

## 2. ANALYSE MATRICIELLE, NORMES

**2.1. Normes vectorielles, matricielles.** Une sémi-norme sur un espace vectoriel  $E$  est la donnée d'une application  $N : E \rightarrow \mathbb{R}$  vérifiant deux axiomes ( $X, Y$  vecteurs de  $E$ ,  $\lambda$  scalaire) :

- <sub>1</sub>  $N(\lambda X) = |\lambda|N(X)$ .
- <sub>2</sub>  $N(X + Y) \leq N(X) + N(Y)$ .

Où pour le scalaire  $\lambda$ , sa valeur absolue  $|\lambda|$  est la valeur absolue de  $\mathbb{R}$  ou le module de  $\mathbb{C}$  selon que corps des scalaires est  $\mathbb{R}$  ou  $\mathbb{C}$ .

Nous dirons qu'il s'agit d'une norme si  $N(X) > 0$  pour tout  $X \neq 0$ . Dans ce cas on notera  $N(X) = \|\cdot\|_*$  avec des indices "\*" lorsque l'on à faire à plusieurs normes dans le même texte.

Une norme détermine une structure d'espace metrique, en posant comme distance entre les vecteurs  $X$ , et  $Y$  la quantité  $N(X - Y)$ . La topologie est celle que vous avez déjà étudié. Dans le cas qui nous intéresse, le cas de dimension finie, cette distance fais de  $E$  un espace normé complet (de Banach).

Pour un espace de Hilbert,  $(E, \langle \cdot, \cdot \rangle)$  la quantité  $\sqrt{\langle X, X \rangle}$  détermine "la" norme.

Pour travailler avec des matrices, nous nous intéressons tout particulièrement aux espaces  $\mathbb{K}^n$  sur lesquels certaines normes sont standard : Si  $X = (x_1, x_2, \dots, x_n)^T$  posons

$$\|X\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}, \quad \|X\|_\infty = \max_j |x_j|.$$

Ces quantités sont visiblement des nombres réels bien définis, qui de plus vérifient les axiomes de norme, sauf éventuellemnt l'inégalité triangulaire, que l'on procède à démontrer.

**Lemme 3.** Pour  $1 \leq p \leq \infty$ , l'application  $X \mapsto \|X\|_p$  est une norme sur le  $\mathbb{K}$ -espace vectoriel  $\mathbb{K}^n$ . (Pour  $\mathbb{K}, \mathbb{R}$  ou  $\mathbb{C}$ ). Et nous remarquons les inégalités :

*Inégalité de Minkowski* :  $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$ .

*Inégalité de Hölder* : si  $1 \leq p \leq \infty$ ,  $|Y^* X| \leq \|X\|_p \|Y\|_q$

si l'on prend  $q$  tel que  $\frac{1}{p} + \frac{1}{q} = 1$ , lorsque  $1 < p < \infty$ , puis  $q = 1$  si  $p = \infty$  et enfin,  $q = \infty$  si  $p = 1$ . (C'est-à-dire, ici,  $\frac{1}{1} + \frac{1}{\infty} = 1$ ).

*Convexité de l'exponentielle* : si  $a, b \geq 0$  sont deux nombres réels, et  $1 < p < \infty$ ,

$$\frac{1}{p} + \frac{1}{q} = 1 \Rightarrow ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q.$$

**Démonstration.** Les énoncés sont fait dans l'ordre inverse de l'ordre de démonstration. Pour le dernier, qui est évident si  $ab = 0$ , on suppose que ni  $a$  ni  $b$  est nul. Nous savons que l'exponentielle réelle  $t \mapsto e^t$  est une fonction convexe, c'est-à-dire que si  $\theta_1 t_1 + \theta_2 t_2$  est un point qui se trouve entre  $t_1$  et  $t_2$  alors,  $e^{\theta_1 t_1 + \theta_2 t_2}$  est plus bas que  $\theta_1 e^{t_1} + \theta_2 e^{t_2}$ . C'est-à-dire que si  $0 \leq \theta_1, \theta_2$ , et  $\theta_1 + \theta_2 = 1$ , alors

