M2 Agrégation - UE 8.

TD : Analyse Numérique matricielle.

Liste non exhaustive des points **à connaître :**

- *Matrices et propriétés classiques.*
- *Réduction des matrices*
- *Décomposition en valeurs singulières*
- *Valeurs propres des matrices, rayon spectral et propriétés, localisation des valeurs propres.*
- *Matrices à diagonale dominante, strictement dominante etc...*
- *Normes matricielles.*
- *Conditionnement d'une matrice, d'un système linéaire.*
- *Quelques problèmes de modélisation amenant à résoudre des systèmes linéaires.*
- *Propriétés de convergence des suites de vecteurs ou matrices suivant le rayon spectral.*

$\sim\sim\sim$

Au niveau de la résolution proprement dite :

- *Méthodes directes de résolution : Élimination de Gauss, Factorisation LU, Cholesky, QR.*
- *Cas d'une matrice réelle symétrique définie positive, lien avec les méthodes de gradient.*
- *Cas des matrices tridiagonales.*
- *Notion de nombre d'opérations.*
- *Méthodes itératives et leur convergence : Jacobi, Gauss-Seidel, relaxation.*
- *Savoir traiter notamment l'exemple de la matrice des différences finies du laplacien.*
- *Moindres carrés (cf. partie optimisation)*

On se donne $n \in \mathbb{N}^*$ et $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. On s'intéresse à la résolution de systèmes linéaires du type $Ax = b$ où $A \in \mathcal{M}_n(\mathbb{K})$, $x \in \mathbb{K}^n$ et $b \in \mathbb{K}^n$.

I. Origine des erreurs et conditionnement de matrices et de systèmes linéaires [LT1], [QSS], [Fi], [Ci], [S]

Pour résoudre un système linéaire, on a rarement la solution exacte sans erreurs. Les sources d'erreurs peuvent être multiples, par exemple : incertitudes sur les coefficients de A et b, erreurs d'arrondis, erreur de précision machine etc... Tout cela fait qu'en pratique on résout plutôt un système linéaire perturbé :

$$(A + \Delta A)(x + \delta x) = (b + \delta b). \tag{1}$$

I.1. **Un exemple classique (R.S. Wilson)**

On s'intéresse à la résolution du système linéaire $Ax = b$ avec :

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \text{ et } b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

A est inversible et on connaît la solution de ce système : c'est $x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$. On considère alors les perturbations suivantes : une perturbation sur les coefficients de la matrice

$$A + \Delta A = \begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix} \text{ c'est à dire } \Delta A = \begin{pmatrix} 0 & 0 & 0.1 & 0.2 \\ 0.08 & 0.04 & 0 & 0 \\ 0 & -0.02 & -0.11 & 0 \\ -0.01 & -0.01 & 0 & -0.02 \end{pmatrix}$$

et une perturbation sur les coefficients du second membre

$$b + \delta b = \begin{pmatrix} 32.01 \\ 22.99 \\ 33.01 \\ 30.99 \end{pmatrix} \text{ c'est à dire } \delta b = \begin{pmatrix} 0.01 \\ -0.01 \\ 0.01 \\ -0.01 \end{pmatrix}.$$

La solution de $Ay = b + \delta b$ est donnée par

$$y = \begin{pmatrix} 1.82 \\ -0.36 \\ 1.35 \\ 0.79 \end{pmatrix} \text{ c'est à dire } \delta x = y - x = \begin{pmatrix} 0.82 \\ -1.36 \\ 0.35 \\ -0.21 \end{pmatrix}.$$

La solution de $(A+\Delta A)z = b$ est donnée par

$$z = \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix} \text{ c'est à dire } \Delta x = z - x = \begin{pmatrix} -82 \\ 136 \\ -35 \\ 21 \end{pmatrix}.$$

Commentaires ?

I.2. **Conditionnement d'une matrice**

I.2.1. *Définition et premières propriétés*

Définition I.1. *Soit $|||\cdot|||$ une norme matricielle subordonnée à une norme sur $\mathbb{K}^n$. Le conditionnement d'une matrice inversible A, associé à cette norme, est le nombre réel :*

$$\text{cond}(A) := |||A|||\,|||A^{-1}|||.$$

On notera en particulier pour $p \in \mathbb{N}^*$, $\text{cond}_p(A) = |||A|||_p\,|||A^{-1}|||_p$, avec $|||A|||_p = \sup_{x\neq 0} \dfrac{\|Ax\|_p}{\|x\|_p}$ la norme matricielle subordonnée à la norme l^p.

