]$		$f[x_0,x_1,x_2,x_3]$	
x_2	f_2		$f[x_1,x_2,x_3]$		
		$f[x_2,x_3]$			
x_3	f_3				
$\vdots$	$\vdots$				$f[x_0,x_1,x_2,x_3,\ldots,x_k]$
x_{k-3}	f_{k-3}				
		$f[x_{k-3},x_{k-2}]$			
x_{k-2}	f_{k-2}		$f[x_{k-3},x_{k-2},x_{k-1}]$		
		$f[x_{k-2},x_{k-1}]$		$f[x_{k-3},x_{k-2},x_{k-1},x_k]$	
x_{k-1}	f_{k-1}		$f[x_{k-2},x_{k-1},x_k]$		
		$f[x_{k-1},x_k]$			
x_k	f_k				

(4.3.7)

Observació 4.3.5. Com que l'ordre en què agafem els punts és irrellevant, podem mostrar que el valor de $f[x_0,\ldots,x_k]$ no canvia per molt que els permutem.

Procés 4.3.6 (Com trobar el polinomi interpolador)**.** *Hem de determinar un polinomi que interpoli els punts $x_0,x_1,x_2,\ldots,x_k$. Per tal de fer-ho, només ens cal tenir en compte que l'estructura del polinomi en qüestió serà anàloga a* (4.2.4) *i hem de construir la taula de 4.3.4 adaptada al cas pertinent. Cadascun dels coeficients de la diagonal superior seran els del polinomi.*

Procés 4.3.7 (Com computar les diferències dividides)**.** *Es computen de la següent manera:*

$$f[x_i,x_{i+1},\ldots,x_{i+j}]=\frac{f[x_{i+1},\ldots,x_{i+j}]-f[x_i,\ldots,x_{i+j-1}]}{x_{i+j}-x_i} \tag{4.3.8}$$

Teorema 4.3.8 (Polinomi interpolador de Newton)**.** *El polinomi*

$$P_n(x)=f_0+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)+\cdots+f[x_0,x_1,\ldots,x_n](x-x_0)\ldots(x-x_{n-1}), \tag{4.3.9}$$

interpola f en els punts $x_0,\ldots,x_n$, on el grau de P_n és $\leq n$ i $P_n(x_i)=f_i$, $i=0,1,\ldots,n$.

Per 4.3.1 i 4.3.8 tenim que

$$P_k(x)=P_{k-1}(x)+f[x_0,\ldots,x_k](x-x_0)\cdots(x-x_{k-1}),\ P_k(x_k)=f(x_k) \tag{4.3.10}$$

i, per tant,

$$f(x_k)-P_{k-1}(x_k)=f[x_0,\ldots,x_k](x_k-x_0)\cdots(x_k-x_{k-1}). \tag{4.3.11}$$

Exemple 4.3.9 (Polinomi interpolador de Newton). Posem un cas concret $P_3(x)$ que es generalitza fàcilment a n. Els seus coeficients són fàcils de computar a través de la norma de Horner. La idea és escriure $P_3(x)$ com:

$$P_3(x) = c_0 + (x - x_0)(c_1 + (x - x_1)(c_2 + (x - x_2)c_3)). \tag{4.3.12}$$

Això es pot computar com $P_3(x) = b_0$ donada la recursió

$$\begin{gathered} b_3 = c_3 \\ b_j = b_{j+1}(x - x_j) + c_j,\ j = 2, 1, 0. \end{gathered} \tag{4.3.13}$$

4.4 Interpolació lineal

En aquesta secció començarem estudiant la influència dels tipus d'error en la interpolació lineal; quan aproximem la funció en l'interval $[x_0, x_1]$ per la recta de punts $(x_0, f(x_0))$ i $(x_1, f(x_1))$.

$$P(x) = f_0 + \frac{x - x_0}{x_1 - x_0}(f(x_1) - f(x_0)). \tag{4.4.1}$$

Notació 4.4.1. $f_0 = f(x_0)$, $f_1 = f(x_1)$. A més:

1. R_X: error en l'argument x de la interpolació;
2. R_{XF}: errors en els valors de f_0 i f_1;
3. R_T: l'error de truncament donat per l'aproximació d'f amb una recta;
4. R_C: d'errors d'arrodoniment durant la computació.

Primerament, estudiarem l'error $R_T = f(x) - P(x),\ x_0 \leq x \leq x_1$.

Teorema 4.4.2. *Sigui f una funció derivable dues vegades en l'interval $[x_0, x_1]$, $x_1 = x_0 + h$. Sigui $P(x)$ el polinomi que interpola els punts $(x_0, f(x_0))$ i $(x_1, f(x_1))$. Per $x_0 \leq x \leq x_1$ l'error de truncament es pot estimar com:*

$$|R_T| = |f(x) - P(x)| \leq \frac{h^2}{8} \max_{x_0 \leq \zeta \leq x_1} |f''(\zeta)|. \tag{4.4.2}$$

Demostració. Tenim, per 4.2.7:

$$R_T = \frac{f''(\zeta(x))}{2!}(x - x_0)(x - x_1), \tag{4.4.3}$$

on $x_0 < \zeta(x) < x_1$. Posem $x = x_0 + uh, u \in [0, 1]$ i usem que $x_1 = x_0 + h$, de tal manera que ens queda $(x - x_0)(x - x_1) = h^2 u(u - 1)$ i:

$$R_T = \frac{f''(\zeta(x))}{2!} h^2 u(u - 1). \tag{4.4.4}$$

Calculem una fita superior de R_T:

$$|R_T| \leq \frac{h^2}{2} \cdot \max_{0 \leq u \leq 1} |u(u - 1)| \cdot \max_{x_0 \leq \zeta \leq x_1} |f''(\zeta)| = \frac{h^2}{8} \max_{x_0 \leq \zeta \leq x_1} |f''(\zeta)|, \tag{4.4.5}$$

on hem utilitzat que el màxim de $u(u - 1)$ és el punt u on la derivada de la funció val 0. ■

Considerem ara l'efecte dels errors en els valors de les funcions donats, R_{XF}. Volem demostrar que els errors són de la mateixa magnitud.

Teorema 4.4.3. *Siguin $\overline{f}_0$ i $\overline{f}_1$ valors aproximats de f_0, f_1, respectivament. L'error induït en l'interpolació lineal satisfà:*

$$|R_{XF}| \leq \varepsilon = \max\{|\overline{f}_0 - f_0|, |\overline{f}_1 - f_1|\}. \tag{4.4.6}$$

Demostració. Sigui $f_i = f(x_i)$, $i = 0, 1$ i $u = \frac{x - x_0}{x_1 - x_0}$. Veiem que els valors reals i pertorbats de l'interpolador són:

$$\begin{aligned} P(x) &= f_0 + u(f_1 - f_0) = (1-u)f_0 + uf_1, \\ \overline{P}(x) &= \overline{f}_0 + u(\overline{f}_1 - \overline{f}_0) = (1-u)\overline{f}_0 + u\overline{f}_1. \end{aligned} \tag{4.4.7}$$

Així, l'error és

$$R_{XF} = \overline{P}(x) - P(x) = (1-u)(\overline{f}_0 - f_0) + u(\overline{f}_1 - f_1). \tag{4.4.8}$$

