Analyse numérique 1

Écriture sous forme normalisée : $0,3141085 \times 10^2$
Ds un ordinateur, les nombres sont stockés sous forme normalisée.
Le nb réellement stocké est une approximation du nb fournit par l'utilisateur. car l'ordi dispose d'un nb limité de place.
virgule flottante en simple précision selon la Standard IEEE-754 $x = (-1)^s \times f \times 2^e$
avec : x : le nombre
s : le signe (0 si >0 ; 1 sinon)
f : Mantisse sur 23 bits
e : exposant codé sur 8 bits.
1 bit caché.
0 : exp = 00000000 = $\pm(0\ldots0)_2 \times 2^{-127}$
∞ : exp = 11111111 si $b_i = 0$
si $\exists b_i \setminus b_i \neq 0$ alors NaN.

+ petit nb :
0 00000001 0...0 : mp
+ grd : 0 11111110 111 ... 111 : mg
$mp \approx 1,2 \times 10^{-35}$ $mg \approx 3,4 \times 10^{38}$
+ nb normalisé : exposant nul et la mantisse a au moins un bit $\neq 0$
La précision eps d'un ordinateur est le + petit nombre positif normalisé tq $1 + eps \neq 1$.
Pour la simple précision, on a $eps = 2^{-23} \approx 1,192 \times 10^{-7}$
$fl(x) = \tilde{x}$ = nb dans l'ordinateur (x vrai).

3 types d'erreur :
- erreur de représentation : $\Delta x = fl(x) - x$
- erreur relative de représentation :
$$\varepsilon(x) = \frac{\Delta x}{fl(x)} = \frac{fl(x) - x}{fl(x)}$$
- erreur relative de précision :
$$\eta(x) = \frac{\Delta x}{x} = \frac{fl(x) - x}{x} \quad \text{donc } fl(x) = x(1+\eta(x))$$
- En règle générale pour l'analyse des erreurs on utilise la valeur absolue de l'erreur de précision relative, qui pour le standard est :

arrondi $\nearrow$ $\eta(x) = \begin{cases} \beta^{1-p} \\ 0,5 \times \beta^{1-p} \end{cases}$ avec p = nb chiffre mantisse
β = base de numération.

Pour calculer l'intervalle de variation $[E_{min}; E_{max}]$ de l'exposant on utilise le biais
$B = \underbrace{0 1 \ldots 1}_{q-1} = 2^{(q-1)} - 1$ avec q = nb chiffre expo
i.e. : contenu expo = biais + $(E)_2$

Simple précision	Double préci
nb bit mantisse = 23 + 1	nb bit M = 52 + 1
$E_{min} = +127$	nb bit E = 11
$E_{max} = 126$	$E_{min} = 1023$

Sources d'erreur
- modélisation math du système C.
- approximation des fonctions analytiques du modèle mathématique
- discrétisation des fonctions obtenues par approximation des fonctions analytique afin qu'elle soient traitées par l'ordi.
- calcul numérique par l'ordi

$a \circledast b = fl(fl(a) \circledast fl(b))$

th erreur des sommes :
$S = x_1 + \ldots + x_n$
erreur absolue : $\Delta s = \Delta x_1 + \Delta^2 x_2 + \ldots + \Delta^n x_n$
erreur relative entrée : $\eta(s) = \frac{x_1}{s}\eta'(x_1) + \ldots + \frac{x_n}{s}\eta^n(x_n)$
On veut borner l'erreur d'entrée :
$$|\eta(s)| \leq \frac{\sum |x_i|}{|\sum x_i|} \times eps.$$
$a \otimes b = fl(fl(a) \otimes fl(b))$
L'erreur est en 2 parties :
- erreur d'entrée : $\eta(a \otimes b)$
- erreur de calcul : $\eta(a \otimes b)$
L'erreur d'entrée dépend de la façon dont sont agencés les calculs.

opération	erreur d'entrée
$a \pm b$	$\eta(a \pm b) = \frac{a}{a \pm b}\eta(a) \pm \frac{b}{a \pm b}\eta(b)$
$a \times b$	$\eta(a \times b) = \eta(a) + \eta(b)$
a / b	$\eta(a/b) = \eta(a) - \eta(b)$
$\sqrt{a}$	$\eta(a) = 0,5 \times \eta(a)$

On peut le borner :
$\eta(p) \leq N \times eps$ avec N nb de facteur
$\eta(Q) \leq 2\, eps$ avec Q quotient.
On évalue une fonction grâce au nb conditionnement
$$k(x) = \left| \frac{\frac{f(x) - f(x')}{f(x)}}{\frac{x - x'}{x}} \right| \sim \left| \frac{f'(x)}{f(x)} \times x \right|$$
Si k est petit, la fonction est bien conditionnée
On évalue de la fin manière un algo en décomposant toutes les fonctions
l'erreur du résultat
$$\Delta y = \begin{bmatrix} \Delta y_1 \\ \Delta y_n \end{bmatrix} \sim \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \ldots \\ \frac{\partial f_1}{\partial x_n} & \ldots \end{bmatrix} \Delta x = J[f(x)]\, \Delta x$$
avec $J[f(x)]$ jacobien
erreur relative :
$$\eta(y) = \begin{bmatrix} \eta_1(y_1) \\ \eta'(y_n) \end{bmatrix} = k(\phi, x).\eta(x)$$ avec k nombre de condition

Algebre linéaire:

norme:
$\|x\| > 0$ et $\|x\| = 0 \Leftrightarrow x = 0$.
$\|\alpha x\| = |\alpha| \, \|x\|$
$\|x+y\| \leq \|x\| + \|y\|$
$\|x-y\| \geq |\|x\| - \|y\||$

Def norme: $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$; $p \geq 1$.

