Introducció a la computació científica
Grau Enginyeria Informàtica

Semestre de Tardor

A. Benseny (Problemes)
Xavier Jarque (Teoria)
J.C. Tatjer, J. Timoneda (Pràctiques)

Departament de Matemàtica Aplicada i Anàlisi
Facultat de Matemàtiques
Universitat de Barcelona

Objectiu

L'objectiu de l'assignatura és proporcionar uns coneixements bàsics i d'utilitat de Càlcul Numèric, eines fonamentals per la modelització científica i per l'anàlisi de dades experimentals.

Les qüestions relatives a la implementació i programació d'alguns dels algorismes numèrics s'aborden a les classes pràctiques de l'assignatura.

Temari

Càlcul Numèric

1. Errors
2. Àlgebra lineal numèrica.
3. Interpolació polinomial
4. Derivació i integració numèrica
5. Zeros de funcions
6. Resolució numèrica d'equacions diferencials ordinàries

Metodologia

Les classes teòrico–pràctiques es dediquen a:

- Exposar alguns algorismes elementals de Càlcul Numèric.
- Resoldre numèricament alguns problemes de caire científic, i que sovint són de difícil o impossible solució analítica.
- Estudiar alguns dels principis teòrics d'aquests algorismes, especialment de l'Anàlisi Matemàtica.
- Veure algunes aplicacions provinents de la modelització.

Metodologia

És molt racomanable que els alumnes disposin de calculadora científica a les classes, i que portin les transparències del curs, les llistes de problemes, etc.

El material de suport (transparències de teoria, problemes, exàmens de cursos anteriors, etc.) es van publicant i actualitzant durant el curs al Campus Virtual.

El desenvolupament del curs inclou per a cada setmana del curs el següent:

- Una hora de classe a l'aula d'ordinadors per treballar els dubtes que sorgeixin de les pràctiques que us proposem.
- Una hora de problemes on es resolen i es discuteixen els problemes de les llistes que es proposen.
- Dues hores de teoria on farem ús de les transparències i la pissarra per introduir els conceptes que treballem durant el curs.

Avaluació

L'avaluació: Tots els detalls de l'avaluació es poden trobar al pla docent.

Les dates de les proves parcials seran dins de les dates disenyades per la facultat a aquest efecte.

Bibliografia

- A. Aubanell, A. Benseny, A. Delshams. *Eines bàsiques de Càlcul Numèric*, Manuals UAB.
- J.D. Faires, R. Burden. *Métodos numéricos*, 3era edición. International Thomson Paraninfo, 2004.

Errors

Un exemple: La massa de la Terra

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

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

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

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

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

Nota $M = 5.9736 \cdot 10^{24} \;\; kg$ (Wikipedia, NASA).
$M = 5.9742 \cdot 10^{24} \;\; kg$ (J.M.A. Danby, *Fundamentals of Celestial Mechanics*, Willmann-Bell, Inc., 1992).

Fonts d'error

- Errors de modelització: els models matemàtics són aproximacions de la realitat.
- Errors de truncament: d'aproximacions numèriques del model matemàtic (algorismes).
- Errors experimentals: deguts a mesures amb una exactitud limitada.
  - Errors aleatoris: les mesures estan afectades per una gran quantitat de factors "aleatoris" (llei normal).
  - Errors sistemàtics: deguts, per exemple, a una cal·libració incorrecta de l'aparell de mesura.
  - Errors aberrants: deguts a errors humans, canvis sobtats en les condicions de l'experiment, etc.
- Errors d'arrodoniment: les operacions descrites per l'algorisme es fan amb un nombre finit de dígits (amb l'ajuda d'una calculadora o un ordinador).

Definició d'errors absolut i relatiu: Sigui x el valor exacte d'una quantitat i $\overline{x}$ el seu valor aproximat.

- Error absolut de x: $e_a = |\overline{x} - x|$; $\quad (e_a := e_a(\overline{x}, x))$
- Error relatiu de x:

$$e_r = \frac{e_a}{x} = \frac{|\overline{x} - x|}{|x|} \quad (e_r := e_r(\overline{x}, x))$$

Fites d'error:

- ε_a és una fita de l'error absolut de x si $e_a \leq \varepsilon_a$;
- ε_r és una fita de l'error relatiu de x si $e_r \leq \varepsilon_r$.

Notació usual:

- $x = \overline{x} \pm \varepsilon_a \iff x \in [\overline{x} - \varepsilon_a, \overline{x} + \varepsilon_a]$
- $x = \overline{x}\,(1 \pm \varepsilon_r) \iff e_r \in [-\varepsilon_r, \varepsilon_r]$

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

$$e_a(\overline{x}, x) := e_a(\sqrt{2}) = 0.0002135...$$

i

$$e_r(\overline{x}, x) := e_r(\sqrt{2}) = \frac{0.0002135...}{\sqrt{2}} \simeq \frac{0.0002135}{1.414} = 0.00015099...$$

En particular tenim que

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

Sigui $b \geq 2$ enter. La representació en base b d'un número real $x \neq 0$ és

$$
\begin{aligned}
x &= \pm a_{q-1} a_{q-2} \ldots a_0 . a_{-1} \ldots_{b)} \\
&= \pm (a_{q-1} b^{q-1} + a_{q-2} b^{q-2} + \ldots + a_0 + a_{-1} b^{-1} + \ldots),
\end{aligned}
$$

on els a_j són nombres naturals amb $0 \leq a_j < b$.

Observació: Mentre els éssers humans treballem, generalment, en base $b = 10$, els ordinadors ho fan en base $b = 2$. Si no fem explícita la base s'enten que $b = 10$.

Exemples:

$$107.125 = (107.125_{10)}) = 1 \cdot 10^2 + 0 \cdot 10^1 + 7 \cdot 10^0 + 1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 5 \cdot 10^{-3}$$

$$0.333\ldots = 3 \cdot 10^{-1} + 3 \cdot 10^{-2} + 3 \cdot 10^{-3} \ldots$$

$$\pi = 3 \cdot 10^0 + 1 \cdot 10^{-1} + 4 \cdot 10^{-2} + 1 \cdot 10^{-3} + 5 \cdot 10^{-4} + \ldots$$

$$0.1_{2)} = 0.5$$

$$0.1_{10)} = 0.000\overline{1100}_{2)}$$

Escrivim $0.1_{10)}$ en base 2.

$$\begin{aligned} 0.1 \times 2 &= 0.2 \quad \text{(posem 0)} \\ 0.2 \times 2 &= 0.4 \quad \text{(posem 0)} \\ 0.4 \times 2 &= 0.8 \quad \text{(posem 0)} \\ 0.8 \times 2 &= 1.6 \quad \text{(posem 1)} \\ 0.6 \times 2 &= 1.2 \quad \text{(posem 1)} \\ 0.2 \times 2 &= 0.4 \quad \text{(posem 0)} \\ 0.4 \times 2 &= 0.8 \quad \text{(posem 0)} \\ &\dots \end{aligned}$$

Per tant

$$0.1_{10)} = 0.000\overline{1100}_{2)}$$

Escrivim $125.1_{10)}$ en base 2. És a dir volem trobar els valors de $a_j,\ j \in \mathbb{Z}$ tals que

$$125.1_{10)} = a_{q-1}a_{q-2}\ldots a_0.a_{-1}a_{-2}a_{-3}\ldots\ {}_{2)}.$$

Primer escriurem $125_{10)}$ en base 2.