Ara, usem ε i el fet que tant u com $1-u$ són més grans que 0. Aleshores,

$$|R_{XF}| \leq (1-u)\varepsilon + u\varepsilon = \varepsilon. \tag{4.4.9}$$

■

Pel que fa a l'error per truncament (error quan interpolem) en la interpolació en el cas general de l'aproximació del valor de la funció, tenim que la banda dreta de la igualtat és l'error de truncament per a $x = x_k$ quan f és aproximada pel polinomi interpolador P_{k-1}, amb punts d'interpolació $x_0, \ldots, x_{k-1}$. Alternativament, si f és k cops derivable, i recuperant (4.3.10), aleshores

$$f(x_k) - P_{k-1}(x_k) = \frac{f^{(k)}(\zeta)}{k!}(x_k - x_0)\cdots(x_k - x_{k-1}). \tag{4.4.10}$$

Teorema 4.4.4. *Si la funció f és k cops derivable contínua en l'interval format pels punts $x_0, \ldots, x_k$, aleshores hi ha un punt ζ en aquest interval tal que*

$$f[x_0, \ldots, x_k] = \frac{f^{(k)}(\zeta)}{k!}. \tag{4.4.11}$$

Per tant, si $f^{(k)}$ no varia molt en l'interval, es pot estimar $\frac{f^{(k)}(\zeta)}{k!}$ per la diferència dividida. Tal estimació es base en aproximar $f^{(k)}$ per la k-èsima derivada de P_k. Equivalentment, P_{k-1} s'obté a partir de P_k negligint el terme de grau més alt i tal terme es pot utilitzar per estimar l'error en l'aproximació P_{k-1}.

Proposició 4.4.5. *Quan una funció s'aproxima pel polinomi interpolador de Newton, l'error de truncament es pot estimar pel primer terme negligit.*

Observació 4.4.6. Això no deixa de ser una generalització especial del teorema del valor mitjà; vegem-ho per al cas $k = 1$:

$$f[x_0, x_1] = \frac{1}{1!}f'(\zeta) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}. \tag{4.4.12}$$

4.5 Interpolació de Lagrange

El polinomi interpolador de Lagrange és una altra manera de representar el polinomi interpolador d'una funció donada f.

Definició 4.5.1 (Polinomi interpolador de Lagrange)**.** El polinomi interpolador de Lagrange de grau $\leq n$, el qual interpola f en $x_0, \ldots, x_n$ és donat per

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x), \tag{4.5.1}$$

on L_i és el polinomi de grau n que compleix

$$L_i(x_j) = \delta_{ij} = \begin{cases} 0, & \text{si } i \neq j, \\ 1, & \text{si } i = j, \end{cases} \tag{4.5.2}$$

i és de la forma:

$$L_i(x) = \frac{(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} = \frac{\prod_{j=0, j\neq i}^{n} (x - x_j)}{\prod_{j=0, j\neq i}^{n} (x_i - x_j)}. \tag{4.5.3}$$

Observació 4.5.2. $P_n(x_i) = f(x_i)$ i el grau del polinomi és, com a màxim, n.

4.6 Interpolació d'Hermite

4.6.1 Polinomi interpolador d'Hermite no generalitzat

Definició 4.6.1 (Polinomi interpolador d'Hermite)**.** És un polinomi P_{2n+1} de grau màxim $2n+1$ tal que

$$P_{2n+1}(x_i) = f(x_i),\ P'_{2n+1}(x_i) = f'(x_i), \quad i = 0, 1, \ldots, n. \tag{4.6.1}$$

Discutirem el cas $n = 1$ per simplicitat. Denotarem aquests dos punts com x_{i-1} i $x_i = x_{i-1} + h_i$ i el polinomi de grau màxim 3 com q_i. Escrivim el polinomi de la següent forma:

$$q_i(x) = a_i + b_i u + c_i u^2 + d_i u^3,\ u = \frac{x - x_{i-1}}{h_i}. \tag{4.6.2}$$

Els coeficients han de satisfer el següent sistema d'equacions lineals:

$$\begin{gathered} q_i(x_{i-1}) = a_i = f_{i-1}, \\ q_i(x_i) = a_i + b_i + c_i + d_i = f_i, \\ h_i q'_i(x_{i-1}) = b_i = h_i f'_{i-1}, \\ h_i q'_i(x_i) = b_i + 2c_i + 3d_i = h_i f'_i, \end{gathered} \tag{4.6.3}$$

on f_j i f'_j són valors donats de f i f'. Aquest sistema té la següent solució única (per tant, és compatible determinat):

$$\begin{gathered} a_i = f_{i-1}, \\ b_i = h_i f'_{i-1}, \\ c_i = 3(f_i - f_{i-1}) - h_i(2f'_{i-1} + f'_i), \\ d_i = 2(f_{i-1} - f_i) + h_i(f'_{i-1} + f'_i). \end{gathered} \tag{4.6.4}$$

Teorema 4.6.2. *Sigui f una funció quatre vegades derivable en l'interval $[x_{i-1}, x_i]$ i sigui q_i el polinomi interpolador d'Hermite. Aleshores:*

$$\max_{x_{i-1}\leq x\leq x_i} |f(x) - q_i(x)| \leq \frac{1}{384} M_i h_i^4 + \frac{1}{4} E_i' h_i + E_i, \tag{4.6.5}$$

on

$$\begin{array}{cc} h_i = x_i - x_{i-1}, & M_i = \max_{x_{i-1}\leq x\leq x_i} |f^{(4)}(x)|, \\ E_i = \max_{j=i-1,i} |f(x_j) - f_j|, & E_i' = \max_{j=i-1,i} |f'(x_j) - f_j'|. \end{array} \tag{4.6.6}$$

Observació 4.6.3. Si $f_j = f(x_j)$, aleshores $E_i = 0$. Per analogia, si $f_j' = f'(x_j)$, aleshores $E_i' = 0$.

Exemple 4.6.4. Volem aproximar $f(x) = \sin x$ per a $x \in [1.1, 1.3]$. Tenim:

x_j	$f_j = \sin x_j$	$f_j' = \cos x_j$
1.1	0.8912	0.4536
1.3	0.9636	0.2675

Quan utilitzem aquests valors en (4.6.2), amb $h_i = 0.2$, obtenim que:

$$q_i(x) = 0.8912 + 0.09702u - 0.01789u^2 - 0.0004827u^3,\ u = \frac{x - 1.1}{0.2}. \tag{4.6.7}$$

Cal notar que l'error de la funció i la seva derivada és zero als punts d'interpolació. Aplicant 4.6.2, ens surt que:

$$\max_{1.1\leq x\leq 1.3} |\sin x - q_i(x)| \simeq 4 \cdot 10^{-6}. \tag{4.6.8}$$

És fàcil veure que $E_i = E_i' = 0$ i amb $f^{(4)}(x) = \sin x$ obtenim la fita superior $4.01 \cdot 10^{-6}$ en l'error de truncament R_T.

4.6.2 Polinomi interpolador d'Hermite generalitzat

Suposem que coneixem els valors $f(x_i)$ i $f'(x_i)$, per a tots els punts $x_0, \ldots, x_n$. Proposem un cas concret $n = 3$.