$$e^{\theta_1 t_1 + \theta_2 t_2} \leq \theta_1 e^{t_1} + \theta_2 e^{t_2}.$$

De plus, par stricte convexité, l'on sait qu'il ne peut y avoir égalité que dans le cas où l'un des  $\theta$  est nul. On l'applique à :  $t_1 = \log(a^p), t_2 = \log(b^q), \theta_1 = \frac{1}{p}, \theta_2 = \frac{1}{q}$ , ce qui donne :

$$e^{\frac{1}{p} \log(a^p) + \frac{1}{q} \log(b^q)} \leq \frac{1}{p} e^{\log(a^p)} + \frac{1}{q} e^{\log(b^q)}.$$

qui évidemment est l'inégalité annoncé.

L'inégalité d'Hölder se déduit facilement, dans le cas  $1 < p < \infty$ . Puisqu'elle est évidente lorsque l'un des vecteurs est nul, supposons le contraire et posons :  $\tilde{X} = \frac{1}{\|X\|_p} X$  et  $\tilde{Y} = \frac{1}{\|Y\|_q} Y$ , ces vecteurs sont de norme 1 et vérifient :

$$|\tilde{Y}^* \tilde{X}| = \frac{|Y^* X|}{\|X\|_p \|Y\|_q}.$$

En effet, notant  $\tilde{X} = (a_i)^T$  et  $\tilde{Y} = (b_i)^T$  nous avons

$$|\tilde{Y}^* \tilde{X}| = \left| \sum_{i=1}^n \bar{b}_i a_i \right| \leq \sum_{i=1}^n |b_i| |a_i|.$$

Somme de produits de nombres réels positifs où nuls auxquels on applique l'inégalité de convexité de l'exponentielle. Ainsi, pour chaque  $i$  nous avons :  $|b_i| |a_i| \leq \frac{1}{p} |a_i|^p + \frac{1}{q} |b_i|^q$ . En additionnant,

$$\sum_{i=1}^n |b_i| |a_i| \leq \frac{1}{p} \sum_{i=1}^n |a_i|^p + \frac{1}{q} \sum_{i=1}^n |b_i|^q = \frac{1}{p} \|\tilde{X}\|_p^p + \frac{1}{q} \|\tilde{Y}\|_q^q = \frac{1}{p} + \frac{1}{q} = 1.$$

Ainsi,  $\frac{|Y^* X|}{\|X\|_p \|Y\|_q} \leq 1$ , comme il fallait démontrer.

Il faut traiter le cas de la norme infini séparément. Or  $|Y^* X| = \left| \sum \bar{y}_i x_i \right| \leq \sum |y_i| |x_i| \leq |y_{i_0}| \sum |x_i| = |y_{i_0}| \|X\|_1$ , pour  $|y_{i_0}|$  le plus grand des  $|y_i|$ . Or celui-ci vaut  $\|Y\|_\infty$ , et le lemme est démontré (en remarquant la symétrie).

Enfin, pour l'inégalité triangulaire, on procède à l'aide d'une astuce de Minkowsky. Pour  $1 < p < \infty$  nous observons que  $\|X + Y\|_p^p = \sum |x_i + y_i| |x_i + y_i|^{p-1} \leq \sum |x_i| |x_i + y_i|^{p-1} + \sum |y_i| |x_i + y_i|^{p-1}$ . On utilise Hölder pour les couples de vecteurs  $x = (|x_i|)_i$  et  $z = (|x_i + y_i|^{p-1})_i$  et  $y = (|y_i|)_i$  et  $z = (|x_i + y_i|^{p-1})_i$  ce qui nous donne les inégalités suivantes :

$$\sum |x_i| |x_i + y_i|^{p-1} \leq \|X\|_p \|z\|_q; \quad \sum |y_i| |x_i + y_i|^{p-1} \leq \|Y\|_p \|z\|_q.$$