Ex 1. *Quelques propriétés du conditionnement*

1) Montrer que $\text{cond}(A) = \text{cond}(\alpha A)$ pour toute matrice A inversible et tout scalaire $\alpha \neq 0$.

2) Montrer que $\text{cond}(A) = \text{cond}(A^{-1})$, pour toute matrice A inversible.

3) Montrer que $\text{cond}(A) \geq 1$.

4) Montrer que $\text{cond}_2(A) = \dfrac{\mu_{max}(A)}{\mu_{min}(A)}$ où $\mu_{max}(A)$ et $\mu_{min}(A)$ sont respectivement la plus grande et la plus petite valeur singulière de A, i.e. respectivement la racine carrée positive de la plus petite et la plus grande valeur propre de la matrice AA^*.

5) Si A est normale (i.e. $AA^* = A^*A$), montrer que $\text{cond}_2(A) = \dfrac{|\lambda_{max}(A)|}{|\lambda_{min}(A)|}$, où $\lambda_{max}(A)$ et $\lambda_{min}(A)$ sont respectivement la plus grande et la plus petite valeur propre en module de A.

6) Montrer que $\text{cond}_2(A) = 1$ si et seulement si $A = \alpha Q$ où α est un scalaire et Q une matrice unitaire.

On dira qu'un système linéaire est *bien conditionné* si le conditionnement de sa matrice est proche de 1. Si son conditionnement est très grand, on dit que le système est *mal conditionné*.

Exemple : Un problème mal conditionné avec la matrice de Hilbert cf. [LT1] p. 164-166 et [S].

I.3. **Influence du conditionnement sur les erreurs**

Ex 2. *Erreurs et conditionnement*

On considère le système $Ax = b$ avec A inversible.

1) On s'intéresse au système perturbé $(A + \Delta A)z = (A + \Delta A)(x + \Delta x) = b$. Montrer que

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \mathrm{cond}(A) \frac{|||\Delta A|||}{|||A|||}.$$

2) On s'intéresse alors à un autre système perturbé $Ay = A(x + \delta x) = b + \delta b$. Montrer que

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A) \frac{\|\delta b\|}{\|b\|}.$$

3) Notons x^* une solution approchée, $e := x - x^*$, l'erreur commise et $r := b - Ax^*$, le résidu. Montrer que

$$\frac{1}{\mathrm{cond}(A)} \frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq \mathrm{cond}(A) \frac{\|r\|}{\|b\|},$$

où $\|\cdot\|$ est une norme vectorielle sur $\mathbb{K}^n$ et cond() le conditionnement qui lui est associé.

Remarque : Ces inégalités sont optimales au sens où il existe des matrices A, des vecteurs b et des perturbations ΔA et δb pour lesquels les inégalités sont des égalités.

Ex 3. *Erreurs relatives et conditionnement*

On considère

$$A = \begin{pmatrix} 1 & 100 \\ 0 & 1 \end{pmatrix}.$$

1) Calculer $\mathrm{cond}_1(A)$.

2) On considère alors les systèmes suivants :

$$Ax = b = \begin{pmatrix} 100 \\ 1 \end{pmatrix}$$

et

$$A(x + \delta x) = b + \delta b = \begin{pmatrix} 100 \\ 0 \end{pmatrix}.$$

Calculer le facteur d'amplification de l'erreur, *i.e.* $\dfrac{\|\delta x\|_1}{\|x\|_1} \dfrac{\|b\|_1}{\|\delta b\|_1}$.

I.4. **Déterminant et conditionnement : deux quantités indépendantes l'une de l'autre !**

Ex 4. *Exemples*

1) On considère la matrice $A = (a_{ij})_{(i,j)\in\{1,\dots,100\}^2}$ diagonale de taille 100×100, définie par $a_{11} = 1$, $a_{ii} = 0.1$ pour $2 \leq i \leq 100$. Déterminer le déterminant et le conditionnement (en norme 2 par exemple) de cette matrice.

2) Soit $n \in \mathbb{N}^*$. On considère la matrice bi-diagonale $B = (b_{ij})_{(i,j)\in\{1,\dots,n\}^2}$ définie par :

$$\begin{cases} b_{ii} = 1, & 1 \leq i \leq n, \\ b_{ii+1} = 2, & 1 \leq i \leq n-1, \\ b_{ij} = 0, & \text{sinon.} \end{cases} \tag{2}$$

On peut montrer que $\text{cond}_1(B) = \text{cond}_\infty(B) = 3(2^n - 1)$ (cf. [LT1] p.144).