Exemple 4.6.5. Interpolem $P(x_i) = f(x_i)$ i $P'(x_i) = f'(x_i)$, per a $i = 0, 1$. Cerquem P de la forma

$$\begin{array}{l} p(x) = a + b(x - x_0) + c(x - x_0)^2 + d(x - x_0)^2(x - x_1), \\ p'(x) = b + 2c(x - x_0) + 2d(x - x_0)(x - x_1) + d(x - x_0)^2, \end{array} \implies \begin{array}{r} p(x_0) = f(x_0) = a,\ p'(x_0) = f'(x_0) = b \\ p(x_1) = a + b(x_1 - x_0) + c(x_1 - x_0)^2 = f(x_1), \\ p'(x_1) = b + 2c(x_1 - x_0) + d(x_1 - x_0)^2 = f'(x_1). \end{array} \tag{4.6.9}$$

Definició 4.6.6 (Problema d'Hermite)**.** Donats $x_i,\ 0 \leq i \leq n$ volem les següents condicions d'interpolació:

$$p^{(j)}(x_i) = a_{ij},\ 0 \leq j \leq k_i - 1,\ 0 \leq i \leq n, \tag{4.6.10}$$

on $k_i - 1$ correspon a uns valors enters k_i fixats i $a_{ij} = f^{(j)}(x_i)$. Si anomenem $m + 1 = \sum_{i=0}^{n} k_i$, m serà el grau màxim del grau del polinomi interpolador.

Teorema 4.6.7 (Teorema d'existència i unicitat del polinomi d'Hermite generalitzat)**.** *Existeix un únic polinomi P_m de grau $\leq m$ tal que satisfà* (4.6.10) *per a qualsevol dels a_{ij}.*

Demostració. Suposem la forma polinomial bàsica $P_n(x) = c_0 + c_1x + c_2x^2 + \cdots + c_mx^m$. Si suposem les condicions tindrem sistema lineal de la forma

$$A \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} a_{00} \\ \vdots \\ a_{nk_{n-1}} \end{pmatrix} \tag{4.6.11}$$

Tindrà solució única si

$$A \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \implies \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \tag{4.6.12}$$

Per tant, P té un zero a x_i de multiplicitat k_i. Com que $\sum_{i=0}^{n} k_i = m + 1 > m$, amb la qual cosa $p \equiv 0$. ■

Teorema 4.6.8 (Error en la interpolació d'Hermite)**.** *Sigui P un polinomi interpolador d'Hermite que interpola en* (4.6.10)*. Aleshores,*

$$f(x) - P(x) = \frac{f^{(m+1)}(\zeta)}{(m+1)!}\omega(x), \tag{4.6.13}$$

on $\omega(x) = (x - x_0)^{k_0} \cdots (x - x_n)^{k_n}$ i $\zeta \in \langle x, x_0, \ldots, x_n \rangle$.

Procés 4.6.9 (Càlcul del polinomi interpolador d'Hermite)**.** *Escrivim:*

$$P(x) = \sum_{j=0}^{m} f[y_0, \ldots, y_j] \prod_{i=0}^{j-1} (x - y_i), \tag{4.6.14}$$

on x_i són y_i de multiplicitat k_i.

Teorema 4.6.10. *Siguin $y_0 \leq \cdots \leq y_m$ $m + 1$ punts diferents. Aleshores:*

$$f[y_0, \ldots, y_m] = \begin{cases} \frac{f[y_1, \ldots, y_m] - f[y_0, \ldots, y_{m-1}]}{y_m - y_0}, & \text{si } y_m \neq y_0, \\ \frac{f^{(m)}(y_0)}{m!}, & \text{si } y_0 = y_m. \end{cases} \tag{4.6.15}$$

4.7 Ús de funcions spline

4.7.1 Fenomenologia de Runge

Quan una funció f és aproximada per un polinomi interpolador es dona que l'error en els punts d'interpolació resulta zero. El fenomen de Runge ens mostra, en síntesi, que no podem estar segurs d'obtenir una millor aproximació a l'augmentar el grau del polinomi interpolador.

Exemple 4.7.1 (Contraexemple de Runge). S'interpola la funció

$$f(x) = \frac{1}{1+25x^2},\ -1 \leq x \leq 1, \tag{4.7.1}$$

per un polinomi P_n de grau $\leq n$ usant una xarxa equidistant de punts:

$$x_i = -1 + \frac{2i}{n},\ i = 0, 1, \ldots, n. \tag{4.7.2}$$

Si dibuixem la gràfica per a diverses n, veurem que a mesura que aquesta creix, l'error esdevé més petit a l'aproximar-nos a $x = 0$, però quan ens acostem a $x = \pm 1$ l'error incrementa a mesura que ho fa n.

4.7.2 Interpolació spline

A la secció anterior, hem vist que la interpolació polinòmica no ens serveix per aproximar correctament funcions que canvien el seu comportament a diferents parts de l'interval. En aquest cas, és millor aproximar-les a partir de diferents polinomis de grau petit en diverses parts de l'interval.

Sigui $[a, b]$ tal interval i el subdividim en els $n+1$ punts $x_0, \ldots, x_n$ tals que $a = x_0 < x_1 < \cdots < x_n = b$. En cadascun d'aquests n intervals $[x_{i-1}, x_i]$ utilitzem un polinomi diferent.

Definició 4.7.2 (*spline*). Una funció s s'anomena *spline* de grau $2m+1$ si s està composta per polinomis de grau $2m+1$ tals que s i les $2m$ primeres derivades siguin contínues.

1. Si $m = 0$, és una *spline* lineal.
2. Si $m = 1$, parlem d'una *spline* cúbica.
3. Si $s(x_i) = f(x_i), i = 0, 1, \ldots, n$, obtenim un *spline* interpolador.

4.8 Aplicacions

4.8.1 Derivació i integració numèriques

Suposem que f és una funció coneguda en els punts $x-h, x$ i $x+h$. Volem computar una aproximació de $f'(x)$.

Definició 4.8.1 (Derivada per la dreta). Si f és aproximada per la línia recta que passa pels punts $(x, f(x))$ i $x+h, f(x+h)$:

$$D_+(h) = \frac{f(x+h) - f(x)}{h}. \tag{4.8.1}$$

Definició 4.8.2 (Derivada per l'esquerra). Si f és aproximada per la línia recta que passa pels punts $(x, f(x))$ i $(x-h, f(x-h))$:

$$D_-(h) = \frac{f(x) - f(x-h)}{h} \tag{4.8.2}$$

Es veu de manera intuïtiva que obtenim una millor aproximació d'$f'(x)$ si utilitzem imatges de la funció en punts simètrics al voltant d'x:

Definició 4.8.3 (Derivada centrada)**.** Si f és aproximada per la línia recta que passa pels punts $(x-h, f(x-h))$ i $(x+h, f(x+h))$:

$$D_0(h) = \frac{f(x) - f(x-h)}{h} \tag{4.8.3}$$

Observació 4.8.4. El significat de les derivades per la dreta, per l'esquerra i centrades s'entenen a partir de la construcció d'un polinomi interpolador:

$$\begin{aligned} p(x) = f(\overline{x}) + f[\overline{x}, \overline{x}+h](x-\overline{x}) = f(\overline{x}) + \frac{f(\overline{x}+h) - f(\overline{x})}{h}(x-\overline{x}) \\ p'(x) = \frac{f(\overline{x}+h) - f(\overline{x})}{h} = D_+(h). \end{aligned} \tag{4.8.4}$$

Actuaríem anàlogament per a derivada per la dreta i la derivada centrada.