Il reste à constater que  $\|z\|_q = \|X + Y\|_p^{p-1}$  parce que  $q(p-1) = p$ . Ainsi, nous avons établi,

$$\|X + Y\|_p^p \leq \|X\|_p \|X + Y\|_p^{p-1} + \|Y\|_p \|X + Y\|_p^{p-1}.$$

En factorisant  $\|X + Y\|_p^{p-1}$  sur le membre de droite et divisant, nous avons l'inégalité triangulaire cherchée.  $\diamond$

Evidement que l'on peut identifier  $M_{n,m}(\mathbb{K})$  avec  $\mathbb{K}^{mn}$  et utiliser une quelconque des normes ci-dessus, pour parler de la norme d'une matrice, **mais** cette démarche n'est pas très utile (sauf pour penser à la topologie). En effet, la propriété intéressante serait  $\|AB\| \leq \|A\| \|B\|$ . Par exemple, pour la "norme" du maximum, qui donnerait à une matrice le maximum des modules de ses éléments, nous avons que la matrice pleine de 1 (tout élément est le nombre 1) en taille  $n$ , à norme 1 mais évidemment la "norme" de son carré est  $n$  et  $n \not\leq 1$  en général!

**Définition 3.** *Norme matricielle* Une norme matricielle sur  $M_n(\mathbb{K})$  est une application  $\|\cdot\| : M_n(\mathbb{K}) \rightarrow \mathbb{R}_+$  telle que

$$\begin{aligned} \|A\| &= 0 \Leftrightarrow A = 0, \\ \|\lambda A\| &= |\lambda| \|A\|, \\ \|A + B\| &\leq \|A\| + \|B\|, \\ \|AB\| &\leq \|A\| \|B\|. \end{aligned}$$

C'est donc juste une norme pour les matrices considérées comme vecteurs, qui vérifie en plus  $\|AB\| \leq \|A\| \|B\|$ . En particulier, nous remarquons que la norme de la matrice identité vérifie,  $\|I\| \geq 1$ . (Car nous pouvons diviser par  $\|I\|$  dans  $\|I\| = \|I^2\| \leq \|I\|^2$ ).

**Proposition 1.** *Soit  $\|\cdot\|$  une norme matricielle sur  $M_n(\mathbb{K})$ , alors*

$$\begin{aligned} \|I\| &\geq 1, \quad \|A^n\| \leq \|A\|^n \quad (n > 0), \quad 1 \leq \|A^{-1}\| \|A\|. \\ \|A\| &< 1, \implies I - A \in GL_n(\mathbb{K}). \end{aligned}$$

Si  $U \in GL_n(\mathbb{K})$ , la boule ouverte de centre  $U$  et rayon  $\|U^{-1}\|^{-1}$  est contenue dans  $GL_n(\mathbb{K})$  qui est donc ouvert.

**Démonstration.** Evident,  $\|I\| = \|U^{-1}U\| \leq \|U^{-1}\| \|U\|$ . Puis, remarquons que

$$\frac{1}{1-x} = \sum_{n=0}^{+\infty} x^n, \quad (1-x) \left( \sum_{n=0}^N x^n \right) = 1 - x^{N+1}.$$

Ainsi si  $1 < p < q$  nous avons

$$\left\| \sum_{n=0}^q A^n - \sum_{n=0}^p A^n \right\| \leq \sum_{n=p+1}^q \|A^n\| \leq \sum_{n=p+1}^q \|A\|^n,$$

qui, lorsque  $\|A\| < \rho < 1$  est majoré par

$$\leq \|A\|^p \sum_{n=1}^{+\infty} \rho^n = \|A\|^p \frac{\rho}{1-\rho}.$$

Ainsi, si  $\varepsilon > 0$ , il existe  $p$  tel que  $\|A\|^p < \varepsilon \frac{1-\rho}{\rho}$  ce qui prouve que la série converge. Puisque  $(1-A) \left( \sum_{n=0}^N A^n \right) = I - A^{N+1} \rightarrow I$  ( $n \rightarrow +\infty$ ) La matrice  $I - A$  est inversible, et son inverse est la série de Von Newman

