CHAPITRE 1

COURS UM6P, ALGÈBRE LINÉAIRE NUMERIQUE

L'objectif de cette série de cours réalisés en novembre 2019 à l'UM6P est de présenter différentes méthodes de résolution d'un système linéaire ou de recherche de valeurs propres et de vecteurs propres d'une matrice. L'impérative nécessité d'une telle construction s'impose à la lecture du nombre d'années (en l'occurence 3.10^{146}) demandées par la résolution d'un système 100×100 par la méthode d'inversion de Cramer avec un ordinateur de 100 Mflops (c'est à dire effectuant 10^8 opérations en une seconde).

Après avoir mis l'accent au paragraphe 1 sur l'importance du bon conditionnement des deux types de problèmes, plusieurs méthodes directes (ou exactes) de résolution d'un système linéaire (noté de manière générale $Au = b$) sont présentées au paragraphe 2, toutes basées sur la triangularisation de la matrice A. Le paragraphe 3 s'intéresse ensuite aux méthodes itératives, donnant une solution approchée du même système. Différentes méthodes de recherche de valeurs propres et de vecteurs propres d'une matrice sont construites ou simplement citées au paragraphe 4. Enfin, le paragraphe 5 fait un lien avec l'optimisation à travers l'étude des problèmes aux moindres carrés.

1.1 Généralités

1.1.1 Difficulté du problème à résoudre

1.1.1.1 Résolution d'un système linéaire

L'exemple suivant (puisé dans [Ci]) sert d'illustration à la difficulté de la résolution numérique d'un système linéaire : soit à résoudre le système 4×4

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

La solution exacte est

$$(u_1, u_2, u_3, u_4) = (1, 1, 1, 1)$$

alors que le système où seul le second membre a été légèrement modifié

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} u_1' \\ u_2' \\ u_3' \\ u_4' \end{pmatrix} = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix}$$

a pour solution

$$(u_1', u_2', u_3', u_4') = (9.2, -12.6, 4.5, -1.1).$$

Ainsi, une simple erreur d'arrondi sur la donnée du second membre peut provoquer une erreur totale sur la solution du système.

1.1.1.2 Recherche de valeurs propres d'une matrice

La recherche de valeurs propres d'une matrice est également un problème numériquement difficile comme le montre l'exemple suivant : la matrice

$$A = \begin{pmatrix} 0 & 0 & 0 & 10^{-4} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

a pour valeurs propres les racines quatrièmes de 10^{-1}, toutes de module 10^{-1} alors que la même matrice où 10^{-4} est remplacé par zéro a pour unique valeur propre zéro. Au vu de ces deux exemples, il s'avère donc indispensable de définir préalablement une grandeur attachée à la matrice A (différente pour les deux problèmes) permettant d'évaluer l'amplification possible sur la solution d'une erreur sur les données.

1.1.2 Conditionnement

1.1.2.1 Définition

Définition 1. *Soit $|||.|||$ une norme subordonnée sur $\mathscr{M}_n(\mathbf{R})$ (c'est à dire une norme d'opérateurs associée à une norme $||.||$ de $\mathbf{R}^n$) On appelle conditionnement de $A \in GL_n(\mathbf{R})$ par rapport à cette norme, le réel positif*

$$\operatorname{cond} A = |||A||| \cdot |||A^{-1}|||.$$

Si de plus A est diagonalisable, on appelle conditionnement spectral de A la valeur

$$\Gamma(A) = \inf_{P^{-1}AP=\operatorname{diag}(\lambda_1,\ldots,\lambda_n)} \operatorname{cond} P.$$

On peut déjà observer quelques propriétés immédiates du conditionnement d'une matrice A : (i) $\operatorname{cond} A \geq 1$, $\operatorname{cond} A^{-1} = \operatorname{cond} A$, $\operatorname{cond} \lambda A = \operatorname{cond} A$. (ii) $\operatorname{cond}_2 Q = 1$ si $Q \in O_n(\mathbf{R})$ où cond_2 désigne le conditionnement associé à la norme euclidienne sur $\mathbf{R}^n$. (iii) $\operatorname{cond}_2 QA = \operatorname{cond}_2 AQ = \operatorname{cond}_2 A$ si $(A, Q) \in GL_n(\mathbf{R}) \times O_n(\mathbf{R})$.