Que dire de son déterminant ? Qu'en conclure ?

$\sim\sim\sim$

Soit $n \in \mathbb{N}^*$. Théoriquement, si la matrice $A \in \mathcal{M}_n(\mathbb{R})$ est inversible, on peut résoudre le système en utilisant les formules de Cramer. Par contre le nombre d'opérations est alors vite rédibitoire ($3(n+1)!$ opérations que l'on peut éventuellement réduire à $n^{3.8}$).

On cherche donc des moyens de pouvoir calculer la solution d'un système linéaire de façon plus efficace en essayant de réduire le nombre d'opérations. Les méthodes se répartissent en deux classes : les méthodes directes (qui reposent sur une factorisation de la matrice) et les méthodes itératives (qui définissent une suite de solutions approchées convergeant vers la solution du système linéaire).

II. Méthodes directes pour la résolution de systèmes linéaires [LT1], [QSS], [Fi], [Ci], [S].

II.1. **Méthodes de descente et de remontée.**

Ces méthodes permettent de résoudre un système linéaire triangulaire inférieur ou supérieur.

Ex 5. *Méthodes de descente et remontée : des briques essentielles...*

1) Écrire l'algorithme de remontée qui permet de résoudre un système $Ax = b$ avec A une matrice triangulaire supérieure inversible.

2) De façon analogue, écrire l'algorithme de descente qui permet de résoudre un système $Ax = b$ avec A une matrice triangulaire inférieure.

Ces méthodes nécessitent un équivalent de n^2 opérations

II.2. **Méthode d'élimination de Gauss**

On cherche à transformer le système $Ax = b$ en un système triangulaire supérieur en multipliant à gauche par une matrice M inversible. On cherchera alors à résoudre le système $MAx = Mb$, avec MA triangulaire supérieure, par une méthode de remontée.

Théorème II.1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$. Il existe (au moins) une matrice inversible M telle que la matrice MA soit triangulaire supérieure.*

Démonstration. Voir exercices section II.6 □

En effectuant les $n-1$ étapes nécessaires à l'élimination, on obtient finalement une matrice $A^{(n)} = E^{(n-1)}P^{(n-1)} \dots E^{(1)}P^{(1)}A$ où $P^{(k)}$ est une matrice de permutation échangeant deux lignes. $A^{(n)}$ est alors triangulaire supérieure. On note $M = E^{(n-1)}P^{(n-1)} \dots E^{(1)}P^{(1)}$.

La méthode d'élimination de Gauss nécessite un équivalent de $\dfrac{2n^3}{3}$

II.3. **Factorisation LU**

On dispose d'un théorème général :

Théorème II.2. *Si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice inversible, alors il existe une matrice de permutation P, et deux matrices L et U, respectivement triangulaire inférieure avec des 1 sur la diagonale et triangulaire supérieure, telles que :*

$$PA = LU.$$

Pour ce qui est de la décomposition LU, on a le théorème suivant :

Théorème II.3. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ telle que les n sous-matrices diagonales*

$$\Delta_k = \begin{pmatrix} a_{1,1} & \cdots & a_{1,k} \\ \vdots & & \vdots \\ a_{k,1} & \cdots & a_{k,k} \end{pmatrix}$$

soient inversibles. Alors il existe une unique factorisation $A = LU$ où L est une matrice triangulaire inférieure, telle que $L_{i,i} = 1$, pour tout $0 \leq i \leq n$ et U est une matrice triangulaire supérieure inversible.

Démonstration. Voir exercice II.6. □

Ainsi, pour résoudre le système $Ax = b$, avec la matrice A vérifiant les hypothèses du théorème, on peut utiliser la décomposition LU de la matrice A, si elle existe. On résout alors successivement les deux systèmes $Ly = b$ et $Ux = y$, qui sont des systèmes triangulaire inférieurs ou supérieurs donc plus faciles à résoudre.

Le nombre d'opérations est équivalent à $\frac{2n^3}{3}$.

Il faut y rajouter la résolution du système par remontée en un équivalent de n^2 opérations.

L'intérêt de la décomposition LU réside dans le fait que l'on peut calculer cette décomposition indépendamment du second membre et l'utiliser pour résoudre plusieurs systèmes avec des seconds membres différents. On peut notamment s'en servir pour faire le calcul de l'inverse d'une matrice.