$$\frac{1}{I-A} = \sum_{n=0}^{+\infty} A^n.$$

Enfin posons  $C = U - U + B = U(I - (I - U^{-1}B))$  qui est inversible si et seulement si,  $I - C$  l'est pour  $C = I - U^{-1}B$  l'est. Or il suffit pour cela, que  $\|C\| = \|U^{-1}(U - B)\| \leq \|U^{-1}\| \|U - B\| < 1$  et pour cela il suffit que  $\|A - U\| < \|U^{-1}\|^{-1}$ .  $\diamond$

**Définition 4. Norme matricielle subordonnée (d'opérateur)** Une norme matricielle sur  $M_n(\mathbb{K})$ ,  $\|\cdot\|$ , est dite subordonnée, s'il existe une norme  $N$  sur  $\mathbb{K}^n$  (et l'on dira que  $\|\cdot\|$  est subordonnée à  $N$  ou que  $\|\cdot\|$  est la norme d'opérateur de l'espace normé  $(\mathbb{K}^n, N)$ ). si pour toute matrice  $A$  :

$$\|A\| = \max_{X \in \mathbb{K}^n, N(X)=1} N(AX).$$

On remarque que dans  $\mathbb{K}^n$  toutes les normes sont équivalentes, et ainsi  $X \mapsto AX$  est continu en norme  $N$  et comme  $N(X) = 1$  détermine un compact, le sup est un maximum et il est atteint en un vecteur de norme 1.

**Lemme 4.** Si  $\|\cdot\|$  est une norme sur  $\mathbb{K}^n$ , il existe une norme (unique, que l'on notera aussi  $\|\cdot\|_?$ ) sur  $M_n(\mathbb{K})$  qui lui soit subordonnée. De plus nous avons

$$\|A\| = \max_{X \neq 0} \frac{\|AX\|}{\|X\|}$$

et toujours  $\|I\| = 1$  puis  $\|AX\| \leq \|A\| \|X\|$ .

**Démonstration.** La formule pour  $\|A\|$  doit être

$$\|A\| = \max_{X \in \mathbb{K}^n, \|X\|=1} \|AX\|$$

si l'on veut qu'elle soit subordonnée à  $\|\cdot\|$ , d'où l'unicité. Il ne reste à démontrer que cette identité définit bien une norme matricielle!

D'une part  $(A+B)X = AX + BX$  et  $\|(A+B)X\| \leq \|AX\| + \|BX\|$  d'où, pour  $X$  parcourant la sphère unité en norme  $\|\cdot\|$  de  $\mathbb{K}$ ,  $\max \|(A+B)X\| \leq \max(\|AX\| + \|BX\|) \leq \max \|AX\| + \max \|BX\|$ . Qui est l'inégalité triangulaire.

L'homogénéité est évidente. De plus, si  $\|A\| = 0$ , pour tout  $X$  de norme 1,  $AX$  est de norme 0 c'est-à-dire nul. La matrice  $A$  est donc nulle, tout vecteur de  $\mathbb{K}$  est, à une dilatation près, de norme 1.

Avant de montrer qu'il s'agit d'une norme matricielle, montrons la formule alternative pour la norme d'opérateur. Ce qui est clair est que

$$\max_{X \in \mathbb{K}^n, \|X\|=1} \|AX\| \leq \sup_{X \neq 0} \frac{\|AX\|}{\|X\|}.$$

Soit  $X \neq 0$  et  $Y = \frac{X}{\|X\|}$ ,  $\|Y\| = 1$  nous avons  $\|AY\| = \frac{\|AX\|}{\|X\|}$ . Ainsi  $\sup_{X \neq 0} \frac{\|AX\|}{\|X\|} = \max_{X \neq 0} \frac{\|AX\|}{\|X\|} \leq \max_{X \in \mathbb{K}^n, \|X\|=1} \|AX\|$ . En plus, si  $X \neq 0$ ,  $\frac{\|AX\|}{\|X\|} \leq \|A\|$ , d'où  $\|AX\| \leq \|A\| \|X\|$  le cas du vecteur nul étant évident. Puisque  $IX = X$ ,  $\|I\| = 1$  lorsque  $\|\cdot\|$  est subordonnée.