Normes utiles:
1 → $\|x\|_1 = \sum |x_i|$
2 → $\|x\|_2 = \sqrt{\sum x_i^2}$
3 → $\|x\|_\infty = \max |x_i|$.

Inégalité de Hölder

$$\sum_{i=1}^{m} |x_i y_i| \leq \|x\|_p . \|y\|_p = \left(\sum |x_i|^p \right)^{1/p} \cdot \left(\sum |y_i|^p \right)^{1/p}$$

Pour $p=2$, on retrouve l'inégalité de Cauchy.

Si $\|A.B\| \leq \|A\|.\|B\|$, on parle de norme sous-multiplicative.

- $\text{lub}(A) = \sup_{x \in \mathbb{R}^m} \frac{\|Ax\|}{\|x\|}$ lub = norme subordonnée.
- $\text{Cond}(A) = \|A\|.\|A^{-1}\| = K(A)$
- matrice régulière = inversible
- matrice singulière = non inversible et $K(A) = \infty$.

$\|A\|_1 = \max_j \sum_{i=1}^{m} |a_{ij}|$ (somme des colonnes)

$\|A\|_2 = \left(\sum_{i,j=1}^{m} |a_{ij}|^2 \right)^{1/2}$

$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|$ somme des lignes.

Une norme matricielle est consistante avec une norme vectorielle si $\|Ax\| \leq \|A\|.\|x\|$.

- norme de Frobenius:
$\|A\|_F = \left(\sum_i \sum_j |a_{ij}|^2 \right)^{1/2} = \sqrt{\text{tr}(A^T A)}$

- **Th:** Soit $\|.\|$ une norme quelconque, alors le rayon spectral d'une matrice carré A est $\leq$ à sa norme : $\rho(A) \leq \|A\|$.
- A orthogonale $\Leftrightarrow A^T = A^{-1}$

Th de Von Neumann:
Soit A avec $\rho(A) < 1$, $(I-A)$ est régulière
$(I-A)^{-1} = \sum_{k=0}^{\infty} A^k$

Th: A matrice carré, alors
- $\lim_{k \to +\infty} A^k = 0$
- $\forall x \in \mathbb{R}^m$, $\lim_{k \to +\infty} A^k x = 0$.
- $\rho(A) < 1$
- $\|A\| < 1$ pour au moins une borne subordonnée.

Th: si A régulière et si $\frac{\|\Delta A\|}{\|A\|} \leq \frac{1}{\text{cond}(A)}$
alors $A + \Delta A$ est aussi régulière.

- Soit le système $Ax = b$
perturbation de b; b devient $b + \Delta b$.
$\Leftrightarrow x \to x + \Delta x$ $\Rightarrow \frac{\|\Delta x\|}{\|x\|} \leq K(A) . \frac{\|\Delta b\|}{\|b\|}$.

Si $K(A) \gg 1$, une petite perturbation de b peut provoquer une grande perturbation de la solution.

- Soit le système ax=b, les perturbations sur A $\Rightarrow A \to A + \Delta A$
$\Leftrightarrow \frac{\|\Delta x\|}{\|x\|} \leq K(A) \frac{\|\Delta A\|}{\|A\|}$.

- Si perturbations sur A et b.
$(A + \Delta A)(x + \Delta x) = b + \Delta b$
$\Rightarrow \frac{\|\Delta x\|}{\|x\|} \leq K(A) . \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta B\|}{\|B\|} \right)$

- Wilkinson: analyse de l'erreur:
$|x \oplus y - fl(x \oplus y)| \leq \text{eps} \, |fl(x \oplus y)|$
$|e_k| \leq \text{eps} . \delta_k$ avec $\delta_k = \delta_{k-1} + |x_k - fl(s_k)|$

- Une matrice A est dite symétrique est définie positive si $x^T A x > 0 \; \forall x \in \mathbb{R}^n$
- Toute matrice régulière A peut se factoriser de manière unique selon le produit $A = LU$ avec

$$L = \begin{pmatrix} 1 & & \\ l_{21} & 1 & 0 \\ \vdots & \ddots & 1 \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} \quad U = \begin{pmatrix} R_{11} & R_{12} & \cdots & \cdots \\ & R_{22} & \ddots & \vdots \\ & & \ddots & \\ 0 & & & R_{mm} \end{pmatrix}$$

- une matrice carré A est dite diagonalement dominante si $|a_{ij}| \geq \sum_{\substack{k=1 \\ k \neq i}}^{m} |a_{ik}|$ et

si $\exists \, i_0$ tq $|a_{i_0 i_0}| > \sum_{\substack{k=1 \\ k \neq i_0}}^{m} |a_{i_0 k}|$

une telle matrice est régulière.