1.1.2.2 Estimations d'erreurs

Théorème 1. *Soient A et A' deux matrices inversibles et b et b' deux vecteurs de $\mathbf{R}^n$ (avec $b \neq 0$).*
(i) Soient u et u' les solutions respectives des systèmes $Au = b$ et $Au' = b'$. On a

$$\frac{||u-u'||}{||u||} \leq \operatorname{cond} A \frac{||b-b'||}{||b||}$$

et cette inégalité est optimale (il existe des cas d'égalité). (ii) Soient u et u' les solutions respectives des systèmes $Au = b$ et $A'u' = b$. On a

$$\frac{||u-u'||}{||u'||} \leq \operatorname{cond} A \frac{|||A-A'|||}{|||A|||}$$

et

$$\frac{||u-u'||}{||u||} \leq \operatorname{cond} A \frac{|||A-A'|||}{|||A|||}\big(1 + O(|||A-A'|||)\big).$$

Démonstration. : Pour démontrer la première inégalité, il suffit d'écrire les deux égalités suivantes dans $\mathbf{R}^n$:

$$\begin{cases} u-u' = A^{-1}(b-b'), \\ Au = b, \end{cases}$$

et d'utiliser la propriété de la norme d'opérateurs

$$\begin{cases} ||u-u'|| \leq |||A^{-1}||| \cdot ||b-b'||, \\ ||b|| \leq |||A||| \cdot ||u||. \end{cases}$$

Les inégalités dans (ii) s'obtiennent en raisonnant de manière analogue □

Théorème 2. *Soit A une matrice diagonalisable de valeurs propres* $(\lambda_1, ..., \lambda_n)$ *et soit* $|||.|||$ *une norme subordonnée telle que pour toute matrice diagonale*

$$|||\text{diag}(d_1, ..., d_n)||| = \max_{1 \leq i \leq n} |d_i|.$$

Soit $A' \in \mathscr{M}_n(\mathbf{R})$ *: le spectre de* A' *est contenu dans l'union des n disques du plan complexe*

$$D_i = \left\{ z \in \mathbf{C} \ / \ |z - \lambda_i| \leq \Gamma(A)|||A' - A||| \right\}, \quad 1 \leq i \leq n$$

Démonstration. : Soit P une matrice inversible telle que $P^{-1}AP = \text{diag}(\lambda_1, ..., \lambda_n)$ et soit λ une valeur propre complexe de A'. On peut toujours supposer que la matrice

$$D = \text{diag}(\lambda_1 - \lambda, ..., \lambda_n - \lambda)$$

est inversible (sinon le résultat est trivialement vérifié).

On a alors

$$\begin{aligned} P^{-1}(A' - \lambda I)P &= D + P^{-1}(A' - A)P \\ &= D \underbrace{(I + D^{-1}P^{-1}(A' - A)P)}_{B}. \end{aligned}$$

En particulier, la matrice B est singulière et par conséquent

$$|||I - B||| \geq 1$$

(sinon la série $\displaystyle\sum_{n=0}^{+\infty} |||I - B|||^n$ serait convergente et $B = I - (I - B)$ inversible).

De plus, par les propriétés de la norme subordonnée

$$|||I - B||| \leq |||D^{-1}||| \cdot |||P^{-1}||| \cdot |||A' - A||| \cdot |||P|||$$

ce qui implique bien

$$\max_{1 \leq i \leq n} |\lambda_i - \lambda| \leq \text{cond}\, P |||A' - A||| \quad \square$$

Remarque : Il existe (fort heureusement) des algorithmes de calcul approché du conditionnement d'une matrice permettant de s'abstenir du calcul de A^{-1} : voir à ce sujet [LaTh1].

1.1.2.3 Préconditionnement

Le Théorème précédent montre que le bon conditionnement d'un système linéaire $Au = b$ est assuré par une valeur faible de cond A. Il peut donc être intéressant avant de résoudre un système $Au = b$ de multiplier A et b par l'inverse d'une matrice C proche de A (et facile à inverser) en espérant réduire ainsi le conditionnement du nouveau système. Par exemple, le choix de la matrice $C = \text{diag}(a_{1,1}, ..., a_{n,n})$ (sous réserve qu'elle soit inversible) s'avère donner de bons résultats (voir [LuPi] pour d'autres exemples).