Remarque II.4. *Si la matrice A est creuse (i.e. avec de nombreux coefficients nuls), alors les deux matrices L et U le sont aussi.*

II.4. **Factorisation de Cholesky d'une matrice**

Dans le cas où la matrice A est symétrique définie et positive, on peut améliorer la décomposition LU. En effet, les deux facteurs de la décomposition de Cholesky peuvent être choisis transposés l'une de l'autre.

Théorème II.5. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive. Alors il existe (au moins) une matrice B triangulaire inférieure inversible telle que $A = BB^T$. De plus, si la matrice B est telle que $B_{i,i} > 0$ pour $1 \leq i \leq n$, la factorisation est unique.*

Démonstration. Voir exercices en section II.6. □

La décomposition de Cholesky de la matrice A demande un nombre d'opérations équivalent à $\frac{n^3}{3}$.

II.5. **Factorisation QR d'une matrice**

Pour les matrices non nécessairement définies positives, il existe la décomposition QR en un produit d'une matrice orthogonale Q et une matrice triangulaire supérieure. Cette décomposition peut par exemple être obtenue par la méthode de Householder ou en utilisant Gram-Schmidt (cependant pour obtenir une factorisation QR numériquement, cette dernière méthode est déconseillée).

Théorème II.6. *Soit $A \in \mathcal{M}_n(\mathbb{R})$, une matrice inversible. Alors, il existe une matrice orthogonale Q et une matrice triangulaire supérieure R telle que $A = QR$.*

Démonstration. cf. [LT1], [Fi] □

Il suffit ensuite de résoudre le système $Rx = Q^T b$ par remontée.

$\sim\sim\sim$

II.6. **Exercices et démonstrations des résultats**

Ex 6. *Élimination de Gauss : démonstration*

La démonstration est basée sur l'algorithme du pivot de Gauss.
1) $1^{\text{ière}}$ itération. Supposons que $a_{1,1} \neq 0$. Montrer que l'on peut trouver $E^{(1)}$ telle que

$$E^{(1)}A = \begin{pmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \cdots & \cdots & \cdots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(1)} & & & & \vdots \\ \vdots & \vdots & & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & & \vdots \\ 0 & a_{n,2}^{(1)} & \cdots & \cdots & \cdots & a_{n,n}^{(1)} \end{pmatrix}.$$

Que faire si $a_{1,1} = 0$?

2) On se place à la k-ième étape de l'algorithme d'élimination de Gauss. Soit

$$A^{(k)} = \begin{pmatrix} a_{1,1}^{(k)} & a_{1,2}^{(k)} & \cdots & \cdots & \cdots & a_{1,n}^{(k)} \\ 0 & a_{2,2}^{(k)} & & & * & \vdots \\ \vdots & \ddots & \ddots & & * & \vdots \\ \vdots & (0) & 0 & a_{k,k}^{(k)} & \cdots & a_{k,n}^{(k)} \\ \vdots & & & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{n,k}^{(k)} & \cdots & a_{n,n}^{(k)} \end{pmatrix}$$

telle que $\det(A^{(k)}) = \pm\det(A)$.

Dans le cas où $a_{k,k}^{(k)} \neq 0$, trouver une matrice $E^{(k)}$ telle que $\det(E^{(k)}) = 1$ et telle que la matrice $A^{(k+1)} = E^{(k)}A^{(k)}$ soit de la forme

$$A^{(k+1)} = \begin{pmatrix} a_{1,1}^{(k)} & a_{1,2}^{(k)} & \cdots & \cdots & \cdots & a_{1,n}^{(k)} \\ 0 & a_{2,2}^{(k)} & & & * & \vdots \\ \vdots & \ddots & \ddots & & * & \vdots \\ \vdots & (0) & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} \\ \vdots & & & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{n,n}^{(k+1)} \end{pmatrix}$$

3) Étudier le cas où $a_{k,k}^{(k)} = 0$.

4) Conclure.