Vecteur et valeurs propres

$Au = \lambda u$

A matrice carrée et λ vp, u vecteur propre, ona :
- λ^k est valeur propre de A^k.
- $1/\lambda$ est vp de A^{-1}
- $\lambda + R$ est vp de $A + RI_m$
- $c\lambda$ vp de cA. $tr(A) = \lambda_1 + \lambda_2 + \dots + \lambda_m$
- si u vecteur propre de A alors $u' = \frac{1}{\|u\|} u$ l'est aussi. de plus $\|u'\| = 1$.
- si B matrice régulière, A vp de $B^{-1}AB$
- si B mat orthogonale ou unitaire, λ est une vp de $B^T AB$
- si A matrice symétrique déf. positive, toutes les vp sont positives ou nulles.
- si A matrice triangulaire, alors les éléments diagonaux sont des vp.
- Soit u un vp droit et v un vp gauche, alors $\langle u.v \rangle = 0$
- Soit A une matrice carrée diagonalisable et soit U tq $U^{-1}AU = diag(\lambda, \dots, \lambda_m)$. soit $B = A + \Delta M$ alors $\exists \lambda$ tq
$$|u' - \lambda| \leq k_\infty(U) . \|\Delta A\|_\infty .$$
- $rang(AB) = rang(B) - dim(N(A) \cap R(B))$
- $rg(AB) \leq \min(rg(A), rg(B))$.
- $rg(A) - rg(B) \leq rg(A+B) + n$.

$rg(A^TA) = R(A^T)$ et $R(AA^T) = R(A)$.
$N(A^TA) = N(A)$ et $N(AA^T) = N(A^T)$

Décomposition en valeur singulière :

Toute matrice rectangulaire $A \in R^{m \times n}$ de rang $R \leq \min(R(A), n)$ admet une factorisation $A = U \Delta V^T$ avec :
- $U = [u_1, \dots, u_m] \in R^{m \times m}$ matrice orthogonale dont les colonnes u_i sont les vecteurs propres de la matrice AA^T
$AA^T = U\Delta V^T (U\Delta V^T) U \Delta \Delta^T U^T$.
- $V = [v_1, \dots, v_m] \in R^{n \times m}$ matrice orthogonale dont les colonnes v_i sont les vecteurs propres de la matrice A^TA
$A^TA = (U\Delta V^T)^T U \Delta V^T = V \Delta^T \Delta V$
- $\Delta \in R^{m \times n}$ la matrice des valeurs singulières $\Delta = \Lambda^{1/2}$ avec Λ étant la matrice diagonale des vp non nulles de la matrice A^TA et aussi des vp non nulles de AA^T.

Soit $A \in R^{m \times n}$ et la DVS $A = U\Delta V^T$ avec les vs $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m > 0$ alors on a $\|A\| = \sigma_1$, $\|A^{-1}\| = \sigma_m$.

Pseudo inverse de $A = U\Delta V^T$ est
$$A^T = V\Delta^{-1}U^T$$

Th :
Le problème des moindres carrés a au moins une solution a_0. Si a_1 est une autre solution, alors $Xa_0 = Xa_1$. Le résidu $r = y - Xa_0$ est unique et il est solution de l'équation $X^T r = 0$.
Chaque point a_0 qui réalise $\min_{z \in R^n} \|y - Xz\|$ est aussi solution du système d'eq. normales.
$a_0 = (X^TX)^{-1}X^T y = X^+ y$.
où X^+ est la pseudo inverse.

Une matrice de permutation P est une matrice dont chaque ligne et chaque colonne a un élément = à 1 et tous les autres sont nuls

Théorème de Gershgorin :

Toutes les valeurs propres de la matrice A se trouvent dans l'union des cercles de Gershgorin
$$C_i = \left\{ z \in R / |z - a_{ii}| \leq \sum_{\substack{k=1 \\ k \neq i}}^{n} |a_{ik}| \right\}$$

Quotient de Rayleigh :

Si A est une matrice carrée, le coef de Rayleigh est :
$$R[x] = \frac{x^T A x}{x^T x}$$
Si A symétrique alors $|R(x)| \leq |\lambda_{max}|$.