1.2 Résolution d'un système linéaire : méthodes directes

1.2.1 Méthode du pivot de Gauss

Cette méthode est basée sur l'observation qu'un système linéaire associé à une matrice inversible et triangulaire (inférieure ou supérieure, ensembles respectivement notés $TI_n(\mathbf{R})$ et $TS_n(\mathbf{R})$) se résout en un nombre réduit d'opérations par un principe de remontée (en l'occurence n^2 opérations élémentaires).

La méthode de Gauss consiste à se ramener à un tel système ou plus précisément, à déterminer une matrice inversible $M \in GL_n(\mathbf{R})$ telle que MA soit une matrice triangulaire supérieure. Il suffit alors de calculer MA et Mb et de résoudre le système triangulaire $MAu = Mb$ pour avoir la solution du système initial $Au = b$. Dans l'algorithme, il ne sera même pas nécessaire de calculer explicitement M. On définit

tout d'abord un principe de transformation élémentaire noté G d'une matrice inversible $A = [a_{i,j}]_{1\leq i,j\leq n} \in GL_n(\mathbf{R})$ et d'un vecteur colonne $b = (b_i)_{1\leq i\leq n} \in \mathbf{R}^n$:

1.2.1.1 Principe de transformation élémentaire

`Etape 1` *: soit $a_{i,1}$ un coefficient non nul et de valeur absolue maximale de la première colonne de A (il en existe un forcément). On échange alors les lignes i et 1 de la matrice A et du vecteur colonne b. On obtient une nouvelle matrice $\tilde{A} = [\tilde{a}_{i,j}]$ et un nouveau vecteur colonne $\tilde{b} = (\tilde{b}_i)_{1\leq i\leq n}$.*
`Etape 2` *: pour tout $j \geq 2$, on remplace dans $\tilde{A}$ et $\tilde{b}$, la ligne j ($\tilde{L}_j$) par la ligne $\tilde{L}_j - \frac{\tilde{a}_{j,1}}{\tilde{a}_{1,1}}\tilde{L}_1$ On note alors $G(A)$ et $G(b)$ la nouvelle matrice et le nouveau vecteur colonne obtenus.* On peut alors proposer l'algorithme suivant de résolution du système

$Au = b$: Algorithme du pivot de Gauss :

`Etape 1` *: on part de la matrice A et du vecteur colonne b auxquels on applique l'algorithme de transformation élémentaire : on obtient alors une matrice $A_1 = G(A)$ et un vecteur colonne $b_1 = G(b)$. On remarque que la première colonne de A_1 est constituée de zéros en dehors de son premier terme. On note alors A'_1 la matrice carrée mineure de taille $(n-1)$ de A_1 obtenue en supprimant la première ligne et la premiere colonne et b'_1 le vecteur colonne de taille $(n-1)$ égal à b_1 privé de sa première valeur (à noter que A'_1 est inversible grâce aux propriétés du déterminant).*
`Etape 2` *: on applique l'algorithme de transformation élémentaire à A'_1 et à b'_1. On obtient ainsi une nouvelle matrice carrée de taille $(n-1)$, $A_2 = G(A'_1)$, et un nouveau vecteur colonne $b_2 = G(b'_1)$ puis en procédant comme dans l'étape 1, une matrice mineure de taille $(n-2)$, A'_2 et un vecteur colonne mineur de taille $(n-2)$, b'_2. ...*
`Etape n` *: on a construit une suite de matrices carrées $A_k = [a^k_{i,j}]$ $(1 \leq k \leq n)$ et de vecteurs colonne $b_k = [b^k_i]$ tous deux de taille $(n-k+1)$. On reconstruit alors une matrice triangulaire supérieure $U = [u_{i,j}]$ et un vecteur colonne β constitués des premières lignes de A_k et de b_k :*

$$U = \begin{pmatrix} a^1_{1,1} & a^1_{1,2} & \dots & a^1_{1,n} \\ 0 & a^2_{1,1} & \dots & a^2_{1,n-1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a^n_{1,1} \end{pmatrix}$$

et

$$\beta = {}^t(b^1_1, ..., b^n_1).$$

La solution u du système $Au = b$ est alors obtenue par résolution du système triangulaire $Uu = \beta$.