La mateixa idea es pot aplicar per a derivades més altes. Per exemple, si volem calcular f'' i f és aproximada per un polinomi interpolador de segon grau definit per tres punts $(x-h, f(x-h)), (x, f(x))$ i $(x+h, f(x+h))$, podem fer els següents càlculs:

$$f''(\overline{x}) \simeq \frac{\dfrac{f(\overline{x}+h) - f(\overline{x})}{h} - \dfrac{f(\overline{x}) - f(\overline{x}-h)}{h}}{2h} = \frac{f(\overline{x}+h) - 2f(\overline{x}) + f(\overline{x}-h)}{2h^2} \tag{4.8.5}$$

4.8.1.1 L'error en derivades numèriques

Hi ha dos tipus d'errors predominants quan utilitzem un polinomi interpolador per a aproximar una derivada: l'error de truncament R_T i R_{XF}. Treballarem el primer.

Quan volem aproximar la derivada d'una funció per un quocient de diferències es produeix un error de truncament R_T. Podem aproximar aquest error mitjançant la seva expansió de Taylor: sigui f una funció dos cops derivable, la derivada de la qual s'aproxima mitjançant:

$$\begin{aligned} R_T = D_+(h) - f'(x) = \frac{1}{h}(f(x+h) - f(x)) - f'(x) \\ = \frac{1}{h}(f(x) + hf'(x) + \frac{1}{2}h^2 f''(\xi) - f(x)) - f'(x) = \frac{1}{2}hf''(\xi) = \mathcal{O}(h), \end{aligned} \tag{4.8.6}$$

on ξ és un punt en l'interval obert $\langle x, x+h\rangle$. Es pot raonar de manera anàloga per a $D_+(h)$ i $D_0(h)$ (ho provarem per al segon cas):

$$\begin{aligned} R_T = D_0(h) - f'(x) = \frac{1}{h}(f(x+h) - f(x-h)) - f'(x) \\ = \frac{1}{h}(f(x) + hf'(x) + \frac{1}{2}h^2 f''(\xi) - f(x) + hf'(x) - \frac{1}{2}f''(x)h^2 + \frac{1}{3!}f'''(\xi)h^3) = \mathcal{O}(h^2), \end{aligned} \tag{4.8.7}$$

Prenem la derivada central i suposem que $\overline{f}(x \pm h) = f(x \pm h)(1 + \delta_\pm),\ |\delta_\pm| \leq \varepsilon$:

$$\overline{D}_0(h) = \frac{\overline{f}(x+h) - \overline{f}(x-h)}{2h} = D_0(h) + \frac{\delta_+ f(x+h) - \delta_- f(x-h)}{2h} = D_0(h) + R_{XF}. \tag{4.8.8}$$

Obtenim la següent estimació:

$$|R_{XF}| \leq \frac{|\delta_+ f(x+h) - \delta_- f(x-h)|)}{2|h|} \leq \frac{|\delta_+ f(x+h)| + |\delta_- f(x-h)|}{2|h|} \leq \frac{2\epsilon}{2|h|} = \frac{\varepsilon}{|h|}. \tag{4.8.9}$$

Si ara volem analitzar $|\overline{D}_0(h) - f'(x)|$, fem:

$$|\overline{D}_0(h) - f'(x)| = |D_0(h) + R_{XF} - f'(x)| \leq |D_0(h) - f'(x)| + |R_{XF}| \leq |f'''(\xi)|\frac{h^2}{6} + \frac{\varepsilon}{|h|}. \quad (4.8.10)$$

El primer terme és decreixent i el segon és creixent a mesura que $h \to 0$. Busquem el mínim per aïllar el valor d'h, considerant $f'''(\xi)$ una constant i $|h| = h$ per simplificació:

$$\frac{d}{dh}\left(|f'''(\xi)|\frac{h^2}{6} + \frac{\varepsilon}{h}\right) = 0 \iff h = \left(\frac{3\varepsilon}{|f'''(\xi)|}\right)^{\frac{1}{3}}. \quad (4.8.11)$$

i veiem que és d'ordre $\varepsilon^{\frac{1}{3}}$.

4.8.2 Extrapolació

Suposem que una funció $F(h)$ es pot calcular per a diferents valors d'$h \neq 0$ i volem trobar el límit de la funció a mesura que $h \to 0$. Utilitzant la derivada central,

$$F(h) = \frac{f(x+h) - f(x-h)}{2h}, \; f'(x) = \lim_{h\to 0} F(h). \quad (4.8.12)$$

Lema 4.8.5. *Volem provar que es pot trobar una bona estimació de $F(0)$ si coneixem el comportament de la funció a mesura que $h \to 0$ i dos valors d'F amb diferents h.*

Demostració. Tenim:

$$\begin{aligned} f(x+h) &= f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \frac{1}{3!}f'''(x)h^3 + \mathcal{O}(h^4) \\ f(x-h) &= f(x) - f'(x)h + \frac{1}{2}f''(x)h^2 - \frac{1}{3!}f'''(x)h^3 + \mathcal{O}(h^4) \\ F(h) &= \frac{f(x+h) - f(x-h)}{2h} = f'(x) + \frac{1}{3!}f'''(x)h^2 + \mathcal{O}(h^4). \\ F(2h) &= f'(x) + \frac{4}{3!}f'''(x)h^2 + \mathcal{O}(h^4). \\ \frac{1}{3!}f'''(x)h^2 &= \frac{1}{3}\left(F(2h) - F(h)\right) + \mathcal{O}(h^4). \end{aligned} \quad (4.8.13)$$

Utilitzant que $F(h) = f'(x) + \frac{1}{3!}f'''(x)h^2 + \mathcal{O}(h^4)$, tenim que:

$$f'(x) = F(h) - \frac{1}{3!}f'''(x)h^2 + \mathcal{O}(h^4). \quad (4.8.14)$$

Amb això, hem aconseguit estimar l'error de truncament amb h^4 en lloc d'amb h^2 sense avaluacions en punt extres. ■

Volem generalitzar aquest procediment. Per a tal efecte, introduïm: $F_1(h) = F(h)$, $F_2(h) = F_1(h) + \frac{1}{3}(F_1(h) - F_1(2h))$. Segons la derivació, hem eliminat el terme h^2 en l'error de truncament R_T, de tal manera que:

$$F_2(h) = f'(x) + \frac{1}{4!}f^{(4)}(x)h^4 + \mathcal{O}(h^6). \quad (4.8.15)$$

Si el resultat de $F_2(h)$ és finit i conegut per a diferents valors d'h podem estimar el segon terme de la següent manera:

$$\begin{aligned} F_2(h) &= f'(x) + \frac{1}{4!}f^{(4)}(x)h^4 + \mathcal{O}(h^6) \\ F_2(2h) &= f'(x) + \frac{16}{4!}f^{(4)}(x)h^4 + \mathcal{O}(h^6) \\ \frac{1}{4!}f^{(4)}(x)h^4 &= \frac{F_2(2h) - F_2(h)}{15} + \mathcal{O}(h^6). \end{aligned} \tag{4.8.16}$$

Trobem l'expressió de $F_3(h)$ d'una manera similar:

$$f'(x) = F_3(h) + \mathcal{O}(h^6) \implies F_3(h) = F_2(h) + \frac{F_2(h) - F_2(2h)}{15}. \tag{4.8.17}$$

El principi superior per estimar $F(0)$ es pot usar de manera generalitzada, quan F s'ha calculat per a dos arguments, h i qh. És sabut que l'error de truncament R_T en F és proporcional a h^p, 2.1.14.