Remarque II.7. *Méthode du pivot partiel : Le choix du pivot peut se révéler important. On peut décider de choisir à chaque étape le pivot le plus grand possible, c'est à dire de prendre à la k-ième étape comme pivot l'élément de la k-ième colonne défini ainsi* $|a_{i_0,k}^{(k)}| = \max_{i \geq k} |a_{i,k}^{(k)}|$*, puis d'effectuer une permutation de la k-ième ligne et de la* i_0*-ième ligne.*

Ex 7. *Importance du choix du pivot [LT]*
Supposons que les nombres soient représentés avec 3 chiffres significatifs ($x = \pm 0.\alpha\beta\gamma * 10^e$ avec $(\alpha, \beta, \gamma) \in [0, 9]^3$ et $e \in \mathbb{Z}$). On considère le système $Au = b$ avec :

$$A = \begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix}$$

et

$$b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

1) Résoudre le système par la méthode de Gauss avec comme pivot la composante 10^{-4} de la première ligne, puis avec la composante 1 de la seconde ligne.

2) Calculer les conditionnements cond_1 et cond_∞ de la matrice exacte obtenue à la première étape de la procédure d'élimination de Gauss pour les deux choix de pivots possibles.

Ex 8. *Démonstration du théorème II.3 et réciproque de la factorisation LU [LT]*

1) Montrer que si une telle décomposition existe, alors elle est unique.

2) Utiliser la méthode d'élimination de Gauss sans permutation pour prouver le théorème.

3) Montrer l'assertion réciproque à celle du théorème.

4) Donner la forme explicite de la matrice L.

Ex 9. *Démonstration du théorème de décomposition de Cholesky et de sa réciproque [LT]*

1) Montrer que si la décomposition existe, alors elle est unique.

2) Montrer l'existence de la matrice B en utilisant la décomposition LU de la matrice A.

3) Montrer l'assertion réciproque au théorème.

Ex 10. *Décomposition LU et de Cholesky pour des matrices bandes*

A est une matrice p-bande si $A_{i,j} = 0$ pour $|i - j| \geq p$. Montrer que la décomposition LU préserve la structure des matrices bandes. En déduire qu'il en est de même pour la décomposition de Cholesky.

III. Méthodes itératives pour la résolution de systèmes linéaires [LT1], [QSS], [Fi], [Ci], [S]

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^n$ un vecteur colonne. On cherche à calculer la solution $x \in \mathbb{R}^n$ du système $Ax = b$, comme limite d'une suite de solutions approchées (x^k).

Définition III.1. *On dit qu'une méthode itérative est convergente, si quel que soit le vecteur $x^0 \in \mathbb{R}^n$ et quel que soit le second membre $b \in \mathbb{R}^n$, la suite $(x^k)_{k\geq 0}$ définie par la méthode itérative converge vers la solution x du système linéaire $Ax = b$.*

III.1. Stratégie de quelques méthodes classiques.

Une stratégie consiste à faire apparaître un problème de point fixe. On écrit A sous la forme $A = M - N$ avec M une matrice inversible. On peut alors réécrire le système $Ax = b$ comme $x = M^{-1}(Nx + b)$.

On va donc utiliser (si elle converge) la suite récurrente définie par la *méthode itérative* suivante :
- on se donne un vecteur initial x^0,
- ensuite, pour $k \geq 0$, on calcule $x^{k+1} = M^{-1}(Nx^k + b)$.

Remarque III.2. *En pratique, à l'étape k, on ne calcule pas M^{-1}, mais à l'itération k, on résout le système linéaire : $Mx^{k+1} = (Nx^k + b)$. On a donc tout intérêt à choisir une "bonne" matrice, typiquement diagonale ou triangulaire.*

III.1.1. *Quelques critères de convergence*

Lemme III.3. *[Cia], [S] Soit $A \in \mathcal{M}_n(\mathbb{R})$.*
(i) Pour toute norme matricielle $|||\cdot|||$, on a $\rho(A) \leq |||A|||$.
(ii) Pour tout $\varepsilon > 0$, il existe au moins une norme matricielle subordonnée $|||\cdot|||$ telle que $|||A||| \leq \rho(A) + \varepsilon$.

Lemme III.4. *[Cia,LT1] Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors $\lim_{k\to+\infty} A^k = 0$ si et seulement si $\rho(A) < 1$.*

Théorème III.5. *Une méthode itérative comme définie ci-dessus, converge si et seulement si $\rho(M^{-1}N) < 1$.*

III.2. Les méthodes itératives classiques

On note $A = \begin{pmatrix} \ddots & & -F \\ & D & \\ -E & & \ddots \end{pmatrix} = D - E - F$ avec D, E et F les matrices associées.
On suppose que D est inversible.