$$\begin{aligned}
125/2 &= 62 \quad (\text{resta } 1)\\
62/2 &= 31 \quad (\text{resta } 0)\\
31/2 &= 15 \quad (\text{resta } 1)\\
15/2 &= 7 \quad (\text{resta } 1)\\
7/2 &= 3 \quad (\text{resta } 1)\\
3/2 &= 1 \quad (\text{resta } 1)\\
1/2 &= 0 \quad (\text{resta } 1)
\end{aligned}$$

Per tant

$$125_{10)} = 1111101_{2)}.$$

Com ja hem calculat abans:

$$0.1_{10)} = 0.000\overline{1100}_{2)}$$

Finalment

$$125.1_{10)} = 1111101.000\overline{1100}_{2)} = 0.1111101000\overline{1100}_{2)} \times 2^7$$

Sigui $x \neq 0$. Llavors normalitzem el número x de la forma següent:

$$x = \pm 0.\alpha_1\alpha_2\ldots\alpha_t\alpha_{t+1}\ldots \times 10^q = \pm m \times 10^q \quad \alpha_j \in \{0, 1, \ldots 9\}$$

on podem suposar que $\alpha_1 \neq 0$. Direm que

- q és l'exponent, un nombre enter;
- m és la mantissa, un nombre real positiu tal que $0.1 \leq m < 1$.

En base $b = 2$ la mantisa és una tira de 0's i 1's

Exemples:

- $g = 9.80665 = 0.980665 \times 10^1$, l'exponent és 1 i la mantissa és 0.980665;
- $G = 6.67428 \times 10^{-11} = 0.667428 \times 10^{-10}$, l'exponent és -10 i la mantissa és 0.667428.

Problema: Els ordinadors poden emmagatzemar una quantitat finita de dígits. Per tant no podem repressentar totes les mantises (ni tots els exponents).

Solució: Si sols podem guardar t xifres significatives (a la mantisa), el número (normalitzat)

$$x = \pm(0.\alpha_1\alpha_2\ldots\alpha_t\alpha_{t+1}\ldots) \times 10^q = \pm m \times 10^q, \text{ amb } \alpha_1 \neq 0,$$

l'arrodonim pel número $\mathrm{fl}_t(x)$ (flotant de x), calculat usant:

- Si $\alpha_{t+1} < 5$, $\mathrm{fl}_t(x) = \pm(0.\alpha_1\alpha_2\ldots\alpha_t)\times 10^q$,
- Si $\alpha_{t+1} \geq 5$, $\mathrm{fl}_t(x) = \pm\mathrm{fl}_t\left((0.\alpha_1\alpha_2\ldots\alpha_t + 10^{-t})\times 10^q\right)$.

Exemple: Sigui $x = 0.999527 \times 10^1$. Llavors:

$$\mathrm{fl}_5(x) = 0.99953\times 10^1, \quad \mathrm{fl}_4(x) = 0.9995\times 10^1, \quad \mathrm{fl}_3(x) = 0.100\times 10^2$$

Format	base (b)	digits (t)	q_{min}	q_{max}	bits
IEEE simple	2	24	-125	128	32
IEEE doble	2	53	-1021	1024	64

Taula: Formats IEEE (simple i doble precisió)

IEEE simple	signe (1)	$e = q + 126$ (8)	mantisa (23)
IEEE doble	signe (1)	$e = q + 1022$ (11)	mantisa (52)

Taula: Distribució de memòria pel format IEEE (simple i doble).

- La mantisa té $t - 1$ dígits ja que suposem que el primer dígit és 1 i no es guarda (números normalitzats).
- Per exemple, en precisió simple, els valors $e = 0$ i $e = 255$ es reserven a NaN (Not a Number) i overflow, respectivament.

IEEE simple	signe (1)	$e = q + 126$ (8)	mantisa (23)
IEEE doble	signe (1)	$e = q + 1022$ (11)	mantisa (52)

Taula: Distribució de memòria pel format IEEE (simple i doble).

- En precisió doble, com que $2^{-53} \approx 1.1 \times 10^{-16}$, podem garantir una precisió de 16 decimals.
- Si un número real x es pot escriure exactament fent servir la precisió de l'ordinador, diem que és un número màquina. Altrament tenim una representació en punt flotant $\mathrm{fl}_t(x)$.

IEEE simple	signe (1)	$e = q + 126$ (8)	mantisa (23)
IEEE doble	signe (1)	$e = q + 1022$ (11)	mantisa (52)

Taula: Distribució de memòria pel format IEEE (simple i doble).

- Considerem precisió simple. Com s'escriu $x = 125.1$?

Sabem que

$$x = 125.1 = 0.1111101000\overline{1100}_{2)} \times 2^7$$

I ara tenim que $e = 7 + 126 = 133 = 10000101_{2)}$. Per tant

IEEE simple	0	10000101	11110100011001100110011

Observem que tal com hem definit les coses tenim que

$$e_a(\mathrm{fl}_t(x), x) = |\mathrm{fl}_t(x) - x| \leq \frac{1}{2} \times 10^{-t} \times 10^q = \frac{1}{2} 10^{q-t} =: \varepsilon_a(\mathrm{fl}_t(x), x)$$

Així $\varepsilon_a = \frac{1}{2} 10^{q-t}$ és una fita de l'error absolut de la representació amb punt flotant amb t xifres i arrodoniment de qualsevol nombre real x.

Si $x \neq 0$, llavors podem escriure (recordem que $|m| \geq 10^{-1}$)

$$e_r(\mathrm{fl}_t(x), x) = \frac{e_a(\mathrm{fl}_t(x), x)}{|x|} \leq \frac{1}{2} \frac{10^{q-t}}{m \times 10^q} \leq \frac{1}{2} 10^{1-t} =: \varepsilon_r(\mathrm{fl}_t(x), x)$$

i per tant $\varepsilon_r(\mathrm{fl}_t(x), x) = \frac{1}{2} \times 10^{1-t}$ és una fita de l'error relatiu de la representació amb punt flotant amb t xifres i arrodoniment de qualsevol nombre real $x \neq 0$. O també que t és el número de xifres significatives.

Exemples:

- $g = 0.980665 \times 10^1$: $t = 6$, $q = 1$.

$$\varepsilon_a = \frac{1}{2}10^{1-6} = \frac{1}{2}10^{-5}, \qquad \varepsilon_r = \frac{1}{2}10^{1-6} = \frac{1}{2}10^{-5}.$$

- $G = 0.667428 \times 10^{-10}$: $t = 6$, $q = -10$.

$$\varepsilon_a = \frac{1}{2}10^{-10-6} = \frac{1}{2}10^{-16}, \qquad \varepsilon_r = \frac{1}{2}10^{1-6} = \frac{1}{2}10^{-5}.$$

- De fet $\epsilon = \frac{1}{2} b^{1-t}$ s'anomena épsilon de la màquina o precisió de la màquina i coincideix amb el número positiu més petit que sumat a 1 dóna diferent de 1, és a dir

$$\epsilon = \min\{\varepsilon : \mathrm{fl}_t(1+\varepsilon) \neq 1\}.$$

- Notem que el fet que $\mathrm{fl}_t(1+\varepsilon) = 1$ no vol dir que per l'ordinador el número ε sigui igual a 0.
- Si treballem amb $t = 3$ tenim que si

$$x = 0.1 \times 10^1 \quad \mathrm{i} \quad y = 0.456 \times 10^{-4}$$

llavors $x + y = x$, però $y \neq 0$.

Els ordinadors, degut a la representació dels nombres en punt flotant, calculen de manera aproximada. Això té implicacions importants com per exemple: l'ordre de les operacions afecta al resultat final.