Factorisation LU :

```
Pour j←1 à n-1
  Pour i←j+1 à n
    A(i,j) = A(i,j) / A(j,j)
    Pour k← j+1 à n
      A(i,k) = A(i,k) - A(i,j) x A(j,k)
    Fin Pour
  Fin Pour
Fin Pour.
```

Inverser Matrice.

```
Pour j=1 à n
 Pour j=1 à n
  Si (i=j)
    Pour k=1 à n
      B'(j,k) = B(j,k) / B(j,j)
    Fin Pour
  Sinon
    Pour k=1 à n
  B'(i,k) = B(i,k) - B(i,j) * B'(j,k)
    Fin Pour
 Fin Si
 Fin P
Fin P.
```

Algo de Cholesky

$L(1,1) = \sqrt{A(1,1)}$

Pour $i = 2$ à n

Pour $j = 1$ à $i-1$

$$L(i,j) = \frac{1}{L(j,j)} \left(A(i,j) - \sum_{k=1}^{j-1} L(i,k) \times L(j,k) \right)$$

Fin Pour

$L(i,i) \leftarrow \left(A(i,i) - \sum_{k=1}^{i} L^2(i,k) \right)^{1/2}$

Fin Pour

Méthode itérative

$$\Pi x^{(k+1)} = N x^{k} + b. \quad \Leftrightarrow x^{(k+1)} = \Pi^{-1} N x^{(k)} + \Pi^{-1}$$

$$A = \begin{bmatrix} D & F \\ E & D \end{bmatrix} = \begin{bmatrix} L & D \\ D & U \end{bmatrix}.$$

Méthode décompo. $A = \Pi - N$ matrice $\Pi^{-1}N$

Jacobi $A = \underbrace{D}_{M} - \underbrace{(E+F)}_{N}$

$$J = \Pi^{-1}N = D^{-1}(E+F) = I - D^{-1}A$$

descrip d'1 itérat°:

$$Dx^{(k+1)} = (E+F)x^{k} + b$$

Gauss-Seide:

$$A = \underbrace{D}_{\Pi} - \underbrace{(E+F)}_{N} \underbrace{(D-E)}_{\Pi} + \underbrace{F}_{N}$$

$$G = \Pi^{-1}N = (D-E)^{-1}F$$

descripto

$$(D-E)x^{(k+1)} = Fx^{(k)} + b.$$

Richardson:

$$A = \underbrace{I}_{\Pi} - \underbrace{(I-A)}_{N}$$

$$R = \Pi^{-1}N = I - A$$

descripto: $x^{(k+1)} = (I-A)x^{k} + b$

Jacobi ~~$A = D - (L+U)$~~.

Fascicule 1 | Chap 1 - Analyse des erreurs

- forme normalisée : $0, \underbrace{b_1 b_2 \dots b_p}_{\text{mantisse}} \times 10^{\dots}$. forme standardisée : en respectant le nb de places . $110,11 \Rightarrow 1,10110 \times 10^{101}$
- nbs stockés ≠ des approximat° car ont chiffres infinis en partie décimale.
- nbs flottants norme IEEE-754 : s(1) e(8) m(23) en binaire
  - séparer et convertir partie entière et partie décimale . réécrire les 2 parties . normaliser
    - on divise par 2 et on lit ↑ restes ↳ on multiplie et on lit ↓ les parties entières
    - exposant ⇒ -1 +127 (on enlève bit caché) . on enlève 1er bit (1) de mantisse
- décodage : . $1, \text{mantisse} \times 2^{\text{exposant} - 127}$ → biais
- 0 → que des zéros . ⊕ petit nb positif stockable normalisé : 0|0 ... 01|0...0 ⇒ 2^{-126} pr base 2
- variat° de e ⇒ ℰ -126 et 127 . ⊕ grd nb normalisé : 0|1 .. 10|1 ... 1| ⇒ $1,111 \times 2^{127}$ pr base 2
- ⊕ petit nb non nul : 2^{-149} ⇒ 0|0 ... 0|0 .. 01
- normalisés : exposant a au ⊖ 1 bit ≠ et au ⊖ 1 = 0 — double précision : 127 → 1023
- sous-normalisés : exposant nul, mantisse a au ⊖ 1 bit ≠ 0
- zéro : tout à 0 (1 ⊕ et 1 ⊖)
- NaN : exposant 1...1 et mantisse au ⊖ 1 bit ≠ 0
- infini : exposant 1...1 et mantisse 0...0
- précision eps d'1 pc est le ⊕ petit nb positif normalisé tq $1 + \text{eps} \neq 1$ ⇒ 2^{-23} pr norme IEEE-754 ⇒ $\text{eps} = 2^{-p}$ (base (β) ↑, bits mantisse ↓)
- 3 types d'erreur :
  - erreur de représentat° : $\Delta x = fl(x) - x$ absolue : $|\Delta x|$ — pas tjrs vrai, leurre calcul ℂ au partiel
  - erreur relative de représentat° : $\varepsilon(x) = \frac{\Delta x}{fl(x)}$
  - ——— de précision : $\eta(x) = \frac{\Delta x}{x}$ ⇒ $fl(x) = x(1+\eta(x))$
- on utilise en général val abs de l'erreur de précision relative . pr IEEE-754 ⇒ $\eta(x) \leq \begin{cases} \beta^{1-p} & \text{si troncature} \\ 0{,}5\,\beta^{1-p} & \text{si arrondi} \end{cases}$ base 16 → 15 = 1 chiffre
- qd on veut juste convertir, on garde m précision (chiffres significatifs)

Chap 2 - Standard IEEE-754

- biais = $B = 2^{(q-1)} - 1$ pr base 2, avec q : bits de l'exposant ⇒ $E_{min} = -\text{biais} + 1$ et $E_{max} = \text{biais}$.
- binaire : biais = 127 (nb à enlever de l'exposant pr passer du standard au nb)
- $\eta(a+b) = \eta^I(a+b) + \eta^C(a+b) = \frac{a}{a+b}\eta^I(a) + \frac{b}{a+b}\eta^I(b) + \eta^C(a+b)$ I → erreur d'entrée C → erreur de calcul.