Teorema 4.8.6 (Extrapolació de Richardson)**.** *Si $F(h) = F(0) + ch^p + \mathcal{O}(h^r), r > p$, amb p conegut i c desconeguda, independents d'h. Aleshores:*

$$F(h) + \frac{1}{q^p - 1}(F(h) - F(qh)) = F(0) + \mathcal{O}(h^r). \tag{4.8.18}$$

Si coneixem una expansió completa de l'error de truncament, podem aplicar iterativament l'extrapolació de Richardson. Suposem:

$$F(h) = F(0) + a_1h^{p_1} + a_2h^{p_2} + \cdots \tag{4.8.19}$$

amb exponents $p_1, \ldots,$ coneguts, però $a_1, \ldots,$ desconeguts. Suposem que hem calculat F per a arguments $q^3h, \ldots, qh, h$. Posem $F_1(h) = F(h)$ i calculem:

$$F_{k+1}(h) = F_k(h) + \frac{1}{q^{p_k} - 1}(F_k(h) - F_k(qh)),\ k = 1, 2, \cdots \tag{4.8.20}$$

En aquesta extrapolació, el terme h^{p_k} es cancel·la en l'expansió, de tal manera que:

$$F_{k+1}(h) = F(0) + \overline{a}_{k+1}h^{p_{k+1}} + \overline{a}_{k+2}h^{p_{k+2}} + \cdots . \tag{4.8.21}$$

Podem organitzar els càlculs en un esquema com el que veurem a continuació. Es calcula fila a fila, i si h és suficientment petita, la diferència entre dos valors adjacents a la mateixa columna ens dona una fita superior per a l'error de truncament.

$$\begin{array}{cccccc} F_1(q^3h) & & & & \\ F_1(q^2h) & F_2(q^2h) & & & \\ F_1(qh) & F_2(qh) & F_3(qh) & & \\ F_1(h) & F_2(h) & F_3(h) & F_4(h) & \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array} \tag{4.8.22}$$

4.8.2.1 L'error en l'extrapolació

Posem $\overline{F}_1(h) = F_1(h) + \epsilon_h = F(h) + \epsilon_h$. Suposem que $|\epsilon_h| \leq \varepsilon$. Aleshores:

$$\overline{F}_2(h) = F_1(h) + \epsilon_h + \frac{F_1(h) + \epsilon_h - F_1(qh) - \epsilon_{qh}}{q^{p_1} - 1} = F_2(h) + \left(\epsilon_h \frac{\epsilon_h - \epsilon_{qh}}{q^{p_1} - 1}\right), \tag{4.8.23}$$

on hem usat (4.8.20) i on podem acotar el segon terme de la següent manera:

$$\left|\epsilon_h \frac{\epsilon_h - \epsilon_{qh}}{q^{p_1} - 1}\right| \leq \left(1 + \frac{2}{q^{p_1} - 1}\right)\varepsilon = \frac{q^{p_1} + 1}{q^{p_1} - 1}\varepsilon. \tag{4.8.24}$$

Anàlogament, podem aplicar el mateix procediment per a $F_3(h)$.

$$\overline{F}_3(h) = F_3(h) + \epsilon_h^{[3]},\ |\epsilon_h^{[3]}| \leq \frac{q^{p_1} + 1}{q^{p_1} - 1}\frac{q^{p_2} + 1}{q^{p_2} - 1}\varepsilon. \tag{4.8.25}$$

D'aquesta manera, veiem que l'efecte dels errors creix a mesura que anem extrapolant. Es pot demostrar, però, que $\forall j$ es dona $|\epsilon_h^{[j]}|$, en cas que $q = 2$ i $p_i = 2i$.

4.8.3 Integració

La integració numèrica és el càlcul de $\int_a^b f(x)dx$. Quan coneixem $f = F'$ solament en uns pocs punts discrets o bé coneixem una fórmula explícita però aquesta no és integrable. En tals casos, voldrem aproximar f per una funció que s'integri fàcilment: un polinomi. Hi ha dos maneres d'obtenir una aproximació acurada:

1. aproximar f per un polinomi interpolador p de grau molt alt,
2. aproximar f per diferents polinomis interpoladors de grau baix.

Com ja vam veure en la subsecció 4.7.2 la segona opció és preferible per aproximar f.

Definició 4.8.7 (La regla del trapezi). L'aproximació de f per la línia de punts $(x_0, f(x_0))$ i $(x_1, f(x_1))$: la integral s'aproxima, doncs, per l'àrea del trapezi:

$$\int_{x_0}^{x_1} f(x)dx = \frac{h}{2}(f(x_0) + f(x_1)) + R_T \simeq \frac{h}{2}(f(x_0) + f(x_1)). \tag{4.8.26}$$

Veiem que l'error de truncament R_T és

$$R_T = \int_{x_0}^{x_1} (f(x) - p(x)) = \int_{x_0}^{x_1} \frac{f''(\xi(x))}{2!}(x - x_0)(x - x_1)dx, \tag{4.8.27}$$

on $\xi(x)$ és un punt en l'interval $\langle x_0, x_1\rangle$. Per simplificar l'expressió fem un canvi de variable $x = x_0 + th$. Com $x_1 = x_0 + h$, obtenim:

$$R_T = h\int_0^1 \frac{f''(\xi(x_0 + th))}{2!}th(t-1)hdt = \frac{1}{2}h^3\int_0^1 f''(\xi(x_0 + th))t(t-1)dt. \tag{4.8.28}$$

Veiem que $t(t-1)$ té signe negatiu constant per a $0 \leq t \leq 1$. Per tant, f'' és contínua en $[x_0, x_1]$ i tenim:

$$R_T = \frac{1}{2}h^3 f''(\xi(x_0 + \hat{t}h))\int_0^1 t(t-1)dt = -\frac{1}{12}h^3 f''(\xi(x_0 + \hat{t}h)), \tag{4.8.29}$$

on $0 < \hat{t} < 1$ i on hem aplicat el teorema següent:

Teorema 4.8.8 (Teorema del valor mitjà del càlcul integral)**.** *Si la funció φ és contínua i la funció ϕ és contínua i no canvia de signe en l'interval $[a,b]$, aleshores existeix un punt $\hat{x}$ dins l'interval tal que*

$$\int_a^b \varphi(x)\phi(x)dx = \varphi(\hat{x})\int_a^b \phi(x)dx,\ \hat{x}\in(a,b). \tag{4.8.30}$$

L'expressió mostra que l'error de truncament és petit si, i només si, h és petit. No ens és molt útil i trobem una altra manera: podem dividir l'interval d'integració $[a,b]$ en subintervals $\{[x_{i-1},x_i]\}_{i=1}^m$, tots de longitud $h=\frac{b-a}{m}$, ja que

$$h = x_i - x_{i-1} \iff x_i = a+ih \iff x_m = b = a+mh \iff h = \frac{b-a}{h}. \tag{4.8.31}$$

La integral, doncs, és la suma de totes les integrals dels subintervals, i cadascuna d'aquestes aproximacions es pot acotar amb la regla del trapezoide:

$$\int_a^b f(x)dx = \sum_{i=1}^m \int_{x_{i-1}}^{x_i} f(x)dx = \sum_{i=1}^m \frac{h}{2}(f(x_{i-1}+f(x_i))) + R_T, \tag{4.8.32}$$