Exemple: Prenem $t = 4$. Volem calcular $a + b + c$, on

$$a = 0.5317 \times 10^{-2},\ b = 0.3387 \times 10^{2},\ c = -0.3381 \times 10^{2}.$$

són números màquina. Aleshores,

$$\mathrm{fl}_4(a + \mathrm{fl}_4(b + c)) = \mathrm{fl}_4(0.5317 \cdot 10^{-2} + 0.6000 \cdot 10^{-1}) = 0.6532 \cdot 10^{-1}$$

$$\mathrm{fl}_4(\mathrm{fl}_4(a + b) + c) = \mathrm{fl}_4(0.3388 \cdot 10^{2} - 0.3381 \cdot 10^{2}) = 0.7000 \cdot 10^{-1}\ .$$

El resultat exacte és $a + b + c = 0.65317 \times 10^{-1}$.

Exemple: Les dues solucions de l'equació $x^2 - 18x + 1 = 0$ són

$$x_{1,2} = 9 \pm \sqrt{80} = \begin{cases} x_1 = 0.17944271909999916 \times 10^2 \\ x_2 = 0.5572809000084121 \times 10^{-1} \end{cases}$$

Si prenem $\sqrt{80} = 8.9443$ (és a dir $t = 5$) sóbté

$$x_1 = 9 + 8.9443 = 17.9443 = 0.179443 \cdot 10^2 \text{ (6 xifres),}$$

$$x_2 = 9 - 8.9443 = 0.0557 = 0.557 \cdot 10^{-1} \text{ (3 xifres!)).}$$

Problema: Al calcular x_2 hi ha una cancel·lació de xifres decimals, perquè restem dues quantitats que són properes (i una és significativament errònia).

Hi ha dos raons (o almenys així ho podem pensar) responsables de la propagació de l'error quan iniciem un procés que implicarà una quantitat significativa de càlculs:

- Degut a les operacions

  Per exemple: tenim dos números màquina x, y i quan els sumem tenim $\mathrm{fl}_t(x+y)$ (és l'operació la que s'equivoca).

- Degut a les dades.

  Per exemple: calculem $\sin(x)$ però de fet estem calculant $\sin(\mathrm{fl}_t(x))$ (l'error és en la dada tot i que el sinus estigui perfectament calculat).

- Les dues alhora...

Per simplificar suposarem que les funcions que apliquem en el procés de càlcul no cometen errors i que per tant tots els errors propagats són deguts a les dades.

Teorema (del valor mig): Sigui $f : [a, b] \to \mathbb{R}$ una funció contínua, i derivable a $]a, b[$. Aleshores existeix un punt $\xi \in (a, b)$, tal que

$$f(b) - f(a) = f'(\xi)(b - a).$$

Aplicació: Fórmula de propagació d'errors (una variable) Sigui $x \in \mathbb{R}$ i sigui $\overline{x} \approx x$.

Del Teorema anterior deduim que

$$e_a(f(\overline{x}), f(x)) := |f(\overline{x}) - f(x)| = |f'(\xi)||\overline{x} - x|, \quad \xi \in < \overline{x}, x > .$$

Per tant també podem escriure

$$e_a(f(\overline{x}), f(x)) \approx |f'(\overline{x})| e_a(\overline{x}, x),$$
$$\varepsilon_a(f(\overline{x}), f(x)) := M \cdot \varepsilon_a(\overline{x}, x), \quad \text{on} \quad M = \max_{\xi \in [\overline{x} - \varepsilon_a, \overline{x} + \varepsilon_a]} |f'(\xi)|.$$

De les darreres expressions tenim que

$$e_r\left(f(\overline{x}), f(x)\right) \approx |\overline{x}| \frac{|f'(\overline{x})|}{|f(\overline{x})|} |e_r\left(\overline{x}, x\right)|, \;\; (f(x) \neq 0).$$

De fet, el terme $\varphi(x) = |x| \dfrac{|f'(x)|}{|f(x)|}$ s'anomena coeficient de propagació (de l'error relatiu) i és el que volem controlar en un procés de càlcul. Si podem acotar-lo, és a dir, si podem dir que

$$|x| \frac{|f'(x)|}{|f(x)|} \leq M$$

per alguna $M > 0$, en un entorn de $\overline{x}$, llavors

$$\varepsilon_r\left(f(\overline{x}), f(x)\right) := M \times \varepsilon_r\left(\overline{x}, x\right).$$

Exemple 1: Tornem a l'exemple anterior on teniem l'equació $x^2 - 18x + 1 = 0$ i les seves arrels $x_{1,2} = 9 \pm \sqrt{80}$. Haviem trobat $\overline{x_1} = 17.9443$, i per tant $x_1 = 17.9443 \pm \frac{1}{2} \times 10^{2-6=-4}$. És conegut que

$$x_2 = f(x_1) = \frac{1}{x_1} \quad \mapsto \quad \left(\overline{x_2} = f(\overline{x_1}) = \frac{1}{\overline{x_1}}\right)$$

Solució: Tenim

$$e_a(\overline{x_2}, x_2) = e_a\left(\frac{1}{\overline{x_1}}, \frac{1}{x_1}\right) \simeq \left|\frac{-1}{x_1^2}\right| e_a(\overline{x_1}, x_1) \simeq \frac{1}{17.9443^2}\frac{1}{2} \times 10^{-4} \simeq 0.16 \times 10^{-6}$$

$$e_r(\overline{x_2}, x_2) = \frac{e_a(\overline{x_2}, x_2)}{x_2} = 17.9443 \times 0.16 \times 10^{-6} \simeq 0.29 \cdot 10^{-5}$$

Així $e_r(\overline{x_2}, x_2) \leq 5 \times 10^{-5}$ garanteix que $\overline{x_2}$ té 5 xifres significatives.

Exemple 2: Tenim $f(x) = \log \cos^2(x)$, i volem acotar l'error comès en avaluar-la en un punt x del qual només coneixem tres xifres correctes $\overline{x} = 0.735$.

Solució:

- $e_a(\overline{x}, x) \leq \varepsilon_a = 1/2 \times 10^{-3}$.
- $f'(x) = -2\tan(x)$.
- Acotem $|f'(\xi)|$ per $\xi \in [0.7345, 0.7355]$ (i.e, $[\overline{x} - \varepsilon_a, \overline{x} + \varepsilon_a]$). Com que la tangent és creixent i positiva a tot l'interval $]0, \pi/2 \simeq 1.5708[$, llavors $|\tan(\xi)| \leq \tan(0.7355) \lesssim 0.905$, i

$$|f'(\xi)| \leq 2 \times 0.905 = 1.810 \ .$$

- Aplicant la fórmula de propagació de l'error:

$$|\varepsilon_a(f(\overline{x} = 0.735), f(x))| = 1.810\ (0.5 \times 10^{-3}) = 0.905 \times 10^{-3}.$$

Pel que fa al coeficient de propagació φ de l'error relatiu tenim que

$$\varphi(x) = x\frac{f'(x)}{f(x)} = x\frac{-2\tan(x)}{\log\cos^2(x)}.$$

Treballant s'obté que $|\varphi(x)| \leq 2.2$ si $x \approx 0.735$ i per tant

$$\varepsilon_r(f(0.735)) \leq 2.2 \times \varepsilon_r(\overline{x}, x) \leq 2.2 \times \frac{0.5 \times 10^{-3}}{0.735} \approx 1.5 \times 10^{-3}.$$