Avec cette formule, il est aisé de prouver que  $\|AB\| \leq \|A\| \|B\|$ . En effet, soit  $X$  un vecteur de norme 1, alors  $\|ABX\| \leq \|A\| \|BX\| \leq \|A\| \|B\| \|X\| = \|A\| \|B\|$  ainsi  $\max_{\|X\|=1} \|ABX\| \leq \|A\| \|B\|$ .  $\diamond$

**Proposition 2** (calcul des normes subordonnées aux normes de Hölder). *Soit  $A = ((a_{ij}))$  une matrice carrée. Pour la norme sup :*

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|, \quad \text{max des norme-1 des lignes.}$$

Pour la norme  $\|\cdot\|_1$  de  $\mathbb{C}$  (ou  $\mathbb{R}$ ) :

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}| = \|A^*\|_\infty, \quad \text{max des norme-1 des colonnes.}$$

Pour la norme euclidienne de  $\mathbb{C}$  :

$$\|A\|_2 = \sqrt{\rho(A^*A)} = \|A^*\|_2, \quad \text{racine du rayon spectral de } AA^*.$$

**Définition 5. Norme de Frobenius** Soit  $A \in M_{mn}(\mathbb{K})$  nous appellerons norme de Frobenius la quantité

$$\|A\|_F := \left( \sum_{i=1, j=1}^{m, n} |a_{ij}|^2 \right)^{\frac{1}{2}}.$$

**Proposition 3** (Propriétés de la norme de Frobenius). *Soient  $A, B \in M_{mn}(\mathbb{K}), C \in M_{nt}(\mathbb{K}), X \in \mathbb{K}^n$ .*

1- La forme  $(A/B)_F = \text{trace}(B^*A)$  donne une structure hilbertienne à  $M_{mn}(\mathbb{K})$  de norme associée la norme de Frobenius.

2- Nous avons  $\|AC\|_F \leq \|A\|_2 \|C\|_F$ ,  $\|AC\|_F \leq \|A\|_F \|C\|_2$ ,  $\|AX\|_F \leq \|A\|_F \|X\|_2$ .

3- Puis  $\|A\|_2 \leq \|A\|_F \leq \sqrt{n} \|A\|_2$ ,  $\|AC\|_F \leq \|A\|_F \|C\|_F$ .

4- Si  $U, V$  sont unitaires,  $\|UA\|_F = \|A\|_F = \|AV\|_F$ .

**Proposition 4** (Théorème de Householder). *Soit  $A \in M_n(\mathbb{C})$  et  $\varepsilon > 0$ , il existe alors une norme vectorielle sur  $\mathbb{C}^n$  telle que pour la norme matricielle subordonnée, l'on ait*

$$\|A\| \leq \rho(A) + \varepsilon.$$

Où  $\rho(\cdot)$  est le rayon spectral.

On en déduit  $\rho(A) = \inf\{\|A\| / \|\cdot\| \text{ est une norme matricielle de } M_n(\mathbb{C})\}$ .

**Démonstration.** Fixons  $\varepsilon$  et  $A$ . La matrice  $A$  se triangularise, soit  $P \in GL_n(\mathbb{C})$ ,  $T = P^{-1}AP$  triangulaire supérieure,  $\rho(A) = \rho(T)$ . Soit  $Q(\mu) = \text{diag}(1, \mu, \mu^2, \dots, \mu^{n-1}) \in GL_n(\mathbb{C})$  si  $\mu > 0$ . Considérons la norme de  $\mathbb{C}^n$   $\|X\|_* := \|Q(\mu)PX\|_2$ . La norme d'opérateur induite par cette norme est  $\|B\|_* = \|Q(\mu)PBQ(\mu)^{-1}\|_2$  si  $B \in M_n(\mathbb{C})$ . On trouvera  $\mu$  de sorte que cette norme d'opérateur soit celle que nous cherchons. La matrice  $Q(\mu)PAP^{-1}Q(\mu)^{-1} = Q(\mu)TQ(\mu)^{-1}$  est triangulaire supérieure et elle a comme diagonale principale la même matrice diagonale  $D$  que celle de  $T$ . L'élément en