Le théorème suivant permet de justifier cet algorithme :

Théorème 3. *En reprenant les notations de l'algorithme du pivot de Gauss, il existe $M \in GL_n(\mathbf{R})$ tel que $U = MA$ et $\beta = Mb$.*

Démonstration. : On considère la famille des matrices de transposition $T^{i,j}$ $(1 \leq i, j \leq n)$ de terme général $t^{i,j}_{k,l}$ valant 1 si $l = \tau(k)$ (où τ est la permutation (i, j)) et 0 sinon, ainsi que la famille des matrices de transvection $E^i(\alpha_{i+1}, ..., \alpha_n)$ $(1 \leq i \leq n$, $(\alpha_j)_{i+1 \leq j \leq n} \in \mathbf{R}^{n-i})$ telle que

$$E^i(\alpha_{i+1}, ..., \alpha_n) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \alpha_{i+1} & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & \alpha_n & \dots & 0 & 1 \end{pmatrix}$$

On remarque aussitôt que ces deux familles sont incluses dans $GL_n(\mathbf{R})$.

En reprenant le principe de transformation élémentaire de A, on montre par un calcul matriciel simple, que $T^{i,j}A$ (respectivement $T^{i,j}b$) correspond à la matrice $\tilde{A}$ (respectivement $\tilde{b}$) de l'étape 1, puis que

$$\begin{cases} G(A) = E^1(-\frac{\tilde{a}_{2,1}}{\tilde{a}_{1,1}}, ..., -\frac{\tilde{a}_{n,1}}{\tilde{a}_{1,1}})\tilde{A} = E^1(-\frac{\tilde{a}_{2,1}}{\tilde{a}_{1,1}}, ..., -\frac{\tilde{a}_{n,1}}{\tilde{a}_{1,1}})T^{i,j}A, \\ G(b) = E^1(-\frac{\tilde{a}_{2,1}}{\tilde{a}_{1,1}}, ..., -\frac{\tilde{a}_{n,1}}{\tilde{a}_{1,1}})\tilde{b} = E^1(-\frac{\tilde{a}_{2,1}}{\tilde{a}_{1,1}}, ..., -\frac{\tilde{a}_{n,1}}{\tilde{a}_{1,1}})T^{i,j}b. \end{cases}$$

On déduit alors aisément de cette observation qu'au terme de l'algorithme du pivot de Gauss, on a $U = MA$ et $\beta = Mb$, où M est un produit de matrices de transpositions et de matrices de transvection, ce qui achève la démonstration du théorème □

Remarques :

1. L'algorithme de Gauss nécessite environ $\frac{n^3}{3}$ additions, $\frac{n^3}{3}$ multiplications et $\frac{n^2}{2}$ divisions avant la résolution du système triangulaire, soit au total un coût algorithmique de l'ordre de $\frac{2n^3}{3}$ opérations. Par exemple, un système 100×100 sera résolu en un dixième de seconde avec un ordinateur fonctionnant à 100Mflops (valeur à comparer avec celle donnée en introduction). Concernant la place mémoire nécessaire, on s'aperçoit aisément que la méthode de Gauss peut être implémentée sans affectation supplémentaire de variables (voir [LaTh1] et la séance B.I.).

2. On parle de stratégie de pivot partiel comme lorsque dans l'étape 1 de l'algorithme de transformation élémentaire, on choisit $a_{i,1} \neq 0$ tel que

$$|a_{i,1}| = \max_{1 \leq j \leq n} |a_{j,1}|.$$

La stratégie de pivot total consisterait à prendre le maximum sur toutes les valeurs de A et donc à autoriser des permutations de colonnes sur A. Ces deux méthodes permettent de minimiser les erreurs d'arrondis dans les divisions intervenant dans l'algorithme (voir [Sch] ou [LaTh1] pour d'autres choix judicieux de pivot). À noter cependant que toutes ces stratégies engendrent un surcoût d'opérations informatiques (de l'ordre de $\frac{n^2 \ln_2 n}{4}$ comparaisons et échanges pour le pivot partiel).

3. Il est également possible de calculer par cette méthode le déterminant de la matrice A (suivant la parité du nombre de changement de lignes, il est égal à $\det(U)$ ou à $-\det(U)$) et l'inverse de A (en transformant A en la matrice identité par opérations élémentaires sur les lignes).