• La méthode de Jacobi. On choisit comme matrice M, la matrice diagonale D, *i.e.* $M = D$ et donc la matrice $N = E + F$. La matrice d'itération de la méthode itérative vaut $J = M^{-1}N =$

$D^{-1}(E+F)$. Les composantes du vecteur x^{k+1} à l'itération k, sont calculées grâce à la formule suivante :

$$x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^k \right), \ 1 \leq i \leq n.$$

• La méthode de Gauss Seidel. On choisit comme matrice M, la matrice triangulaire inférieure $M = D - E$ et donc la matrice $N = F$. La matrice d'itération vaut $\mathcal{L}_1 = (D-E)^{-1}F$. Les composantes du vecteur x^{k+1} à la k-ième itération valent donc :

$$x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{k+1} - \sum_{j>i} a_{ij} x_j^k \right), \ 1 \leq i \leq n.$$

• La méthode de relaxation. Soit $\omega \in \mathbb{C}^*$. On choisit comme matrice M, la matrice triangulaire inférieure $M = \dfrac{1}{\omega} D - E$ et donc la matrice $N = \left(\dfrac{1}{\omega} - 1 \right) D + F$. La matrice d'itération vaut : $\mathcal{L}_\omega = (D - \omega E)^{-1}((1-\omega)D + \omega F)$. Les composantes du vecteur x^{k+1} à la k-ième itération valent donc :

$$x_i^{k+1} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{k+1} - \sum_{j>i} a_{ij} x_j^k + \left(\frac{1}{\omega} - 1 \right) a_{ii} x_i^k \right), \ 1 \leq i \leq n.$$

ce qui donne :

$$x_i^{k+1} = x_i^k + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{k+1} - \sum_{j \geq i} a_{ij} x_j^k \right), \ 1 \leq i \leq n.$$

III.3. **Résultats de convergence**

On a un premier résultat sur l'ensemble des paramètres ω possibles s'il y a convergence.

Théorème III.6. *Si la méthode de relaxation converge avec une matrice $A \in \mathcal{M}_n(\mathbb{C})$ et un paramètre $\omega \in \mathbb{C}^*$, alors $|\omega - 1| < 1$.*

III.3.1. *Cas des matrices à diagonale strictement dominante.*

Définition III.7. *Une matrice est à diagonale strictement dominante si*

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

Théorème III.8. *Si $A \in \mathcal{M}_n(\mathbb{R})$ est à diagonale strictement dominante, alors*
(i) A est une matrice inversible,
(ii) la méthode de Jacobi converge,
(iii) la méthode de relaxation avec $0 < \omega \leq 1$ converge.

III.3.2. *Cas des matrices définies positives.*

Théorème III.9. *Supposons que $A = M - N$ et $M^* + N$ soient symétriques définies positives. Alors $\rho(M^{-1}N) < 1$.*

Théorème III.10. *Si $A \in \mathcal{M}_n(\mathbb{R})$ est symétrique définie positive, alors la méthode de relaxation avec $|\omega - 1| < 1$ converge.*

III.4. tests d'arrêts, vitesse de convergence et taux de convergence

On se donne $\|\cdot\|$ et $|||\cdot|||$ la norme subordonnée associée. On note B la matrice d'itération associée à une méthode donnée. On suppose que la méthode itérative qui y est associée converge.

III.4.1. *Test d'arrêt*

Le test d'arrêt usuel consiste à se donner $\varepsilon > 0$ petit et à arrêter les itérations lorsque

$$\frac{\|r^k\|}{\|b\|} \leq \varepsilon, \tag{3}$$

avec $r^k = b - Ax^k$.

III.4.2. *Vitesse et taux de convergence*

S'il y a convergence, on peut chercher à étudier la vitesse avec laquelle l'erreur $e^k := x - x^k$ tend vers 0 lorsque k tend vers $+\infty$.

Définition III.11. *On définit le facteur moyen de réduction de l'erreur par itération par la formule*

$$\sigma[k] := \left(\frac{\|e^k\|}{\|e^0\|}\right)^{\frac{1}{k}}, \ \textit{si } x^0 \neq x.$$

Plus $|||B|||$ est petite, plus la convergence de $(x^k)_{k\in\mathbb{N}}$ vers x est rapide.