Teorema (del valor mig en varies variables). Sigui G un obert de $\mathbb{R}^n$, i $f : G \to \mathbb{R}$ una funció diferenciable sobre G. Siguin $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ dos punts de $G \subset \mathbb{R}^n$ tals que el segment que els uneix està contingut a G. Aleshores existeix un punt ξ d'aquest segment tal que

$$f(y) - f(x) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\xi)(y_i - x_i).$$

Aplicació: Fórmula de propagació d'errors (varies variables) Sigui $x \in \mathbb{G} \subset \mathbb{R}^n$, sigui $\overline{x} \approx x$.

$$e_a(f(\overline{x}), f(x)) \approx \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(\overline{x}) \right| e_a(\overline{x_i}, x_i)$$

$$\varepsilon_a(f(\overline{x}), f(x)) := \sum_{i=1}^{n} M_i \times \varepsilon_a^i(\overline{x_i}, x_i), \quad \text{on} \quad M_i = \max_{\xi \in [\overline{x} - \varepsilon_a, \overline{x} + \varepsilon_a]^n} \left| \frac{\partial f}{\partial x_i}(\xi) \right|$$

Corol·lari 1: $f(x,y) = x + y$

$$\varepsilon_a(\overline{x}+\overline{y}, x+y) = \varepsilon_a(\overline{x}, x) + \varepsilon_a(\overline{y}, y)$$
$$\varepsilon_r(\overline{x}+\overline{y}, x+y) = \left|\frac{x}{x+y}\right| \varepsilon_r(\overline{x}, x) + \left|\frac{y}{x+y}\right| \varepsilon_r(\overline{y}, y)$$

Corol·lari 2: $f(x,y) = xy$

$$\varepsilon_a(\overline{xy}, xy) \approx |y|\, \varepsilon_a(\overline{x}, x) + |x|\, \varepsilon_a(\overline{y}, y)$$
$$\varepsilon_r(\overline{xy}, xy) \approx \varepsilon_r(\overline{x}, x) + \varepsilon_r(\overline{y}, y)$$

Corol·lari 3: $f(x,y) = x/y$

$$\varepsilon_a(\overline{x}/\overline{y}, x/y) \approx 1/|y|\, \varepsilon_a(\overline{x}, x) + |\frac{x}{y^2}|\, \varepsilon_a(\overline{y}, y)$$
$$\varepsilon_r(\overline{x}/\overline{y}, x/y) \approx \varepsilon_r(\overline{x}, x) + \varepsilon_r(\overline{y}, y)$$

Exemple: Volem calcular la massa de la Terra, via $M = \dfrac{gR^2}{G}$, i estimar l'error propagat a partir de les dades no exactes

$$g = 9.80665 \; , G = 6.67428 \cdot 10^{-11} \; , R = 6371.0 \cdot 10^3$$

Solució:

- Suposem que les dades són correctes fins l' última xifra significativa, de forma que (ull amb la notació!):

$$\varepsilon_a(g) = \frac{1}{2}10^{-5} \; , \varepsilon_a(G) = \frac{1}{2}10^{-16} \; , \varepsilon_a(R) = \frac{1}{2}10^{2} \; .$$

- De la fórmula de propagació d'errors en vàries variables:

$$\varepsilon_a(M) \approx \left|\frac{\partial M}{\partial g}\right| \varepsilon_a(g) + \left|\frac{\partial M}{\partial G}\right| \varepsilon_a(G) + \left|\frac{\partial M}{\partial R}\right| \varepsilon_a(R),$$

on, en el nostre cas:

$$\frac{\partial M}{\partial g} = \frac{R^2}{G} \; , \frac{\partial M}{\partial G} = -\frac{gR^2}{G^2} \; , \frac{\partial M}{\partial R} = \frac{2gR}{G} .$$

- Avaluem les derivades parcials de M en les dades:

$$\frac{\partial M}{\partial g} \approx 6.0815 \times 10^{23} \, ,$$
$$\frac{\partial M}{\partial G} \approx -8.9357 \times 10^{34} \, ,$$
$$\frac{\partial M}{\partial R} \approx 1.8722 \times 10^{18} \, .$$

- Finalment, substituint en la fórmula de propagació de l'error:

$$\varepsilon_a(M) \approx 1.0112 \times 10^{20}.$$

Per tant, aproximadament,

$$M \approx 5.9639 \times 10^{24} \pm 1.0112 \times 10^{20} \iff M \in [5.9637 \times 10^{24}, 5.9641 \times 10^{24}].$$

Observació: L'aritmètica intervalar dóna una perspectiva diferent (al Teorema del valor mig) del problema d'acotar la propagació de l'error comès en avaluar una fórmula amb dades aproximades.

Exemple: El mateix que hem vist. La massa de la Terra: $M = \dfrac{gR^2}{G}$.

- $g = 9.80665$, $\varepsilon_a = \dfrac{1}{2}10^{-5} \Rightarrow g \in [9.806645, 9.806655]$;
- $G = 6.67428 \cdot 10^{-11}$,
  $\varepsilon_a = \dfrac{1}{2}10^{-16} \Rightarrow G \in [6.674275 \cdot 10^{-11}, 6.674285 \cdot 10^{-11}]$;
- $R = 6371.0 \cdot 10^3$, $\varepsilon_a = \dfrac{1}{2}10^2 \Rightarrow R \in [6.37095 \cdot 10^6, 6.37105 \cdot 10^6]$.

Aleshores

$$M \in [5.96381\ldots \times 10^{24}, 5.96401\ldots \times 10^{24}] \to M \in [5.9638 \times 10^{24}, 5.9641 \times 10^{24}]$$

Compareu amb

$$M \in [5.9637 \times 10^{24}, 5.9641 \times 10^{24}].$$

Algorismes estables i inestables

Volem calcular

$$R_n = \int_0^1 x^n e^{(x-1)}\,\mathrm{dx}.$$

Exercici: Observeu que:

- $R_0 = 1 - \exp(-1)$.
- $R_n > 0$ per a tot $n \geq 0$.
- $R_n = 1 - nR_{n-1}$ (integració per parts).

Algorisme de càlcul:

- $R_0 = 1 - \exp(-1)$.
- $R_j = 1 - jR_{j-1}$.

j	R_j
1	3.6787944117144233340e-01
2	2.6424111765711533320e-01
3	2.0727664702865400041e-01
4	1.7089341188538398834e-01
5	1.4553294057308008829e-01
10	8.3877070058292702020e-02
15	5.9033793641901866607e-02
18	-2.9453670751536265502e-02

Taula: $R_{18} < 0$ no té cap xifra significativa. Tots els càlculs han estat fets en format double de C.

Estudiem que ens ha passat: Ha d'haver estat un problema de cancel·lacions ja que totes les operacions són sumes, restes i productes.

Sigui $e_0 = |\overline{R_0} - R_0|$ l'error absolut inicial en la dada R_0. Llavors:

$$e_i = |\overline{R}_i - R_i| = |1 - i\overline{R}_{i-1} - 1 + iR_{i-1}| = i|\overline{R}_{i-1} - R_{i-1}| = ie_{i-1}.$$

Deduim doncs que per n gran es compleix

$$|e_n| = n!e_0,$$

i l'error és arbitrariament gran independentment de e_0.