Si f'' és contínua en (a,b) i $a<\eta_i<b$:

$$\begin{gathered}\min_{1\le i\le m} f''(\eta_i) \le f''(\eta_i) \le \max_{1\le i\le m} f''(\eta_i)\\ \min_{1\le i\le n} f''(\eta_i) \le f''(\eta_i) \le \frac{1}{m}\sum_{i=1}^m f''(\eta_i) \le \max_{1\le i\le m} f''(\eta_i)\end{gathered} \tag{4.8.33}$$

i, per tant, $\exists\eta\in(a,b)$ tal que $\frac{1}{m}(\sum_{i=1}^m f''(\eta_i)) = f''(\eta)$. Per tant:

$$R_T = -\frac{h^2}{12}\frac{b-a}{m}\left(\sum_{i=1}^m f''(\eta_i)\right) = -\frac{b-a}{12}h^2 f''(\eta). \tag{4.8.34}$$

4.8.3.1 Mètode de Newton-Côtes

El mètode de Newton-Côtes es basa en substituir f per un polinomi interpolador i integrar-ho. El polinomi interpola f a

$$x_i = a+ih,\ i=0,1,\dots,n;\ h=\frac{b-a}{n}. \tag{4.8.35}$$

Això significa que $P_n(x)$ de grau $\le n$ és determinat de tal manera que $P_n(x_i)=f(x_i), i=0,1,\dots,n$. Obtenim:

$$\int_a^b f(x)dx = \int_a^b P_n(x)dx + R_T. \tag{4.8.36}$$

Per al cas $n=1$ obtenim la regla del trapezi. Per a una n arbitràriament gran:

$$R_T = \int_a^b \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)\cdots(x-x_n)dx. \tag{4.8.37}$$

Això ens demostra que si f és, de fet, un polinomi interpolador de grau màxim n, aleshores $R_T=0$. També es pot veure que la integral s'expressa com una combinació lineal de valors de la funció:

$$\int_a^b P_n(x)dx = \sum_{i=0}^n A_i f(x_i). \tag{4.8.38}$$

Per veure-ho, representem $P_n(x)$ com un polinomi interpolador de Lagrange $P_n(x) = \sum_{i=0}^{n} f(x_i)L_i(x)$, on $L_i(x)$ és un polinomi de grau n. Se segueix d'aquest fet que $A_i = \int_a^b L_i(x)dx$. Posem

$$\int_a^b f(x)dx = \sum_{i=0}^{n} f(x)dx = \sum_{i=0}^{n} A_i f(x_i) + R_T. \tag{4.8.39}$$

i requerim que $R_T = 0$ per a $f(x) = p_k(x)$, $k = 0, 1, \ldots, n$, on p_k és un polinomi de grau k. Això porta a un sistema d'equacions lineals.

4.8.3.2 Simpson

Teorema 4.8.9 (Regla de Simpson)**.** *Es compleix que*

$$\begin{aligned} \int_a^b f(x)dx &= S(h) + R_T, \\ S(h) &= \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{m-2} + 4f_{m-1} + f_m), \end{aligned} \tag{4.8.40}$$

on m és un nombre natural parell, $h = \frac{b-a}{m}$ i $f_j = f(a + jh)$. Si $f^{(4)}$ és contínua, es dona que

$$R_T = -\frac{b-a}{180}h^4 f^{(4)}(\eta), a < \eta < b. \tag{4.8.41}$$

Demostració. Proposarem un exemple, el generalitzarem per al cas n i obtindrem el resultat desitjat.

Exemple 4.8.10 (Regla de Simpson per a $n = 2$)**.** Posant $n = 2$ tenim que

$$\int_{x_0}^{x_2} f(x)dx = A_0 f(x_0) + A_1 f(x_1) + A_2 f(x_2) + R_T, \tag{4.8.42}$$

amb $x_0 = a$, $x_1 = a + h$ i $x_2 = b = a + 2h$. Per a obtenir un sistema d'equacions lineals no molt complex utilitzarem els polinomis $p_k(x) = (x - x_1)^k$, de tal manera que la sortida serà:

$$\begin{gathered} \textstyle\int_{x_0}^{x_2} 1dx = 2h = A_0 + A_1 + A_2, \\ \textstyle\int_{x_0}^{x_2} (x - x_1)dx = 0 = -hA_0 + hA_2 \\ \textstyle\int_{x_0}^{x_2} (x - x_1)^2 dx = \frac{2}{3}h^3 = h^2 A_0 + h^2 A_2. \end{gathered} \tag{4.8.43}$$

El sistema té la solució $A_0 = A_2 = \frac{h}{3}$ i $A_1 = \frac{4h}{3}$.

Es pot demostrar que segueix sent $R_T = 0$ en el cas $n = 3$. Per al cas d'un polinomi de quart grau, extraiem la regla de Simpson:

$$\begin{aligned} \int_{x_0}^{x_2} f(x)dx &= \frac{h}{3}(f(x_0) + 4f(x_1) + f(x_2)) + R_T, \\ R_T &= -\frac{1}{90}h^5 f^{(4)}(\eta),\ x_0 < \eta < x_2. \end{aligned} \tag{4.8.44}$$

Observació 4.8.11. Es pot demostrar que l'error de truncament per a un polinomi de grau quatre no és zero, sinó que

$$R_T = -\frac{1}{90}h^5 f^{(4)}(\eta), x_0 < \eta < x_2. \tag{4.8.45}$$

Es pot demostrar, també, que existeixen I_n, aproximacions d'$I = \int_a^b f(x)$, tals que calculades amb el mètode de Newton-Côtes no convergeixen a I a mesura que $n \to \infty$.

Per tant, en lloc d'un polinomi interpolador de grau alt en $[a, b]$ subdividirem l'interval d'integració i usarem un polinomi interpolador de grau baix a cada subinterval. Si el nombre de subintervals és parell $m = 2q$, podem usar (4.8.44) en cadascun dels subintervals $[x_0, x_2], [x_2, x_4], \ldots, [x_{m-2}, x_m]$ i obtenim:

$$\begin{aligned} \int_a^b f(x)dx &= \sum_{i=1}^{\frac{m}{2}} \int_{x_{2i-2}}^{x_{2i}} f(x)dx \\ &= \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{m-2} + 4f_{m-1} + f_m) + R_T, \end{aligned} \tag{4.8.46}$$

on

$$R_T = -\frac{h^5}{90}\sum_{i=1}^{\frac{m}{2}} f^{(4)}(\eta_i),\ x_{2i-2} < \eta_i < x_{2i}. \tag{4.8.47}$$

Suposant que $f^{(4)}$, procedim amb la regla del trapezi i obtenim el resultat desitjat:

$$R_T = -\frac{h^4}{180}\frac{\frac{b-a}{2}}{\frac{m}{2}}\sum_{i=1}^{\frac{m}{2}} f^{(4)}(\eta_i) \tag{4.8.48}$$

■

4.8.3.3 **Mètode de Romberg**

Anteriorment, hem aproximat $I = \int_a^b f(x)dx$ per

$$T(h) = h(\frac{1}{2}f_0 + f_1 + \cdots + f_{m-1} + \frac{1}{2}f_m), \tag{4.8.49}$$

on $h = \frac{b-a}{m}$ i $f_j = f(x_j) = f(a + jh),\ j = 0, 1, \ldots, m$. Hem demostrat que l'error és $\mathcal{O}(h^2)$ si f'' és contínua. Podem, de fet, mostrar que

Teorema 4.8.12 (Fórmula d'Euler-Maclaurin)**.** *Si f és $(2k+2)$ vegades derivable, $f \in Ck^{2k+2}$, aleshores es compleix que:*

$$T(h) = \int_a^b f(x)dx + a_1h^2 + a_kh^{2k} + \mathcal{O}(h^{2k+2}), \tag{4.8.50}$$

on els coeficients $a_1, \ldots, a_k$ són independents d'h.