Chap 3 - Propagat° des erreurs → erreurs sur calculs numériques par ordinateur

- erreur d'1 opérat° ⊗ est dûe aux erreurs de représentat° de a et b et erreur de représentat° du résultat : $a \otimes b = fl(fl(a) \otimes fl(b))$
  ↳ erreur d'entrée = erreur de représentat° du résultat : $\eta^I(a \otimes b)$
  + erreur dûe au calcul : $\eta^C(a \otimes b)$
- SOMME $S = x_1 + x_2 + \dots$ ⇒ $\Delta^I S = \Delta^I x_1 + \Delta^I x_2 + \dots$ et $\eta^I(S) = \frac{x_1}{S}\eta^I(x_1) + \frac{x_2}{S}\eta^I(x_2) + \dots$
- PRODUIT $P = x_1 . x_2$ → $\Delta^I P = P.\left(\frac{\Delta^I x_1}{x_1} + \frac{\Delta^I x_2}{x_2}\right)$ et $\eta^I(P) \simeq \eta^I(x_1) + \eta^I(x_2)$
- DIVISION $Q = \frac{a}{b}$ ⇒ $\Delta^I Q = \frac{b\Delta^I a - a\Delta^I b}{b^2}$ et $\eta^I(Q) = \eta^I(a) - \eta^I(b)$
- RACINE : $\eta^I(\sqrt[k]{a^b}) = \frac{b}{k}\eta^I(a)$
- $\eta^C((a+b)+c) = \eta^C(a+b) + \eta^C(a+b+c)$
- qualité numérique d'1 fct f = nb condition : $K(x) = \dfrac{\left|\frac{f(x) - f(x')}{f(x)}\right|}{\left|\frac{x - x'}{x}\right|} \simeq \left|\frac{f'(x)}{f(x)} . x\right|$
- K petit ⇒ fct bien conditionnée (sinon mal conditionnée)
- K est évalué par stabilité de l'algorithme ⇒ si fct utilisées par algo ≠ bien conditionnées.
- erreur du résultat $y = [y_1, \dots, y_m]^T$ d'un algorithme avec n entrées $x = [x_1, \dots, x_n]^T$ est :

$$\Delta y = \begin{bmatrix} \Delta y_1 \\ \Delta y_m \end{bmatrix} \simeq \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1} & \dots & \frac{\partial \phi_1}{\partial x_n} \\ \frac{\partial \phi_m}{\partial x_1} & \dots & \frac{\partial \phi_m}{\partial x_n} \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta x_n \end{bmatrix} = J[\phi(x)] . \Delta x$$

- Erreur relative du résultat :

$$\eta(y) = \begin{bmatrix} \eta(y_1) \\ \eta(y_m) \end{bmatrix} \simeq \begin{bmatrix} K(\phi_1, x_1) & \dots & K(\phi_1, x_n) \\ K(\phi_m, x_1) & \dots & K(\phi_m, x_n) \end{bmatrix} \begin{bmatrix} \eta(x_1) \\ \eta(x_n) \end{bmatrix} = K(\phi, x) . \eta(x)$$

- $fl(x) = [fl(x_1) \dots, fl(x_n)]$. Algorithme : $\phi : \mathbb{R}^n \to \mathbb{R}^m$ avec $\phi = \phi^{(r)} \circ \dots \circ \phi^{(1)}$,
- déroulement d'1 algo : $x = x^{(0)} \to \phi^{(1)}(x^{(0)}) = x^{(1)} \to \dots \to \phi^{(r)}(x^{(r-1)}) = x^{(r)} = y$
- $fl(x^{(k+1)}) = fl\left(\phi^{(k+1)}(x^{(k)})\right) \Rightarrow \Delta x^{(k+1)} = fl(x^{(k+1)}) - x^{(k+1)}$
- 1 algo A est ⊗ crédible que A' si $Er(A, x) \leq Er(A', x)$.
- Algo numériquement stable ⇒ erreurs intermédiaires d'arrondi $H_k x^{(k)}$ ≠ du m ordre de grandeur que erreur inhérente.