1.2.2 Méthode de factorisation LU

Cette méthode consiste à déterminer, en reprenant l'algorithme du pivot de Gauss, une factorisation de la matrice A (possible seulement dans certains cas) :

$$A = LU$$

avec $(L,U) \in TI_n(\mathbf{R}) \times TS_n(\mathbf{R})$ (L pour lower et U pour upper). L'intérêt d'une telle factorisation est manifeste lorsqu'il est nécessaire de résoudre l'équation $Au = b$ avec plusieurs seconds membres : on est alors ramemé à chaque fois à résoudre successivement deux systèmes triangulaires $Lw = b$ et $Uu = w$.

On montre en conservant les notations de l'algorithme du pivot de Gauss qu'une condition suffisante de l'existence de cette factorisation est :

$$a_{1,1} \neq 0 \quad \text{et} \quad \forall k \in \{1,...,n-2\}, \quad a_{2,2}^k \neq 0, \tag{1}$$

autrement dit, qu'il n'est jamais nécessaire au cours de l'algorithme d'effectuer une permutation de lignes dans le principe de transformation élémentaire (si on décide de renoncer à toute stratégie de pivot). Dans ce cas, la matrice M du Théorème précédent est en effet un produit de matrices de transvection :

$$M = E^{n-1}(\alpha_n^{n-1})E^{n-2}(\alpha_{n-1}^{n-2}, \alpha_n^{n-2})...E^2(\alpha_3^2,...,\alpha_n^2)E^1(\alpha_2^1,...,\alpha_n^1)$$

avec

$$\begin{cases} \alpha_j^1 = -\dfrac{a_{j,1}}{a_{1,1}} \quad (2 \leq j \leq n), \\ \alpha_j^i = -\dfrac{a_{j-i+2,2}^{i-1}}{a_{2,2}^{i-1}} \quad (2 \leq i \leq n-1, i+1 \leq j \leq n) \end{cases}$$

et est donc en particulier une matrice triangulaire inférieure. La relation $U = MA$ implique alors

$$A = LU$$

avec $L = M^{-1} \in TI_n(\mathbf{R})$ (l'inverse d'une matrice triangulaire inférieure est encore une matrice triangulaire inférieure grâce au principe de remontée) et $U \in TS_n(\mathbf{R})$.

Remarque : D'un point de vue pratique, il est essentiel de remarquer que L est aisément calculable (au contraire de M). En raisonnant par exemple sur les endomorphismes associés, on vérifie que

$$L = M^{-1} = E^1(-\alpha_2^1,...,-\alpha_n^1)...E^{n-2}(-\alpha_{n-1}^{n-2},-\alpha_n^{n-2})E^{n-1}(-\alpha_n^{n-1})$$

est égal à la matrice

$$L = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ -\alpha_2^1 & 1 & \ddots & & \vdots \\ -\alpha_3^1 & -\alpha_3^2 & 1 & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ -\alpha_n^1 & -\alpha_n^2 & \dots & -\alpha_n^{n-1} & 1 \end{pmatrix}.$$

Le théorème suivant donne une condition suffisante sur A, plus opérationnelle que la relation (1), d'existence et d'unicité d'une factorisation LU :

Théorème 4. *: soit $A = [a_{i,j}] \in GL_n(\mathbf{R})$ tel que pour tout $k \in \{1, ..., n\}$, la matrice*

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

soit inversible. Il existe un unique couple $(L, U) \in TI_n(\mathbf{R}) \times TS_n(\mathbf{R})$ vérifiant

$$\forall i \in \{1, ..., n\},\ l_{i,i} = 1$$

et tel que

$$A = LU.$$

Démonstration. : Existence : on montre que la matrice A vérifie la relation (1). Pour cela, on démontre par récurrence qu'au terme de l'étape k ($1 \leq k \leq n-1$) de l'algorithme du pivot de Gauss, aucun échange de lignes n'a été nécessaire et qu'on a donc la situation suivante :

$$\begin{pmatrix} a_{1,1} & * & \dots & \dots & \dots & * \\ 0 & a_{2,2}^1 & * & \dots & \dots & * \\ \vdots & \ddots & \ddots & \ddots & \dots & \vdots \\ 0 & \dots & 0 & a_{2,2}^k & \dots & * \\ 0 & \dots & 0 & a_{3,2}^k & \dots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & * & \dots & * \end{pmatrix} = E^k(\alpha_{k+1}^k, ..., \alpha_n^k)...E^1(\alpha_2^1, ..., \alpha_n^1)A, \qquad (2)$$