ligne  $i$  colonne  $j$  de cette matrice est  $\mu^{i-j}t_{ij}$ . Nous avons donc  $\lim_{\mu \rightarrow +\infty} Q(\mu)TQ(\frac{1}{\mu}) = D$  d'où  $\lim_{\mu \rightarrow +\infty} \|A\|_* = \|D\|_2$  Fixons  $\mu$  assez grand pour que  $\|A\|_* < \|D\|_2 + \varepsilon$ ,

$$\|A\|_* < \|D\|_2 + \varepsilon = \sqrt{\rho(D^*D)} + \varepsilon = \max |t_{ii}| + \varepsilon = \rho(A) + \varepsilon.$$

## 2.2. Localisation des valeurs propres.

**Théorème 7** (Théorème de Gerschgorin-Hadamard). *Soit  $A = (a_{ij}) \in M_n(\mathbb{C})$ . Nous avons toujours*

$$sp(A) \subset \bigcup_{i=1}^n D_i$$

pour les disques de Gerschgorin :  $D_i = \{z \in \mathbb{C} / |z - a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}|\}$

**Démonstration.** Soit  $\lambda \in sp(A)$ , et  $X$  un vecteur propre associé. Posons  $Y = \frac{X}{\|X\|_\infty}$  qui est aussi un vecteur propre associée,  $\|Y\|_\infty = 1$ . Nous avons, donc, qu'il existe un indice, noté  $l \in [1 : n]$  telle que  $|y_l| = 1$  et évidemment  $(AY)_l = \lambda y_l$  d'où l'on tire que

$$\sum_{j=1}^n a_{lj}y_j = \lambda y_l,$$

où précisément,  $\sum_{j=1, j \neq l}^n a_{lj}y_j + a_{ll}y_l = \lambda y_l$  qui donne  $y_l(\lambda - a_{ll}) = \sum_{j=1, j \neq l}^n a_{lj}y_j$ . Or  $|y_j| \leq |y_l| \leq 1$  qui par triangulaire donne

$$|\lambda - a_{ll}| \leq |y_l(\lambda - a_{ll})| \leq \sum_{j=1, j \neq l}^n |a_{lj}|.$$

C'est-à-dire que  $\lambda$  est dans le disque  $D_l$ .

mèD

**Corollaire 1.** *Pour une matrice  $A = (a_{ij})$  nous avons la borne du rayon spectral suivante*

$$\rho(A) \leq \sup_{1 \leq i \leq n} \left( \sum_{j=1}^n |a_{ij}| \right).$$

**Démonstration.** Si  $\lambda \in sp(A)$  il existe  $1 \leq i \leq n$  tel que  $\lambda \in D_i$  c'est-à-dire  $|\lambda - a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}|$ . Ainsi, puisque  $|\lambda| - |a_{ii}| \leq |\lambda - a_{ii}|$  nous avons  $|\lambda| - |a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}|$  et en passant  $|a_{ii}|$  au deuxième membre,

$$|\lambda| \leq \sum_{j=1}^n |a_{ij}|$$

le maximum en  $i$  est donc un majorant.

mèD

**Définition 6.** *Une matrice  $A = (a_{ij}) \in M_n(\mathbb{C})$  est dite "à diagonale dominante" si pour chaque  $i \in [1 : n]$  nous avons*

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

*Elle sera dite "à diagonale dominante stricte" si toutes ces égalités sont strictes.*

Si la matrice  $A$  est à diagonale dominante stricte, il s'en suit qu'aucun des  $n$  disques de Gerschgorin ne peut contenir 0 d'où  $0 \notin sp(A)$ .