Procés 4.8.13 (Mètode de Romberg). *Quan 4.8.12 se satisfà, podem usar l'extrapolació de Richardson: posem $T_1(h) = T(h)$ i suposem que hem calculat $T_1(2h), T_1(4h), \ldots$. En tal cas, els valors calculats usant que*

$$T_{r+1}(h) = T_r(h) + \frac{T_r(h) - T_r(2h)}{2^{2r} - 1},\ r = 1, 2, \ldots, k \tag{4.8.51}$$

satisfan

$$T_r(h) = I + R_T, \quad I = \int_a^b f(x)dx,\ R_T = \begin{cases} \mathcal{O}(h^{2r}),\ r \leq k+1, \\ \mathcal{O}(h^{2k+2}),\ r > k+1. \end{cases} \tag{4.8.52}$$

Recordem que $S(h) = \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{m-2} + 4f_{m-1} + f_m)$, (4.8.44), i veiem que $T_2(h) = S(h)$:

$$\begin{aligned} T(h) &= h(\frac{1}{2}f_0 + f_1 + f_2 + \cdots + f_{m-2} + f_{m-1} + \frac{1}{2}f_m), \\ T(2h) &= 2h(\frac{1}{2}f_0 + f_2 + \cdots + f_{m-2} + \frac{1}{2}f_m), \end{aligned} \tag{4.8.53}$$

i $T_2(h)$ és equivalent a la fórmula de Simpson:

$$T_2(h) = T(h) + \frac{T(h) - T(2h)}{3} = \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{m-2} + 4f_{m-1} + f_m) = S(h). \tag{4.8.54}$$

Per a $h = \frac{b-a}{2}$, $T_2(h)$ és equivalent a la fórmula de Newton-Côtes per a $n = 2$.

$$\begin{array}{c|c|c|c|c} h & T(h) & & & \\ & & \frac{4T(\frac{h}{2}) - T(h)}{3} = T_2(h) & & \\ \frac{h}{2} & T(\frac{h}{2}) & & \frac{16T_2(\frac{h}{2}) - T_2(h)}{15} & \\ & & \frac{4T(\frac{h}{4}) - T(\frac{h}{2})}{3} = T_2(\frac{h}{2}) & & \frac{64T_3(\frac{h}{2}) - T_3(h)}{63} \\ \frac{h}{4} & T(\frac{h}{4}) & & \frac{16T_2(\frac{h}{4}) - T_2(\frac{h}{2})}{15} & \\ & & \frac{4T(\frac{h}{8}) - T(\frac{h}{4})}{3} = T_2(\frac{h}{4}) & & \\ \frac{h}{8} & T(\frac{h}{8}) & & & \end{array} \tag{4.8.55}$$

Per $r > 2$ no hi ha fórmula de Newton-Côtes, la qual és equivalent a $T_r(h)$. Es pot demostrar que, a mesura que $h \to 0$, $T_r(h)$ convergeix a I. Com a l'extrapolació de Richardson, l'error de truncament de cada $T_r(h)$ està fitat per $|T_r(h) - T_r(2h)|$,

$$|R_T| \leq |T_r(h) - T_r(2h)| = |-a_1h^2 - 15a_2h^4 - \cdots|. \tag{4.8.56}$$

De fet, $\frac{R_T}{3} = -a_1h^2 - 5a_2h^4 - \cdots$ i la cota de l'error de $T_r(h)$ depèn directament de $T_{r+1}(h)$. Pel que fa R_{XF},

Proposició 4.8.14. *Si els valors de la funció no tenen errors absoluts més grans que ε, podem acotar l'error en les entrades R_{XF} per*

$$|R_{XF}| \leq (b-a)\varepsilon. \tag{4.8.57}$$

Demostració. Tenim que per a $f_j = f(x_j)$ es dona l'error $fl(f_j) = \overline{f_j}$. Així doncs, per a aquests valors aproximats:

$$\overline{T}(h) = h(\frac{1}{2}\overline{f}_0 + \overline{f}_1 + \overline{f}_2 + \cdots + \overline{f}_{m-2} + \overline{f}_{m-1} + \frac{1}{2}\overline{f}_m) \tag{4.8.58}$$

Suposem que $|\overline{f}_j - f_j| \leq \varepsilon,\ j = 0, \ldots, m$. En altres paraules, si els valors de la funció no tenen errors absoluts més grans que ε, podem acotar l'error en les entrades R_{XF} de la següent manera:

$$|R_C| \leq |\overline{T}(h) - T(h)| \leq h(\sum_{i\in\{0,m\}} \frac{1}{2}|\overline{f}_i - f_i| + \sum_{j=1}^{m-1} |\overline{f}_j - f_j|) \leq h\varepsilon(1 + (m-1)) = hm\varepsilon = (b-a)\varepsilon. \tag{4.8.59}$$

■

Observació 4.8.15. A diferència del procediment de l'extrapolació de Richardson, es pot demostrar que l'efecte dels errors en els valors de la funció no augmenta durant les extrapolacions successives en el mètode de Romberg.

V

Resolució d'equacions en una variable amb mètodes iteratius

5.1

Introducció als mètodes iteratius

En aquest capítol trobarem mètodes par a calcular $f(x) = 0$, on $f : \mathbb{R} \longrightarrow \mathbb{R}$. Normalment, la solució a aquesta equació no es pot expressar de manera analítica: són un exemple les equacions transcendents, com ara les exponencials o les trigonomètriques. En canvi, per a $f(x)$ quadràtica sí que es pot expressar analíticament fins a grau 4, i per a graus superiors es pot demostrar que no és possible en general.

Definició 5.1.1 (Zero)**.** Aquells valors $z \in \mathbb{R}$ tal que $f(z) = 0$ s'anomenen zeros de la funció.

Sigui x^* una arrel d'una funció no lineal f. Podem escriure $f(x) = (x - x^*)^q g(x)$, amb $g(x^*) \neq 0$. L'exponent q és la *multiplicitat de l'arrel*. Se segueix que

$$f'(x) = q(x - x^*)^{q-1} g(x) + (x - x^*)^q g'(x), \tag{5.1.1}$$

i si $q > 1$, aleshores $f'(x^*) = 0$. A no ser que s'indiqui altrament, suposarem que x^* és una arrel simple amb $q = 1$ (en aquest cas, $f'(x^*) \neq 0$.

Procés 5.1.2. *Bàsicament, el procés de càlcul d'arrels consta de tres etapes:*

1. *localització: es busquen regions que continguin els zeros,*
2. *separació: es vol una regió que només contingui un zero,*
3. *aproximació: usant mètodes iteratius, que generen successions $(x_k)_{k=1}^{\infty}$ que convergeixen al zero x^*, $\lim_{k\to\infty} x_k = x^*$.*

Definició 5.1.3 (Mètode iteratiu)**.** És aquell mètode que genera una successió $(x_k)_{k=1}^{\infty}$ que compleix certes condicions de convergència, com ara que convergeixi en un punt en concret i ho faci de la manera més ràpida possible.