Fascicule 2. Algèbre linéaire – Chap 1. Algèbre linéaire et perturbations

Vectorielles :

- norme : $\|x\| \geq 0$ et $\|x\|=0 \Leftrightarrow x=0$, $\|\alpha x\| = |\alpha|\|x\|$, $\|x+y\| \leq \|x\|+\|y\|$, $\|x-y\| \geqslant |\|x\|-\|y\||$
- $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \Rightarrow \|x\|_1 = \sum_{i=1}^{n} |x_i|$, $\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} = \sqrt{x^T x}$, $\|x\|_\infty = \max_{i=1\dots n} |x_i|$
- inégalité de Holder :

$$\sum_{i=1}^{n} |x_i y_i| \leq \|x\|_p \|y\|_q \Rightarrow \sum_{i=1}^{n} |x_i y_i| \leq \sqrt{\sum_{i=1}^{n} |x_j|} \cdot \sqrt{\sum_{i=1}^{n} |y_j|} \quad \text{(Cauchy avec } p=q=2\text{)}$$

$A = \begin{pmatrix} a_{11} & a_{1m} \\ a_{n1} & a_{nm} \end{pmatrix}$ $A(n,m)$

Matricielles :

- 4 définitions avec $x \to Ax$ sauf dernière. Si $\|AB\| \leq \|A\| \cdot \|B\| \Rightarrow$ norme sous-multiplicative
- $\|A\| = \sup_{x \in \mathbb{R}^n, \|x\|=1} \|Ax\| = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|}{\|x\|} \Rightarrow$ norme subordonnée : $\text{lub}(A) = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|}$
- $\|A\|_1 = \max_{j=1\dots n} \sum_{i=1}^{n} |a_{ij}|$ (somme colonnes) $\quad \|A\|_2 = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}$ (diag) $\quad \|A\|_\infty = \max_{i=1\dots n} \sum_{j=1}^{n} |a_{ij}|$ (somme lignes)
- Norme Frobenius : $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\text{tr}(A^t A)}$ $\quad A \in \mathbb{R}^{m \times n}$
- Un pb est bien conditionné si 1 petite variat° des entrées (données) n'entraine pas 1 grde variat° des sorties (résultats).
- Scalaires : $\eta(x) = \frac{\Delta x}{x} = \frac{m(x)-x}{x}$. Vecteurs : $\eta(x) = \frac{m(\|x\|_\infty) - \|x\|_\infty}{\|x\|_\infty}$. Conditionnement : $K(A) = \|A\| \cdot \|A^{-1}\|$
- A matrice carrée. $\lim_{k \to +\infty} A^k = 0$ et THÉO VON NEUMANN : matrice $I-A$ régulière $\Leftrightarrow (I-A)^{-1} = \sum_{k=0}^{\infty} A^k$
- Système linéaire : Si A et b ont perturbat°, $Ax=b \Rightarrow (A+\Delta A)(x+\Delta x) = b+\Delta b$

ob machine : $m(x) = x(1+\eta) = x+\eta(x)$ $|\eta| < \text{eps}$.

- erreur en retard pr 1 colonne de C : $m(C_j) = (A+\Delta A)b_j$ avec $\|\Delta A\| \leq \gamma_n \|A\|$ pour $C = A.B$

complexité : $C = A.B \Rightarrow c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \Rightarrow n^3$ multiplicat° et $n^2(n-1)$ additions $\Rightarrow$ complexité $O(n^3)$.

Définitions :
- appli linéaire $\Rightarrow f(\alpha u + \beta u') = \alpha f(u) + \beta f(u')$. appli. lin. de U dans V = forme linéaire
- A régulière $\Leftrightarrow A^{-1}$ existe (sinon singulière)
- A (carrée) orthogonale $\Leftrightarrow A^t A = I \Leftrightarrow A^{-1} = A^t$
- A symétrique $\Leftrightarrow A = A^t$. définie positive $\Leftrightarrow x^t A x > 0 \quad \forall x \in \mathbb{R}^n$
- Toute matrice régulière A peut se factoriser de manière unique $A = LU$ avec :

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \ell_{21} & 1 & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & 1 \end{pmatrix} \qquad R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & r_{nn} \end{pmatrix}$$

L unique si on met des 1 sur diagonale

- rayon spectral : $\rho(A) = \max \{|\lambda|, \lambda \text{ val propre de } A\}$
- $A(m,n) \Rightarrow n = \dim \text{Im}(A) + \dim \text{Ker}(A)$.

Chap 2. Méthodes de résolution de systèmes linéaires $Ax=b$ $x=?$

- diagonalisat° : $A = PDP^{-1}$
- factorisat° LU : $A = LU \Leftrightarrow$ on résout à la place 2 systèmes $Ly=b$ puis $Ux=y$.
- pour trouver L et U, factorisat° de GAUSS

Chap 3_ Méthodes itératives linéaires $Ax=b$

- on veut décomposer A en $A=M-N$.
- méthodes non relaxées: A peut s'écrire $A=D-E-F=D+L+U$ avec $A=\begin{bmatrix} D & F \\ -E & \ddots \end{bmatrix}=\begin{bmatrix} \ddots & U \\ L & D \end{bmatrix}$

	méthode	Décomp. $A=M-N$	Matrice $M^{-1}N$	Descriptᵒ d'une itération
par $A=D-E-F$	Jacobi	$A=D-(E+F)$	$J=M^{-1}N=D^{-1}(E+F)$ $=I-D^{-1}A$	$Dx^{(k+1)}=(E+F)x^{(k)}+b$
	Gauss-Seidel	$A=(D-E)-F$	$G=M^{-1}N=(D-E)^{-1}F$	$(D-E)x^{(k+1)}=Fx^{(k)}+b$
	Richardson	$A=I-(I-A)$	$R=M^{-1}N=I-A$	$x^{(k+1)}=(I-A)x^{(k)}+b$
par $A=D+L+U$	Jacobi	$A=D-(L+U)$	$J=M^{-1}N=-D^{-1}(L+U)$	$Dx^{(k+1)}=-(L+U)x^{(k)}+b$
	Gauss-Seidel	$A=(D+L)-U$	$G=M^{-1}N=(D+L)^{-1}U$	$(D+L)x^{(k+1)}=Ux^{(k)}+b$