**Corollaire 2.** *Une matrice à diagonale dominante stricte est inversible.*

**On remarquera** que lorsque l'on note  $A_p$  la matrice tirée de la matrice  $A$  en ne retenant que les éléments dont les deux indices sont inférieurs ou égaux à  $p$  :  $A_p = A_{[1:p][1:p]}$ ; si la matrice  $A$  est à diagonale dominante (resp. stricte) il en va de même pour  $A_p$ . Ainsi pour une matrice à diagonale dominante stricte  $A$ , chaque  $A_p$  est inversible.

**2.3. Conditionnement d'une matrice.** Quand on veut résoudre numériquement  $AX = B$  avec  $A \in M_n(\mathbb{C})$ ,  $1 \ll n$ , divers facteurs influent sur la précision du résultat :

- Incertitudes sur les données (expérimentales) de  $A$  et/ou  $B$ .
- Erreur dans la représentation des coefficients de  $A$  et  $B$  avec un nombre fini de décimales (ordinateur).
- Erreurs d'arrondi lors des calculs.
- Erreurs prévisibles de l'algorithme, notamment lors des calculs approchés par des méthodes itératives.

Supposons d'abord une méthode donnée, de calcul de  $X$  pour le problème  $AX = B$ . Par des erreurs de représentation, nous résolvons en fait  $(A + \delta A)Y = B + \delta B$  avec  $\delta A$  et  $\delta B$  des quantités (matricielles) inconnues. On supposera posséder une borne du type  $\|\delta A\| \leq 2^{-N}$  tenant compte de la précision acquise dans l'obtention des données et leur codage. Évidemment nous pouvons considérer que le résultat du calcul  $Y$  est lié avec le véritable résultat cherché  $X$  par  $Y = X + \delta X$  et il s'agit de discuter des informations à priori sur "l'erreur absolue"  $\delta X$  commise. Nous nous intéressons à sa norme, ou tout du moins à l'erreur relative en norme  $\frac{\|\delta X\|}{\|X\|}$ .

**Exemple extrême.** Supposons les données fournies par l'expérimentation suivantes :

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \quad B = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 21 \end{pmatrix} \Rightarrow X = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

ce qui se vérifie sans peine. L'expérimentateur nous avertit qu'il y a des incertitudes, qui ajoutées à la qualité de notre méthode (de codage et calcul) se resument à dire que l'on résout  $(A + \delta A)Y = B + \delta B$  avec

$$\delta A = \begin{pmatrix} 0 & 0 & 0,1 & 0,2 \\ 0,08 & 0,4 & 0 & 0 \\ 0 & -0,2 & -0,11 & 0 \\ 0,01 & -0,01 & 0 & -0,08 \end{pmatrix} \quad \delta B = \begin{pmatrix} 0,01 \\ -0,01 \\ 0,01 \\ -0,01 \end{pmatrix}$$

les calculs donnent pour solution de  $(A + \delta A)Y = B + \delta B$  un

$$Y = \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}$$

qui est vraiment trop trop éloigné de la valeur de  $X$ . Il s'agit de comprendre ce phénomène.

**Proposition 5** (Perturbation du deuxième membre). *Soit  $X \in \mathbb{C}^n$  la solution de  $AX = B$  pour  $A \in M_n(\mathbb{C})$  inversible et  $0 \neq B \in \mathbb{C}^n$  donnés. Notons  $X + \delta X$  la solution  $Y$  de  $AY = B + \delta B$  pour chaque  $\delta B \in \mathbb{C}^n$ . Si  $\|\cdot\|$  est une norme matricielle subordonnée, alors :*

$$\frac{\|\delta X\|}{\|X\|} \leq \|A\| \cdot \|A^{-1}\| \frac{\|\delta B\|}{\|B\|}.$$

De plus, pour  $A$  matrice inversible,  $R, \varepsilon > 0$  donnés, il existe  $B$  de norme égale à  $R$  et un  $\delta B$  de norme inférieure à  $\varepsilon$  pour lesquels il y a égalité.