5.2

Mètode de la bisecció

Quan coneixem un interval que conté una arrel, el podem anar acotant successivament a través del mètode de la bisecció. Suposem $[a, b]$ un interval amb $f(a)f(b) < 0$ i m el punt intermig, $m = \frac{a+b}{2}$.

Definició 5.2.1 (Mètode de la bisecció)**.** Si $f(m)f(b) > 0$, aleshores l'arrel es troba en el subinterval $[a, m]$; altrament, es troba en $[m, b]$. Aquest procés de divisió es pot repetir successivament fins arribar a la solució.

Observació 5.2.2. És evident que amb aquest mètode, les successions de l'aproximació sempre convergeixen, però la velocitat a què ho fan és terriblement lenta: al k-èsim pas l'interval té una longitud $\frac{b-a}{2^k}$ (excepte en el cas que el punt mig de l0'interval és, en efecte, una arrel d'f). Per tant, aquest mètode no refinat s'hauria d'evitar.

5.3 Mètode de Newton-Raphson

Com més informació d'f usem, obtindrem una convergència més ràpida. Necessitarem que f sigui derivable una vegada. Primer donarem la definició de Newton-Raphson i a continuació n'explicarem la deducció:

Definició 5.3.1 (Mètode de Newton-Raphson).

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},\ k = 0, 1, \dots \tag{5.3.1}$$

Demostració. Suposem una aproximació de l'arrel x_0. Extraiem la recta tangent $f(x) - f(x_0)$ a f en el punt $(x_0, f(x_0))$ i definim x_1, la següent aproximació a l'arrel, com la intersecció entre $f(x) - f(x_0)$ i l'eix d'abscisses:

$$f(x) - f(x_0) = f'(x_0)(x - x_0) \iff x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}, \tag{5.3.2}$$

i posant $f(x) = 0$ obtenim la intersecció amb l'eix d'abscisses. Podem procedir iterativament amb aquest procediment i queda generalitzat per la successió del mètode de Newton-Raphson. ■

En lloc d'usar derivació geomètrica podem fer-ho des d'un punt de vista més analític i propi d'aquesta assignatura: per a una aproximació x_k a l'arrel donada, busquem h tal que $f(x_k + h) = 0$, suficientment petita; així, obtindrem una aproximació força bona per als dos primers termes de l'expansió de Taylor al voltant d'x_k:

$$f(x_k + h) = f(x_k) + f'(x_k)h + \mathcal{O}(h^2) = 0 \iff h \simeq \frac{f(x_k + h) - f(x_k)}{f'(x_k)},\ f'(x_k) \neq 0. \tag{5.3.3}$$

Clarament, aquesta és simplement una aproximació d'h, $h = x^* - x_k$, així que la denotem per h_k i posem:

$$h_k = -\frac{f(x_k)}{f'(x_k)}, \tag{5.3.4}$$

i ens queda $x_{k+1} = x_k + h_k$ com la següent aproximació a l'arrel.

Observació 5.3.2. En aquest mètode usem valors tant de la funció com de la seva derivada. Com que usem més informació d'f, la convergència serà, *a priori*, més ràpida.

5.4 Mètode de la secant

Considerem l'equació $f(x) = 0$, amb f derivable. Suposem dos aproximacions de l'arrel, x_0, x_1 i podem aproximar la funció per la recta secant que passa pels punts $(x_0, f(x_0))$ i $(x_1, f(x_1))$ i prenem la intersecció entre la secant i l'eix d'abscisses com la següent aproximació x_2. La secant té l'equació:

$$f(x) - f(x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0), \tag{5.4.1}$$

i posant $f(x) = 0$ i obtenim la intersecció amb l'eix d'abscisses,

$$x_2 = x_1 - f(x_1)\frac{x_1 - x_0}{f(x_1) - f(x_0)}. \tag{5.4.2}$$

Aquest procediment es pot generalitzar, de tal manera que obtenim el mètode de la secant:

Definició 5.4.1 (Mètode de la secant)**.**

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})},\ k = 1, 2, \ldots \tag{5.4.3}$$

Observació 5.4.2. El mètode de la secant també pot obtenir-se del mètode de Newton-Raphson usant l'aproximació:

$$f'(x_k) \simeq \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}. \tag{5.4.4}$$

5.5 Ordre de convergència

Donada la successió $\{x_k\}_{k=0}^{\infty}$ que convergeix a x^*, considerem, equivalentment, que $\{x_k - x^*\}_{k=0}^{\infty}$ convergeix a 0. Prenem una aplicació ϕ tal que $x_{k+1} = \varphi(x_k)$, amb el punt fix $\varphi(x^*) = x^*$. D'aquesta manera, podem escriure:

$$x_k - x^* = \varphi(x_{k-1}) - \varphi(x^*), \tag{5.5.1}$$

i mitjançant el teorema del valor mitjà obtenim que:

$$x_k - x^* = \varphi'(\xi_k)(x_{k-1} - x^*), \tag{5.5.2}$$

on ξ_k es troba entre x_{k-1} i x^*. Si

$$|\varphi'(\xi_k)| \leq m < 1 \tag{5.5.3}$$

per alguna constant $m \in \mathbb{R}$, aleshores:

$$|x_k - x^*| \leq m|x_{k-1} - x^*| < |x_{k-1} - x^*|. \tag{5.5.4}$$

La condició $|\varphi'(x)| \leq m < 1$ proper a x^* és una condició suficient per convergència, ja que:

$$|x_k - x^*| \leq m|x_{k-1} - x^*| \leq \cdots \leq m^k|x_0 - x^*|, \tag{5.5.5}$$

i $m^k \to 0$ per $k \to \infty$. Formalment, podem proposar el teorema del punt fix següent:

Teorema 5.5.1 (Teorema del punt fix)**.** *Posem $\mathcal{I}$ un interval al voltant d'x^*, tal que:*

$$\mathcal{I} = \{x \in \mathbb{R} \mid |x - x^*| \leq \delta\}. \tag{5.5.6}$$

Suposem que la funció φ té un punt fix x^* i que $|\varphi'(x)| \leq m < 1$, $\forall x \in \mathcal{I}$. Si $x_0 \in \mathcal{I}$, aleshores:

1. $x_k \in \mathcal{I}$, $k = 1, 2, \ldots,$
2. $\lim_{k\to\infty} x_k = x^*,$
3. *x^* és la única arrel en $\mathcal{I}$ de l'equació $x = \varphi(x)$.*

Bibliografia

[Ric81] John Rischard Rice. *Matrix computations and mathematical software*. Inf. tèc. 1981.

[ABD91] Anton Aubanell, Antoni Benseny i Amadeu Delshams. "Eines bàsiques de càlcul numèric". A: *Universitat Autònoma de Barcelona, Bellaterra* (1991).

[Bur02] Richard L. Burden. *Análisis numérico / Richard L. Burden, J. Douglas Faires*. spa. México [etc, 2002.

[EWN04] Lars Elden, Linde Wittmeyer-Koch i Hans Bruun Nielsen. "Introduction to Numerical Computation-analysis and MATLAB illustrations". A: (2004).

Índex terminològic