Solució: Observem que es compleix

$$R_j = 1 - jR_{j-1} \rightarrow R_{j-1} = \frac{1 - e_j}{j}$$

Així, quan estudiem l'error s'obté

$$e_{j-1} = \frac{1}{j} e_j \Rightarrow e_k = \frac{1}{(n+k)(n+k-1)\cdots(n+1)} e_{n+k}.$$

Si ara podem aproximar el valor de R_{n+k} per k gran (en el nostre cas $R_{n+k} \equiv 0$ és una bona aproximació) el procés invers ens dóna R_n i l'error de propagació es fa petit a cada pas.

j	E_j	e_j
40	0	2.3e-2
35	2.7046289710763393 72e-02	2.9e-10
30	3.1279673932168072 79e-02	7.4e-18
25	3.7086214423739237 43e-02	4.3e-25
20	4.5544884075818053 98e-02	6.8e-32
18	5.0119854958094255 12e-02	1.83-34
15	5.9017540879297773 76e-02	3.6e-38
10	8.3877070103394166 25e-02	1.02e-43

Taula: Els càlculs fets amb un algorisme estable (l'algorisme invers).

Són problemes on la solució depèn de manera molt sensible de les dades (a les pel·licules això es diu efecte papallona).

Exemple:
El sistema d'equacions

$$\begin{aligned} 2.0000x + 0.6667y &= 2.6667, \\ 1.0000x + 0.3333y &= 1.3333, \end{aligned}$$

té solució $x = 1.0000$, $y = 1.0000$, mentre que el sistema

$$\begin{aligned} 2.0000x + 0.6665y &= 2.6667, \\ 1.0000x + 0.3333y &= 1.3333, \end{aligned}$$

té solució $x = 1.6666$, $y = -1.0000$.

Àlgebra lineal numèrica

En general un sistema de n equacions és un sistema del tipus

$$\begin{aligned} F_1(x_1,\dots x_n) &= 0 \\ F_2(x_1,\dots x_n) &= 0 \\ &\dots \\ F_n(x_1,\dots x_n) &= 0 \end{aligned}$$

on les funcions F_i poden ser lineals, algebràiques o analítiques. Així tenim

$x + 2 - \frac{x}{3} = 0 \quad \rightarrow$ una eq. lineal (Aillant)

$x^2 - 2x + 4 = 0 \quad \rightarrow$ una eq. no lineal (Newton)

$\begin{matrix} x + y = 0 \\ x - 2y = 1 \end{matrix} \quad \rightarrow$ un sistema d'eq. lineals (Gauss...)

$\begin{matrix} x^2 + y = 0 \\ x - 2y\sin(x) = 1 \end{matrix} \quad \rightarrow$ un sistema d'eq. no lineals (Newton)

Un sistema de m equacions lineals amb n incògnites s'escriu:

$$\begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & b_2 \\ \cdots \quad \cdots \quad \cdots & & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n & = & b_m \end{array}$$

La seva formulació matricial pren la forma $Ax = b$, on

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

- A és la matriu de coeficients,
- b és el vector de dades,
- x és el vector d'incògnites (o solucions, un cop resolt!!).

També definim els residus (que ens dóna informació de l'error comès en la resolució) com el vector

$$e = Ax - b = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{pmatrix}.$$

Podem doncs escriure

$$e_a = \|e\|, \quad e_r = \frac{\|e\|}{\|b\|}, \quad \text{on} \quad \|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$$

Suposem que volem resoldre el sistema lineal $Ax = b$, amb A una matriu quadrada. I suposem que el sistema té solució!!.

- Mètodes directes: Mòdul els errors d'arrodoniment i cancel·lació, busquen (i donen) la solució exacta del problema (Gauss, LU,...).

- Mètodes iteratius: Si $x \in \mathbb{R}^n$ és la solució exacta, aquests mètodes generen una successió de vectors x^k de tal forma que $||x^k - x|| \to 0$ quan $k \to \infty$, i llavors triem k de tal forma que la precisió sigui prou bona (solució amb tolerància prefixada).

- Matrius quadrades diagonals: Si el sistema $Ax = b$ és tal que $(a_{ij} = 0,\ i \neq j)$:

$$A = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}$$

amb $a_{jj} \neq 0,\ \forall j = 1, \ldots n$ (sistema compatible determinat) llavors la resolució és trivial:

$$x_j = \frac{b_j}{a_{jj}}, \quad j = 1, \ldots, n.$$

Requerim un total de n operacions.

- Matrius quadrades triangulars (superior): Si el sistema $Ax = b$ és tal que ($a_{ij} = 0,\ i < j$):

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}$$

amb $a_{jj} \neq 0,\ \forall j = 1, \ldots n$ (sistema compatible determinat) llavors la resolució és també molt fàcil:

$$x_j = \frac{1}{a_{jj}} \left(b_j - \sum_{k=j+1}^{n} a_{jk} x_k \right),\ j = n, n-1, \ldots 1.$$

Requerim un total de n^2 operacions (exercici).

Objectiu: Fer servir la teoria d'espais vectorials (àlgebra lineal) per convertir un sistema qualsevol d'equacions lineals (compatible determinat) $Ax = b$ en un sistema d'equacions lineal $Cx = d$ tal que

- $Cx = d$ té les mateixes solucions que $Ax = b$, i
- C és una matriu triangular superior (o diagonal).

Exemple 1: Volem resoldre el sistema lineal

$$\left\{ \begin{array}{ccccccc} x & & & + & z & = & 1, \\ x & + & 0.0001y & + & 2z & = & 2, \\ x & + & y & + & z & = & 0, \end{array} \right.$$

usant aritmètica de punt flotant amb $t = 4$ dígits i arrodoniment.
Escrivim la matriu ampliada

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & 1 \\ 1 & 0.0001 & 2 & 2 \\ 1 & 1 & 1 & 0 \end{array} \right).$$

Mètode de Gauss

Exemple 1 (continuació)

Apliquem el mètode de Gauss. Recordem que $t = 4$!!

Pas 1:

- Fila 2 $\rightarrow$ Fila 2$-$ Fila 1.
- Fila 3 $\rightarrow$ Fila 3$-$ Fila 1.

Obtenim la matriu

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & 1 \\ 0 & 0.0001 & 1 & 1 \\ 0 & 1 & 0 & -1 \end{array}\right).$$

Pas 2:

- Fila 3 $\rightarrow$ Fila 3 $- \frac{1}{0.0001}$ Fila 2.

Obtenim la matriu

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & 1 \\ 0 & 0.0001 & 1 & 1 \\ 0 & 0 & -10000 & -1 - \frac{1}{0.0001} \times 1 = -10000 \end{array}\right).$$

Cal resoldre el sistema lineal

$$\left\{\begin{array}{rcrcrcr} x & & & + & z & = & 1, \\ & & 0.0001y & + & z & = & 1, \\ & & & - & 10000z & = & -10000, \end{array}\right.$$

que és compatible determinat i la solució és $x^T = (0, 0, 1)$. Tenim que:

- $e^T = |(Ax - b)^T| = (0, 0, 1)$.
- $e_a = 1$ i $e_r = 0.4472\ldots$
- La solució exacta és $x^T = (1.0001, -1, -0.0001)$. Per tant no tenim cap xifra significativa!
- En el Pas 2 el terme $a_{22} = 0.0001$ és relativament molt petit.