La quantité  $\|A\| \cdot \|A^{-1}\|$  est dite "le nombre de conditionnement de la matrice  $A$  dans la norme  $\|\cdot\|$ " et sera noté  $c_{\|\cdot\|}(A)$  ou  $c(A)$  lorsque aucune ambiguïté n'est à craindre. Nous avons donc

$$\frac{\|\delta X\|}{\|X\|} \leq c(A) \cdot \frac{\|\delta B\|}{\|B\|}.$$

**Démonstration.** Puisque  $AX + A\delta X = A(X + \delta X) = B + \delta B = AX + \delta B$ ,  $\delta X = A^{-1}\delta B$  et

$$\|\delta X\| \leq \|A^{-1}\| \cdot \|\delta B\|.$$

Mais  $B = AX$ , d'où  $\|B\| \leq \|A\| \cdot \|X\|$  et  $0 < \frac{\|B\|}{\|A\|} \leq \|X\|$ . Ainsi  $\frac{1}{\|X\|} \leq \frac{\|A\|}{\|B\|}$  et

$$\frac{\|\delta X\|}{\|X\|} \leq \frac{\|A^{-1}\| \cdot \|\delta B\|}{\|X\|} \leq \|A^{-1}\| \cdot \|\delta B\| \frac{\|A\|}{\|B\|}.$$

Pour l'optimalité, considérons  $X_0$  tel que  $\|AX_0\| = \|A\| \cdot \|X_0\|$ ,  $B := AX_0$ . On pourra prendre  $B$  de la norme que l'on veut. Considérons ensuite  $Z_0$  tel que  $\|A^{-1}Z_0\| = \|Z_0\|$  et quitte à le multiplier  $\delta B := 10^{-N}Z_0$  nous avons un  $\delta B$  de norme aussi petite que l'on veut, et c'est facile de vérifier que pour ces données, l'erreur relative de la solution calculée est exactement le produit du nombre de conditionnement par l'erreur relative du deuxième membre. L'inégalité est optimale

**On remarquera** que l'on peut dire que l'erreur relative dans la solution d'un système linéaire n'est pas "pire" que celle du deuxième membre seulement lorsque le conditionnement vaut 1. Le nombre de conditionnement est donc *un facteur d'amplification* des erreurs.

**Une matrice est d'autant mieux conditionnée, que son nombre de conditionnement approche 1. Toujours  $1 \leq c(A)$ .**

**Une matrice  $A$  est bien conditionnée si  $c(A) = 1$ . Les matrices unitaires sont bien conditionnées en norme spectrale, ou de Frobenius.**

**Proposition 6** (Perturbation de la matrice). *Soit  $X \in \mathbb{C}^n$  la solution de  $AX = B$  pour  $A \in M_n(\mathbb{C})$  inversible et  $0 \neq B \in \mathbb{C}^n$ . Notons  $X + \delta X$  la solution  $Y$  de  $(A + \delta A)Y = B$  alors*

$$\frac{\|\delta X\|}{\|X + \delta X\|} \leq c(A) \cdot \frac{\|\delta A\|}{\|A\|}.$$

**Démonstration.** Nous avons  $AX + A\delta X + \delta A(X + \delta X) = B = AX$  d'où  $A\delta X = -\delta A(X + \delta X)$ . Ainsi  $\delta X = -A^{-1}\delta A(X + \delta X)$  et  $\|\delta X\| = \|A^{-1}\delta A(X + \delta X)\| \leq \|A^{-1}\| \cdot \|\delta A\| \cdot \|(X + \delta X)\|$  En divisant par  $\|X + \delta X\|$  et exprimant  $\|A^{-1}\| = \frac{c(A)}{\|A\|}$  on a conclu. Evidemment, lorsque  $B$  est non nul, la solution  $Y$  ne peut être nulle.