Définition III.12. *On définit le taux moyen de convergence pour k itérations ($k \in \mathbb{N}^*$) comme le nombre :*

$$R_k(B) := -\ln(|||B^k|||^{\frac{1}{k}}).$$

Définition III.13. *On définit le taux de convergence asymptotique de la matrice d'itération B comme le nombre :*

$$R(B) := -\ln(\rho(B)).$$

Proposition III.14. *Le taux moyen de convergence (définition III.12) est inversement proportionnel au nombre d'itérations nécessaires pour atteindre la précision ε.*

Démonstration. cf. exercices. □

III.5. **Méthode de gradients**

On se place dans le cas où la matrice A est symétrique définie positive. On calcule alors la solution du système $Ax = b$ en minimisant la fonction $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}$ définie par $\mathcal{J}(x) = \frac{1}{2}(Ax, x) - (b, x)$. Pour cela, on utilise des méthodes de descente. Le principe est de fixer de façon itérative des directions de descente $d_k \in \mathbb{R}^n$ et des pas de descente $\rho_k \in \mathbb{R}$. On écrit les itérés $x^{k+1} = x^k - \rho_k d_k$. Dans le cas des méthodes de gradient, la direction de descente est le gradient $d_k = \nabla\mathcal{J}(x^k)$ ou une combinaison de certains gradients.

$\rightsquigarrow$ *Voir aussi les TD/TP sur l'optimisation.*

• Méthode de gradient à pas constant ou à pas optimal.

Dans ce cas, la direction de descente est exactement le gradient $d_k = \nabla\mathcal{J}(x^k)$. Si on choisit le pas $\rho_k = \rho^*$ constant, il s'agit de la méthode de gradient à pas constant. Le principe de la méthode de gradient à pas optimal est d'optimiser à chaque étape le pas de descente en minimisant la fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(\rho) = \mathcal{J}(x^k - \rho d_k)$. On trouve que $\rho_k = \dfrac{\|r_k\|_2^2}{(Ar_k, r_k)}$ avec $r_k = b - Ax^k$.

Théorème III.15. *Si la matrice est symétrique définie positive, alors la méthode de gradient à pas optimal est convergente avec un facteur de convergence égal à* $\dfrac{\frac{\lambda_1}{\lambda_n} - 1}{\frac{\lambda_1}{\lambda_n} + 1}$, *si* λ_1 *est la plus grande valeur propre de* A *et* λ_n *la plus petite valeur propre de* A.

Remarque III.16. *Dans le cas de la méthode du gradient conjugué, on définit ainsi les itérations successives : on suppose que les* x^l, $0 \leq l \leq k$ *sont déjà calculés et on cherche* $\mathcal{J}(x^{k+1}) = \min\limits_{x \in x^k + G_k} \mathcal{J}(x)$, *où* G_k *est le sous-espace engendré par les* $\nabla\mathcal{J}(x^l)$, $0 \leq l \leq k$. *Cet algorithme se termine en au plus* n *itérations.*

III.6. **Exercices**

Ex 11. *Condition de convergence d'une méthode itérative.* Démontrer le théorème III.5.

Ex 12. *Démonstration du théorème III.8[LT2]*

1) Montrer (i) par l'absurde ou en utilisant les disques de Gerschgörin.

2) Calculer les coefficients de la matrice J de la méthode de Jacobi et montrer que $\||J|\|_\infty < 1$. Montrer alors (ii).

3) Montrer que si λ est une valeur propre de $\mathcal{L}_\omega$, alors $\lambda + \omega - 1$ est une valeur propre de $\omega D^{-1}(F + \lambda E)$.

4) On suppose que $|\lambda| \geq 1$. Majorer $|\lambda + \omega - 1|$, grâce au théorème de Gerschgörin. Majorer $|\lambda|$. En déduire (iii).

Ex 13. *Démonstration du théorème III.9*
On définit la norme vectorielle $\|x\|_A^2 = x^* Ax$, pour $x \in \mathbb{R}^n$.

1) Montrer que $M^{-1}Nx = x - y$, avec $y = M^{-1}Ax$.
2) En déduire que $\|M^{-1}Nx\|_A < \|x\|_A$ et conclure.

Ex 14. *Test d'arrêt et conditionnement [LT2]*

On se place dans le cadre de la section III.4. On suppose qu'on a (3).

1) Montrer que $\dfrac{\|e^k\|}{\|x\|} \leq \varepsilon \text{cond}(A)$, où cond est le conditionnement associé à $\|\cdot\|$.

2) Montrer que $\sigma[k]$ définit à la définition III.11 vérifie

$$\sigma[k] \leq \|B^k\|^{\frac{1}{k}},$$

pour tout $k \geq 0$.

3) En déduire que pour avoir $\dfrac{\|e^k\|}{\|e^0\|} \leq \varepsilon$, il suffit d'imposer $\|B^k\| \leq \varepsilon$.