Exemple 2: Volem resoldre el sistema lineal

$$\begin{array}{rrrrcr}
0.000135\,x_1 & +0.00201\,x_2 & -2.07\,x_3 & +3.04\,x_4 & = & 0.345 \\
4.21\,x_1 & +0.0432\,x_2 & -0.183\,x_3 & -12.3\,x_4 & = & 87.4 \\
-2.72\,x_1 & +205\,x_2 & -12.1\,x_3 & +324\,x_4 & = & -434 \\
0.00871\,x_1 & +332\,x_2 & +23.4\,x_3 & +0.0456\,x_4 & = & 23
\end{array}$$

(treballem amb precissió simple)

$$\left(\begin{array}{cccc|c} 1.3499999e-04 & 2.0099999e-03 & -2.0699999e+00 & 3.0400000e+00 & 3.4500000e-01 \\ 4.2100000e+00 & 4.3200001e-02 & -1.8300000e-01 & -1.2300000e+01 & 8.7400002e+01 \\ -2.7200000e+00 & 2.0500000e+02 & -1.2100000e+01 & 3.2400000e+02 & -4.3400000e+02 \\ 8.7099997e-03 & 3.3200000e+02 & 2.3400000e+01 & 4.5600001e-02 & 2.3000000e+01 \end{array}\right)$$

Pas 1

- (Fila 2) $=$ (Fila 2) $-$ ($3.1185188e+04$) * (Fila 1)
- (Fila 3) $=$ (Fila 3) $-$ ($-2.0148148e+04$) * (Fila 1)
- (Fila 4) $=$ (Fila 4) $-$ ($6.4518517e+01$) * (Fila 1)

$$\left(\begin{array}{cccc|c} 1.3499999e-04 & 2.0099999e-03 & -2.0699999e+00 & 3.0400000e+00 & 3.4500000e-01 \\ 0.0000000e+00 & -6.2639027e+01 & 6.4553152e+04 & -9.4815266e+04 & -1.0671489e+04 \\ 0.0000000e+00 & 2.4549777e+02 & -4.1718766e+04 & 6.1574371e+04 & 6.5171113e+03 \\ 0.0000000e+00 & 3.3187033e+02 & 1.5695332e+02 & -1.9609068e+02 & 7.4111187e-01 \end{array}\right)$$

Pas 2

- (Fila 3) $=$ (Fila 3) $-$ ($-3.9192462e+00$) * (Fila 2)
- (Fila 4) $=$ (Fila 4) $-$ ($-5.2981400e+00$) * (Fila 2)

$$\left(\begin{array}{cccc|c} 1.3499999e-04 & 2.0099999e-03 & -2.0699999e+00 & 3.0400000e+00 & 3.4500000e-01 \\ 0.0000000e+00 & -6.2639027e+01 & 6.4553152e+04 & -9.4815266e+04 & -1.0671489e+04 \\ 0.0000000e+00 & 0.0000000e+00 & 2.1128094e+05 & -3.1003000e+05 & -3.5307082e+04 \\ 0.0000000e+00 & 0.0000000e+00 & 3.4216859e+05 & -5.0254066e+05 & -5.6538305e+04 \end{array}\right)$$

Pas 3

- (Fila 4) $=$ (Fila 4) $-$ ($1.6194959e+00$) * (Fila 3)

$$\left(\begin{array}{cccc|c} 1.3499999e-04 & 2.0099999e-03 & -2.0699999e+00 & 3.0400000e+00 & 3.4500000e-01 \\ 0.0000000e+00 & -6.2639027e+01 & 6.4553152e+04 & -9.4815266e+04 & -1.0671489e+04 \\ 0.0000000e+00 & 0.0000000e+00 & 2.1128094e+05 & -3.1003000e+05 & -3.5307082e+04 \\ 0.0000000e+00 & 0.0000000e+00 & 0.0000000e+00 & -4.4835208e+02 & 6.4136877e+02 \end{array}\right)$$

Pas 4

- Sistema compatible determinat

Solució del sistema triangular

$$x = \begin{pmatrix} 1.6482010e+01 \\ 2.2871000e-01 \\ -2.2662044e+00 \\ -1.4305025e+00 \end{pmatrix}$$

Residu (estimació de l'error):

$$e = Ax - b = \begin{pmatrix} 1.4901161e-07 \\ -9.0408325e-03 \\ 7.2631836e-03 \\ 1.9132614e-02 \end{pmatrix}, \; e_a = 0.2237 \times 10^{-1}, \; e_r = 0.5046 \times 10^{-4}$$

Sigui $Ax = b$. Volem trobar l'algorisme general que se segueix dels dos casos particulars que hem vist.

Objectiu: Per cada $k = 1, \ldots n - 1$, modifiquem la matriu $A = A^1$

$$A^1 \mapsto A^2 \mapsto \ldots \mapsto A^n.$$

de forma que:

Els elements de la matriu A^k, que denotem per a_{ij}^k, sota l'element diagonal de les columnes entre la 1 i la k, siguin tots zero.

Suposem que som a la columna k
(aqui hi ha un bucle!)

- Per cada fila sota la fila k, és a dir per $\ell = k+1, \ldots, n$

$$m_{\ell k} = \frac{a_{\ell k}^{k}}{a_{kk}^{k}} \quad \text{(multiplicador fila } \ell\text{)}$$

$$a_{\ell j}^{k+1} = a_{\ell j}^{k} - m_{\ell k} a_{kj}^{k+1}, \; j = k+1, \ldots, n \quad \text{(bucle!)}$$

$$b_{\ell}^{k+1} = b_{\ell}^{k} - m_{\ell k} b_{k}^{k}$$

El resultat final és que el sistema $A^n x = b^n$ té les mateixes solucions que $Ax = b$ i és triangular superior.

En el pas k-èssim de l'eliminació gaussiana, per calcular cada multiplicador $m_{jk} = a_{jk}/a_{kk},\ j = k+1 \ldots n$, hem de dividir per l'element diagonal a_{kk}, anomenat pivot.

- Si $a_{kk} = 0$, podem canviar d'ordre dues files (equacions), per aconseguir un pivot diferent de zero (si no és posible, vol dir que el determinant del sistema és zero).
- Si $|a_{kk}| \neq 0$ però és petit, pot crear inestabilitat numèrica.

En el que al pas $k-$èssim es pren com a pivot el coeficient de valor absolut més gran entre els $a_{\ell k}$ ($\ell = k, \ldots, n$).

- Calculem $\bar{\ell}$ tal que

$$|a_{\bar{\ell}k}| > |a_{\ell k}|, \quad \ell = k, \ldots, n, \;\; \bar{\ell} \neq \ell.$$

- Intercanviem la fila ℓ amb la fila k.
- Apliquem Gauss a la columna k.

Exemple 1: Volem resoldre el sistema lineal

$$\left\{\begin{array}{ccccccc} x & & & + & z & = & 1, \\ x & + & 0.0001y & + & 2z & = & 2, \\ x & + & y & + & z & = & 0, \end{array}\right.$$

usant aritmètica de punt flotant amb 4 dígits i arrodoniment.
Escrivim la matriu ampliada

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & 1 \\ 1 & 0.0001 & 2 & 2 \\ 1 & 1 & 1 & 0 \end{array}\right).$$

Apliquem el mètode de Gauss amb pivotatge maximal per columnes.

Pas 1: No pivotem

- Fila 2 $\to$ Fila 2$-$ Fila 1.
- Fila 3 $\to$ Fila 3$-$ Fila 1.

Obtenim la matriu

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & 1 \\ 0 & 0.0001 & 1 & 1 \\ 0 & 1 & 0 & -1 \end{array}\right).$$

Pas 2: Pivotem 2 i 3

- Fila 2 $\leftrightarrow$ Fila 3.

Obtenim la matriu

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0.0001 & 1 & 1 \end{array}\right).$$

Pas 3:

- Fila 3 $\rightarrow$ Fila 3 $-$ 0.0001 Fila 2.

Obtenim la matriu

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

La solució del sistema és $x^T = (0, -1, 1)$.