↑ matrice d'itérat°

Fascicule 3_ Valeurs et vecteurs propres

Chap 1 - Valeurs et vecteurs propres

- Le pb algébrique des valeurs propres consiste en la solᵒ du système de $n+1$ éqᵒ linéaires $(A-\lambda B)u=0 \Rightarrow Au=\lambda Bu$.
- Si $B=I$, on obtient le pb des val propres: trouver $\lambda\in\mathbb{R}$ tq $Au=\lambda u$ avec $A\in\mathbb{R}^{n\times n}$, $u\in\mathbb{R}^n$, $u\neq 0$.
- $A(u)=Au$. P matrice de passage de $A\to B$ et $T: B\to A$. $B=T^{-1}AP$. Si $E=F \Rightarrow B=P^{-1}AP$.
- λ val propre de $A \Rightarrow \lambda^k$ val propre de A^k. Si A sym. déf. positive, toutes les val. propres sᵗ réelles positives.
- Si A nilpotente $(\exists k>0$ tq $A^k=0) \Rightarrow tr(A)=0$. $tr(A)=\sum \lambda_A$.
- diagonalisatᵒ: $D=P^{-1}AP$ est semblable à A, $A=PDP^{-1}$. Si val propres distinctes $\Rightarrow$ diagonalisable. Sinon regarder dim des sous-espaces propres.
- Théo de Schur: (trigonalisatᵒ) il $\exists$ une matrice unitaire U et $\wedge$ matrice triangulaire sup T tq $A=U^tTU$
- $A\mapsto B=U^TAU$ une transf. unitaire de similitude et A normale $\Rightarrow U^tAU=Diag(\lambda_1 \dots \lambda_n)$.
- Soit A une matrice carrée d'ordre n. On associe à chaq ligne i de la matrice le disque de Geršgorin, D_i dont le centre est l'élément diagonal a_{ii} et le rayon la somme des modules des autres éléments de la ligne: $D_i^{(e)}=\{z\in\mathbb{C} / |z-a_{ii}|\leq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\}$ où $R_i^{(e)}=\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|$ est le rayon de $D_i^{(e)}$
- Toutes les val. propres appartiennent à au ⊖ un des $\bigcup_{j\neq i}$ disques.
- u vect pr. de A et λ val. propre de $A \Leftrightarrow Au=\lambda u$.
- estimatᵒ des erreurs en cas de perturbatᵒ:
- on reprend u $(Au=\lambda u)$ et v tq $v^TA=\mu v^T$. les val. propres λ et μ sᵗ identiques. On a: $v^tu=0$ (orthogonaux).
- matrice perturbée: $A(\varepsilon)=A+\Delta A$, Si λ val propre de $A \Rightarrow |\lambda(\varepsilon)-\lambda| = O(\varepsilon^{1/m})$ où m= multiplicité de λ.
si λ est simple $\to \lambda=\lim_{|\varepsilon|\to 0}\frac{\lambda(\varepsilon)-\lambda}{\varepsilon}=\frac{v^tAu}{v^tu}$

Chap 2_ Décomposition en valeurs singulières

- $R\to$ image et $N\to$ noyau
- $Ker(A)=\{0\}$ ssi $rg\,A=n$
- E un espace. $F, G\subset E$ sᵗ complémentaires $\Leftrightarrow E=G+F$ et $G\cap F=\{\phi\} \Leftrightarrow E=G\oplus F$
- $E=G\oplus F$. Si $x=y+z$ avec $x\in E$, $y\in G$ et $z\in F$, alors y est la projectᵒ de x dans G et z celle de x dans F.
- projectᵒ orthogonale de x sur $Vect(u)$: $\frac{u}{\|u\|}\left(\frac{u}{\|u\|}\right)^t x$ dans $Vect(u^\perp) \Rightarrow P_{u^\perp}=I-P_u$.
projecteur orthogonal P_u de $Vect(u) \Rightarrow P_u=\frac{uu^t}{u^tu}$

- A inversible $\Leftrightarrow \det A \neq 0$ ($\Leftrightarrow$ A est une base possible de l'espace)

exemple de transformation :

$f_1 = e_2 \quad f_2 = e_1 + 2e_2 \quad f_3 = -e_1 - 3e_2 + e_3$

$$B = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 2 & -3 \\ 0 & 0 & 1 \end{pmatrix} \qquad P = B^t = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ -1 & -3 & 1 \end{pmatrix} \quad P^{-1} = \begin{pmatrix} -2 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \text{ et } A' = P^{-1}AP$$