Ex 15. *Taux moyen de convergence [LT2]*

On se place dans le cadre des notations de la section III.4.
1) Montrer que $|||B^k|||^{\frac{1}{k}} < 1$ pour k suffisamment grand.

2) Montrer que le taux moyen de convergence (définition III.12) est inversement proportionnel au nombre d'itérations nécessaires pour atteindre la précision ε (au sens de la question 3 de l'exercice 14).

Ex 16. On se place dans le cadre des notations de la section III.4 avec B une matrice hermitienne. On choisit $\|\cdot\|_2$ comme norme vectorielle. Montrer que dans ce cas $R_k(B) = -\ln(\rho(B)) = R(B)$.

Ex 17. Montrer que le nombre d'itérations pour réduire l'erreur d'un facteur ε vérifie

$$k \geq -\frac{\ln(\varepsilon)}{R(B)}.$$

Ex 18. *Sur les méthodes de gradient*

1) Pourquoi a-t-on besoin de l'hypothèse A symétrique définie positive ?
2) Justifier le choix du gradient comme direction de descente.
3) Écrire l'algorithme obtenu avec la méthode du gradient à pas constant pour la résolution du système $Ax = b$, dans le cas de la fonction $\mathcal{J}$ de la section III.5.

4) Même question avec la méthode du gradient à pas optimal. On montrera que dans ce cas, la solution de la fonction à minimiser est donnée de façon simple et explicite.

Ex 19. *Une matrice issue d'une discrétisation par différences finies [S, Fi]*

Soit A_h la matrice $\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & (0) & \\ & \ddots & \ddots & \ddots & \\ (0) & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}$ de taille $n \in \mathbb{N}^*$ qui discrétise par différences finies (de pas $h > 0$) l'opérateur $-\frac{\partial^2}{\partial x^2}$ sur $[-1, 1]$ avec les conditions au bord $u(-1) = u(1) = 0$. On note $A := h^2 A_h$

1) Quelles sont les valeurs propres (et vecteurs propres associés) de A. En déduire $\text{cond}_2(A)$.

2) En déduire les rayons spectraux des matrices des méthodes de Jacobi et de Gauss-Seidel, en utilisant la relation $\rho(\mathcal{L}_1) = \rho(J)^2$ valable pour les matrices tridiagonales [LT2].

3) [S] Donner un développement asymptotique en fonction de n :
- du rayon spectral de la matrice de la méthode de Jacobi J ;
- du rayon spectral de la matrice de la méthode de Gauss-Seidel $\mathcal{L}_1$;
- du rayon spectral de la matrice de la méthode de relaxation $\mathcal{L}_{\omega^*}$, avec $\omega^* = \frac{2}{1+\sqrt{1-\rho(J)^2}}$, en utilisant que pour ce paramètre $\rho(\mathcal{L}_{\omega^*}) = \omega^* - 1$.

4) Déterminer alors un équivalent du nombre d'itérations nécessaires pour chacune des trois méthodes de la question 3).

5) Calculer, si elle existe, la décomposition LU de cette matrice et en donner le dombre d'opérations.

Ex 20. *Quelques contre-exemples*

1) Montrer que la matrice $\begin{pmatrix} 1 & 3/4 & 3/4 \\ 3/4 & 1 & 3/4 \\ 3/4 & 3/4 & 1 \end{pmatrix}$ est définie positive. Montrer que la méthode de Jacobi appliquée à cette matrice ne converge pas.

2) Montrer que la méthode de Jacobi appliquée à la matrice $\begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$ converge, mais que la méthode de Gauss-Seidel ne converge pas.

3) Montrer que la méthode de Gauss-Seidel appliquée à la matrice $\begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$ converge, mais que la méthode de Jacobi ne converge pas.

Références :

[Ci] P. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation,* Dunod.

[Fi] F. Filbet, *Analyse numérique,* Dunod.

[LT1] P. Lascaux, R. Théodor, *Analyse numérique matricielle appliquée à l'art de l'ingénieur, tome 1,* Dunod.

[LT2] P. Lascaux, R. Théodor, *Analyse numérique matricielle appliquée à l'art de l'ingénieur, tome 2,* Dunod.

[QS] A. Quarteroni, F. Saleri *Calcul scientifique* Springer.

[QSS] A. Quarteroni, R. Sacco, F. Saleri *Méthodes numériques pour le calcul scientifique* Springer.

[S] M. Schatzman, *Analyse Numérique,* Dunod.