- $\mathbf{e}^T = (0, 0, 0)$.
- En el Pas 2 hem canviat el pivot...

Exemple 2:

$$\left(\begin{array}{cccc|c} 1.3499999e-04 & 2.0099999e-03 & -2.0699999e+00 & 3.0400000e+00 & 3.4500000e-01 \\ 4.2100000e+00 & 4.3200001e-02 & -1.8300000e-01 & -1.2300000e+01 & 8.7400002e+01 \\ -2.7200000e+00 & 2.0500000e+02 & -1.2100000e+01 & 3.2400000e+02 & -4.3400000e+02 \\ 8.7099997e-03 & 3.3200000e+02 & 2.3400000e+01 & 4.5600001e-02 & 2.3000000e+01 \end{array}\right)$$

<u>Pas 1</u> Pivotem 1 $\leftrightarrow$ 2

$$\left(\begin{array}{cccc|c} 4.2100000e+00 & 4.3200001e-02 & -1.8300000e-01 & -1.2300000e+01 & 8.7400002e+01 \\ 1.3499999e-04 & 2.0099999e-03 & -2.0699999e+00 & 3.0400000e+00 & 3.4500000e-01 \\ -2.7200000e+00 & 2.0500000e+02 & -1.2100000e+01 & 3.2400000e+02 & -4.3400000e+02 \\ 8.7099997e-03 & 3.3200000e+02 & 2.3400000e+01 & 4.5600001e-02 & 2.3000000e+01 \end{array}\right)$$

- (Fila 2) $=$ (Fila 2) $-$ ($3.2066506e-05$) * (Fila 1)
- (Fila 3) $=$ (Fila 3) $-$ ($-6.4608073e-01$) * (Fila 1)
- (Fila 4) $=$ (Fila 4) $-$ ($2.0688835e-03$) * (Fila 1)

$$\left(\begin{array}{cccc|c} 4.2100000e+00 & 4.3200001e-02 & -1.8300000e-01 & -1.2300000e+01 & 8.7400002e+01 \\ 0.0000000e+00 & 2.0086146e-03 & -2.0699940e+00 & 3.0403943e+00 & 3.4219739e-01 \\ 0.0000000e+00 & 2.0502791e+02 & -1.2218233e+01 & 3.1605319e+02 & -3.7753253e+02 \\ 0.0000000e+00 & 3.3199991e+02 & 2.3400377e+01 & 7.1047269e-02 & 2.2819180e+01 \end{array}\right)$$

Pas 2 Pivotem 2 $\leftrightarrow$ 4

$$\left(\begin{array}{cccc|c} 4.2100000e+00 & 4.3200001e-02 & -1.8300000e-01 & -1.2300000e+01 & 8.7400002e+01 \\ 0.0000000e+00 & 3.3199991e+02 & 2.3400377e+01 & 7.1047269e-02 & 2.2819180e+01 \\ 0.0000000e+00 & 2.0502791e+02 & -1.2218233e+01 & 3.1605319e+02 & -3.7753253e+02 \\ 0.0000000e+00 & 2.0086146e-03 & -2.0699940e+00 & 3.0403943e+00 & 3.4219739e-01 \end{array}\right)$$

- (Fila 3) $=$ (Fila 3) $-$ ($6.1755413e-01$) * (Fila 2)
- (Fila 4) $=$ (Fila 4) $-$ ($6.0500456e-06$) * (Fila 2)

$$\left(\begin{array}{cccc|c} 4.2100000e+00 & 4.3200001e-02 & -1.8300000e-01 & -1.2300000e+01 & 8.7400002e+01 \\ 0.0000000e+00 & 3.3199991e+02 & 2.3400377e+01 & 7.1047269e-02 & 2.2819180e+01 \\ 0.0000000e+00 & 0.0000000e+00 & -2.6669233e+01 & 3.1600931e+02 & -3.9162460e+02 \\ 0.0000000e+00 & 0.0000000e+00 & -2.0701356e+00 & 3.0403938e+00 & 3.4205934e-01 \end{array}\right)$$

Pas 3: No pivotem

- (Fila 4) = (Fila 4) − (7.7622615e − 02) * (Fila 3)

$$\left(\begin{array}{cccc|c} 4.2100000e+00 & 4.3200001e-02 & -1.8300000e-01 & -1.2300000e+01 & 8.7400002e+01 \\ 0.0000000e+00 & 3.3199991e+02 & 2.3400377e+01 & 7.1047269e-02 & 2.2819180e+01 \\ 0.0000000e+00 & 0.0000000e+00 & -2.6669233e+01 & 3.1600931e+02 & -3.9162460e+02 \\ 0.0000000e+00 & 0.0000000e+00 & 0.0000000e+00 & -2.1489075e+01 & 3.0740986e+01 \end{array}\right)$$

Pas 4

- Sistema compatible determinat

Solució del sistema triangular

$$x = \begin{pmatrix} 1.6479750e+01 \\ 2.2877161e-01 \\ -2.2662601e+00 \\ -1.4305402e+00 \end{pmatrix}$$

Residus (estimació de l'error)

$$\mathbf{e} = \begin{pmatrix} -7.1525574e-07 \\ 7.6293945e-06 \\ 3.0517578e-05 \\ 9.5367432e-06 \end{pmatrix}, \; e_a = 0.3287\ldots \cdot 10^{-4}, \; e_r = 0.7416\ldots 10^{-7}$$

Descomposició LU

Motivació

Suposem que volem resoldre simultàneament els sistemes $Ax = b^i, \ i = 1 \dots p$ (la mateixa matriu A però p diferents termes independents). Farem

Suposem que som a la columna k
(aqui hi ha un bucle!)

- Per cada fila sota la fila k, és a dir per $\ell = k+1, \dots, n$

$$m_{\ell k} = \frac{a_{\ell k}^k}{a_{kk}^k} \quad \text{(multiplicador fila } \ell\text{)}$$

$$a_{\ell j}^{k+1} = a_{\ell j}^k - m_{\ell k} a_{kj}^{k+1}, \ j = k+1, \dots, n \quad \text{(bucle!)}$$

$$b_\ell^{i,k+1} = b_\ell^{i,k} - m_{\ell k} b_k^{i,k}, \ i = 1, \dots, p \quad \text{(bucle!)}$$

Observem que tot el que fem és extendre Gauss a tots els termes independents (bucle $i = 1, \dots, p$).

Dues consideracions previes:

- Per resoldre un sistema lineal d'equacions per Gauss d'ordre n (amb $p << n$) el nombre d'operacions és (exercici):

$$\frac{2}{3}n^3 + o(n^2) \quad (\text{ordre} \quad n^3)$$

- Molts cops, tot i que haurem de resoldre $Ax = b^i, \ i = 1 \dots p$ (la mateixa A però p diferents termes independents), no sabem els valors de b^j fins que no hem resolt el sistema

$$Ax = b^i, \ i = 1, \dots, j-1$$

- Volem trobar un mètode que guardi la informació rellevant per tal que poguem resoldre diversos sistemes lineals que comparteixen la mateixa matriu A fent menys operacions.

Sigui A una matriu quadrada amb $\det(A) \neq 0$. Volem trobar matrius L (lower) i U (upper) tals que

- $A = LU$.
- L és una matriu triangular inferior amb $1's$ a la diagonal.
- U és triangular superior.

En cas que ho pogem fer llavors és equivalent resoldre $Ax = b$ a resoldre el sistema $LUx = b$, i aquest és equivalent a resoldre dos sistemes lineals triangulars (en un ordre concret):

- $Ly = b$.
- $Ux = y$. Això ens dóna x.