- Décomposit° en val. singulières de A : on peut factoriser $A = U \Delta V^t$ avec : (DVS)
  - U vecteurs propres de AA^t
  - V vecteurs propres de A^tA
  - Δ matrice des val. singulières, $\Delta = \sqrt{\Lambda} = \sqrt{\text{diag}(\lambda_1, \dots \lambda_n)}$ val. propres non nulles de A^tA ou AA^t

- $B = \begin{pmatrix} 0 & A \\ A^t & 0 \end{pmatrix}$ $\Rightarrow$ val. sing de A st $\sigma_1 \dots \sigma_r$ ($\Leftrightarrow$ val. sing. non nulles de B st $\sigma_1, \dots, \sigma_r, -\sigma_1, \dots, -\sigma_r$.

- Soit $A = U \Delta V^T$ avec val. sing. $\sigma_1 \geq \dots \geq \sigma_r > 0 \Rightarrow \|A\| = \sigma_1$. Si $m = n$ et A régulière, $\|A^{-1}\| = \sigma_n$.
  ↳ $K(A) = \frac{\sigma_1}{\sigma_n}$.

Chap 3. Pseudoinverse et moindres carrés

- Pseudoinverse de $A = U \Delta V^T \Rightarrow A^+ = V \Delta^{-1} U^T$ Si Δ diagonale $\Rightarrow \Delta^{-1} = \begin{pmatrix} \frac{1}{a_{11}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{a_{nn}} \end{pmatrix}$
- $K^+(A) = \|A\| \|A^+\|$ (nb conditi° de A pour inversion gauche)
- Soit $Ax = b$ avec perturbat° $A(x + \Delta x) = b + \Delta b$. L'erreur relative est bornée par :

$$\frac{\|\Delta x\|}{\|x\|} \leq K^+(A) \frac{\|\Delta b\|}{\|b\|}$$

Partiel année dernière :

- ⊕ petits et ⊕ grand nb normalisés représentables. $B = 2^{q-1}$ (q bits exposant) $\Rightarrow x_{max} = 0|1\dots10|1\dots1 = 1,11\dots1 \times \beta^{\exp - B}$
  $x_{min} = 0|0\dots01|0\dots0$
- ⊕ grde erreur : $|\eta(x)| = \beta^{1-p}$ avec β = base et p = bits mantisse si troncature, si arrondi $\Rightarrow 0,5\beta^{1-p}$
- eps est le ⊕ petit positif tq $1 + eps \neq 1$: $1 \to 0|01\dots1|0\dots0 = 1,0\dots0 \times 2^{\exp - B}$
  $1 + eps \to 0|01\dots1|0\dots01 = 1,0\dots01 \times 2^{\exp - B}$
  $\Rightarrow eps = 0,00\dots01_2$
- calcul par ordi à précision finie. Dc dès que le terme dépasse la capacité de représentat° de la machine, il est converti à 0 et son ajout à la somme ne change pas la valeur de la somme. Il y a dc convergence numérique.

- $$A = \begin{pmatrix} x & x & & \\ x & \ddots & \ddots & 0 \\ & \ddots & \ddots & x \\ 0 & & x & x \end{pmatrix} \quad L = \begin{pmatrix} 1 & & & \\ x & \ddots & & 0 \\ & \ddots & \ddots & \\ 0 & & x & 1 \end{pmatrix} \quad U = \begin{pmatrix} x & x & & \\ & \ddots & \ddots & 0 \\ & & \ddots & x \\ 0 & & & x \end{pmatrix}$$

L est triang. inf. avec des 1 sur la diagonale et ne comporte qu'1 sous-diag. non nulle car hérite de la structure bande de A.
U est triang. sup et ne comporte qu'une sur diag. non nulle car idem.

- On peut alors calculer quelques u_{ij} et l_{ij}. On peut démontrer des résultats ($u_{ii} = \frac{i+1}{i}$) par récurrence, d'où $\det(A) = \det(LU) = \det(L) \cdot \det(U) = \prod_{i=1}^{n} \frac{i+1}{i} = n+1$.

- $$J = \frac{1}{2} \begin{bmatrix} 0 & 1 & 0 \\ 1 & \ddots & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

- ⊕ petite val propre λ_1 de A. On calcule $Au^{(i)} = \lambda_i u^{(i)}$
- conditionnement de A :

$$cond_2(A) = \frac{\|\lambda_1\|_2}{\|\lambda_n\|_2}$$

- méthode converge ? oui car consistante et $\rho(J) < 1$
- $\rho(J)^2 = \rho(G)$ car A tridiagonale. $\rho(G) < \rho(J)$ donc Gauss-Seidel est méthode ⊕ rapide.
- $A = \begin{pmatrix} 5 & 1 & \\ 0 & 6 & 1 \\ 1 & 0 & -5 \end{pmatrix}$ on cherche disques et on trouve 1 val. pr. ds le disque $(-5, 1)$ et 2 ds 2 disques qui se recoupent $(5, 1)$ et $(6; 1)$
- u vecteur propre si $Au = \lambda u$, dc vecteur propre doit être multiple.