avec

$$a_{1,1} \neq 0,\ a_{2,2}^1 \neq 0, ..., a_{2,2}^{k-1} \neq 0.$$

Cette relation est vraie lorsque $k = 1$ (car le pivot peut être pris égal à $a_{1,1}$ supposé non nul). Supposons la relation (2) vérifiée jusqu'à un certain $k \in \{1, ..., n-2\}$. On observe qu'en tronquant chacune des matrices de l'égalité (2) aux $k+1$ premières lignes et colonnes, on a :

$$\begin{pmatrix} a_{1,1} & * & \dots & * \\ 0 & a_{2,2}^1 & \dots & * \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_{2,2}^k \end{pmatrix} = E^k(\alpha_{k+1}^k)...E^1(\alpha_2^1, ..., \alpha_{k+1}^1)\Delta_{k+1}$$

En prenant le déterminant de cette nouvelle expression, il vient :

$$a_{1,1}a_{2,2}^1...a_{2,2}^{k-1}a_{2,2}^k = \det\Delta_{k+1} \neq 0,$$

ce qui implique avec l'hypothèse de récurrence que $a_{2,2}^k \neq 0$. Ce dernier élément peut donc être pris comme nouveau pivot et l'expression (2) demeure encore valable au rang $k+1$. Unicité : on suppose qu'on peut écrire $A = L_1U_1 = L_2U_2$. En particulier, l'égalité $\underbrace{L_2^{-1}L_1}_{\in TI_n(\mathbf{R})} = \underbrace{U_2U_1^{-1}}_{\in TS_n(\mathbf{R})}$ implique que $L = L_2^{-1}L_1$ est une matrice diagonale (car triangulaire inférieure et supérieure) égale à la matrice identité I_n (car $l_{i,i} = 1$). Ainsi, $L_2 = L_1$ puis $U_2 = U_1$.

Dans le cas particulier important des matrices tridiagonales, les matrices L et U peuvent être explicitées (lorsqu'elles existent). La démonstration du théorème suivant est laissée au lecteur en exercice :

Théorème 5. *Soit A une matrice tridiagonale :*

$$A = \begin{pmatrix} b_1 & c_1 & 0 & \dots & 0 \\ a_2 & b_2 & c_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & \dots & 0 & a_n & b_n \end{pmatrix}.$$

On définit la suite $(\delta_k)_{0\leq k\leq n}$ par les relations

$$\begin{cases} \delta_0 = 1 \\ \delta_1 = b_1, \\ \delta_k = b_k\delta_{k-1} - a_kc_{k-1}\delta_{k-2} \quad (2 \leq k \leq n). \end{cases}$$

Alors $\delta_k = \det\Delta_k$ *(avec les notations du Théorème 4) et si tous les* δ_k *sont non nuls,* $A = LU$ *avec*

$$L = \begin{pmatrix} 1 & 0 & & \dots & 0 \\ a_2\frac{\delta_0}{\delta_1} & 1 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & a_n\frac{\delta_{n-2}}{\delta_{n-1}} & 1 \end{pmatrix}$$

et

$$U = \begin{pmatrix} \frac{\delta_1}{\delta_0} & c_1 & 0 & \dots & 0 \\ 0 & \frac{\delta_2}{\delta_1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & c_{n-1} \\ 0 & \dots & \dots & 0 & \frac{\delta_n}{\delta_{n-1}} \end{pmatrix}.$$

Remarque : 1. Le fait de factoriser une matrice A mal conditionnée n'améliore pas la situation : soit par exemple à résoudre le système $Au = b$ avec $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ et

$$A = \begin{pmatrix} 10^{-20} & 1 \\ 1 & 0 \end{pmatrix}$$

et dont la solution exacte est $u = \begin{pmatrix} 1 \\ 1-10^{-20} \end{pmatrix}$.