Important: La descomposició LU d'una matriu A és independent del sistemes d'equacions lineals. Depèn només de la matriu A.

Proposició: Sigui A una matriu $n \times n$. Suposem que $\det(A) \neq 0$.

(a) Si existeix la descomposició LU de la matriu A, llavors aquesta és única.

(b) Si es pot fer triagularització de Gauss de la matriu A sense pivotatge de cap tipus llavors A admet descomposició LU i a més:

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ m_{n1} & m_{n2} & \cdots & 1 \end{pmatrix} \quad \text{i} \quad U = \begin{pmatrix} a_{11}^1 & a_{12}^1 & \cdots & a_{1n}^1 \\ 0 & a_{22}^2 & \cdots & a_{2n}^2 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn}^n \end{pmatrix}$$

on m_{ij} són els multiplicadors del mètode de Gauss (sense pivotatge) i a_{ij}^k són els elements de la matriu A^k en el procés d'eliminació Gaussiana.

Exemple: Considerem la matriu

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

Apliquem el mètode de Gauss.

<u>Pas 1</u>:

- Fila 3 $\rightarrow$ Fila 3 $- m_{31}$ Fila 1, on $m_{31} = \frac{1}{1} = 1$.

Obtenim la matriu

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

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

Pas 2:

- Fila 3 $\rightarrow$ Fila 3 $-$ m_{32} Fila 2, on $m_{32} = \frac{2}{2} = 1$.

Obtenim la matriu

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

Per tant el resultat de la descomposició és

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{i} \quad U = \begin{pmatrix} 1 & -2 & 3 \\ 0 & 2 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

- (Determinants) Per calcular el determinant d'una matriu A de tamany significatiu el que farem és triangularitzar la matriu fent ús del mètode de Gauss (amb o sense pivotatge) i observar que

$$\det(A) = (-1)^H a_{11}^1 \times a_{22}^2 \times \cdots \times a_{nn}^n$$

 on H és el nombre de canvis de files o columnes que haguem realitzat a l'aplicar el mètode de Gauss (pivotage maximal!!)
- (Inverses) Si es possible no calcular inverses ho evitem. Per exemple per calcular $M^{-1}v$ on M és una matriu no singular i v és un vector el que farem és resoldre el sistema d'equacions lineal $Mx = v$ i llavors es té que $x = M^{-1}v$.
- (Inverses) En qualsevol cas, si volem calcular la inversa de A resoldrem n sistemes lineals $Ax_i = e_i$ on e_i és el vector i–èssim de la base canònica. Aleshores A^{-1} és la matriu que té per columnes els vectors x_i.

Volem resoldre $Ax = b$ via trobar una sequència de vectors $x^k \mapsto x$ quan $k \mapsto \infty$. Suposem que els elements de la diagonal de la matriu A no són zero. Podem escriure A de la següent forma

$$A = D(L + I + U),$$

on D és diagonal (la diagonal de la matriu A), I és la identitat, L és triangular inferior amb zeros a la diagonal i U és triangular superior amb zeros a la diagonal (és fàcil deduir les matrius L i U en funció de la matriu A). Llavors

$$Ax = b \iff D(L+I+U)x = b \iff (L+I+U)x = D^{-1}b \iff x = -(L+U)x+D^{-1}b.$$

I ara busquem x solució de l'equació anterior de la forma següent:

- Agafem $x^0 \in \mathbb{R}^n$ *qualsevol*.
- Resolem recursivament les equacions:

$$x^{k+1} = -(L + U)x^k + D^{-1}b.$$

La matriu $B = -(L + U)$ s'anomena matriu d'iteració i la condició de convergència és $\rho(B) < 1$, on $\rho(B)$ és el mòdul màxim dels valors propis de B.

Alternativament, també es pot escriure el mètode de Jacobi en coordenades, i s'obté la fórmula recurrent:

- Agafem $x^0 \in \mathbb{R}^n$ *qualsevol*.
- $$x_i^{k+1} = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij}x_j^k\right) \quad k = 0,\dots \quad i = 1,\dots,n.$$

Sigui A una matriu quadrada $n \times n$. Direm que $v \neq 0$ és un vector propi de A si existeix un número real λ tal que

$$Av = \lambda v.$$

El número λ s'anomena valor propi de A de vector propi v. Tot i que no és un fet general assumirem durant tot aquest capítol que

- A admet n vectors propis (reals) linealment independents
- Els n valors propis corresponents (comptant multiplicitat) són reals i s'ordenen segons el seu mòdul com

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|$$

És a dir, assumim, que només n'hi ha un de mòdul màxim.

Exemple: Els valors propis de la matriu

$$\begin{pmatrix} 4 & 0 & 4 \\ 8 & 4 & 0 \\ 16 & 0 & 4 \end{pmatrix}$$

són $\lambda_1 = 12$, $\lambda_2 = 4$ i $\lambda_3 = -4$. Els vectors propis corresponents són, respectivament,

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad v_3 = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}$$

Problema: Calcular $\lambda_1 = 12$, el valor propi de mòdul màxim, fent servir el mètode de la potència.

El mètode de la potència considera les iteracions

$$x^{(0)} \in \mathbb{R}^n$$
$$x^{(k+1)} = Ax^{(k)}, \ k \geq 0$$

Es compleix que, genèricament (la convergència no està sempre garantida per tant sempre haurem de posar un límit en el procés iteratiu),

$$q_i^{(k)} = \frac{x_i^{(k+1)}}{x_i^{(k)}} \to \lambda_1 \ (\forall i = 1, \ldots, n)$$
$$y^{(k)} = \frac{x^{(k+1)}}{\lambda_1^k} \to v,$$

on $v \neq 0$ és un vector propi associat al valor propi λ_1.

Observació: La velocitat de convergència depèn del qocient $|\frac{\lambda_1}{\lambda_2}| > 1$. Quan $|\frac{\lambda_1}{\lambda_2}| \gtrsim 1$, la velocitat pot ser arbitràriament lenta.

Excepte que $|\lambda_1| = 1$ la successió λ_1^k o bé tendeix a zero o bé no està acotada. Això implica que en el procés iteratiu (no mostrem aqui els detalls) hi hagi inconvenients numèrics (overflows, etc).

Per tal d'evitar-los a cada pas del procés iteratiu considererem

$$z^{(k)} = \frac{x^{(k)}}{||x^{(k)}||}$$

on $||\cdot||$ és la norma euclidiana. Així farem

$$x^{(k+1)} = Az^{(k)}$$

i podrem concloure que, genèricament,

$$\frac{x_i^{(k+1)}}{z_i^{(k)}} \to \lambda_1 \;\; (\forall i = 1, \ldots, n)$$

$$z^{(k)} \to v.$$

Observació: El mètode de la potència ens dóna el valor propi de mòdul més gran (assumint és únic). Però, i els altres? Veiem esquemàticament una resposta (pel cas de matrius simètriques).

Observació inversa: Una primera resposta és observar que si A es pot invertir (altrament $\lambda = 0$ és també valor propi de A), i existeix un únic valor propi de modul mínim μ, llavors podem aplicar el mètode de la potència a la matriu A^{-1} i trobar-lo.

Deflació: La idea bàsica dels processos de deflació (en general) és, donada una matriu de la que sabem un valor propi λ i un vector propi v (mètode de la potència), trobar una nova matriu (més simple, de dimensió menor) que tingui els mateixos valors propis de l'anterior, tret del que ja sabem λ.

Aquest tema no el tractarem.