On peut décomposer A en

$$A = \begin{pmatrix} 10^{-20} & 0 \\ 1 & -10^{20} \end{pmatrix}\begin{pmatrix} 1 & 10^{20} \\ 0 & 1 \end{pmatrix}.$$

La solution exacte de $Lw = b$ est $w = \begin{pmatrix} 10^{20} \\ 1-10^{-20} \end{pmatrix}$. Si on approche w par $w' = \begin{pmatrix} 10^{20} \\ 1 \end{pmatrix}$, la solution exacte de $Ux = w'$ est $x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ très éloigné de la solution exacte u de $Au = b$. 2. La factorisation LU permet de calculer l'inverse d'une matrice A en environ $\frac{8n^3}{3}$ opérations ($\frac{2n^3}{3}$ opérations pour la factorisation et $2n(n^2)$ pour les $2n$ systèmes triangulaires).

1.2.3 Méthode de factorisation Cholesky

La méthode de factorisation de Cholesky correspond au cas particulier de la méthode LU pour une matrice symétrique définie positive :

Théorème 6. *Soit $A \in SDP_n(\mathbf{R})$ ($^tA = A$ et $\forall u \in \mathbf{R}^n \setminus \{0\}$, $^tuAu > 0$). Il existe une unique matrice $B = [b_{i,j}] \in TI_n(\mathbf{R})$ vérifiant*

$$\forall i \in \{1,...,n\}, \quad b_{i,i} > 0$$

et telle que

$$A = B\,{}^tB.$$

Démonstration. : Existence : on montre aisément que si $A \in SDP_n(\mathbf{R})$, alors $\Delta_k \in SDP_k(\mathbf{R})$ pour tout $k \in \{1,...,n\}$ (avec les notations du Théorème 4). En particulier, $\det\Delta_k \neq 0$ et le Théorème 4 s'applique : soient donc les matrices $L = [l_{i,j}]_{1\leq i,j\leq n} \in TI_n(\mathbf{R})$ et $U = [u_{i,j}]_{1\leq i,j\leq n} \in TS_n(\mathbf{R})$ telles que

$$\forall i \in \{1,...,n\}, \quad l_{i,i} = 1$$

et

$$A = LU.$$

D'après la démonstration de ce même théorème, on remarque en outre que

$$\forall i \in \{1,...,n\}, \quad u_{i,i} > 0$$

à l'aide de la relation (2) tronquée aux $k+1$ premières lignes :

$$\forall k \in \{1,...,n-1\}, \quad \det\Delta_{k+1} = \prod_{j=1}^{k+1} u_{j,j} > 0.$$

Soit alors la matrice diagonale $\Lambda = \mathrm{diag}(\sqrt{u_{1,1}},...,\sqrt{u_{n,n}})$. On peut écrire

$$A = LU = (L\Lambda)(\Lambda^{-1}U) = BC$$

avec $(B,C) \in TI_n(\mathbf{R}) \times TS_n(\mathbf{R})$. Le caractère symétrique de A se traduit par $(^tC)(^tB) = BC$ soit encore

$$\underbrace{B^{-1}(^tC)}_{\in TI_n(\mathbf{R})} = \underbrace{C(^tB)^{-1}}_{\in TS_n(\mathbf{R})}.$$

$B^{-1}(^tC)$ est donc une matrice diagonale dont le i-ième terme est égal à $\dfrac{1}{b_{i,i}}c_{i,i} = \dfrac{\sqrt{u_i}}{\sqrt{u_i}} = 1$. Comme précédemment, $B^{-1}(^tC) = I_n$ soit $C = {}^tB$ et l'existence de la factorisation de

Cholesky est démontrée. Unicité : on remarque que pour toute matrice B solution,

en notant

$$\Delta = \mathrm{diag}(b_{1,1}, ..., b_{n,n})$$

on peut écrire $A = LU$ où $L = B\Delta^{-1}$ et $U = \Delta^t B$ remplissent les conditions du Théorème 4. L'unicité de la factorisation LU implique donc que $\Delta^t B$ est une matrice indépendante de B : comme $b_{i,i} > 0$, on en déduit successivement que les termes diagonaux de B sont uniquement déterminés, soit Δ puis enfin B □

Remarque : En pratique, la matrice B est obtenue par la méthode dite des coefficients indéterminés qui consiste à résoudre le système $n^2 \times n^2$

$$a_{i,j} = \sum_{k=1}^{n} b_{i,k} b_{j,k} = \sum_{k=1}^{\min(i,j)} b_{i,k} b_{j,k}, \quad (1 \leq i, j \leq n)$$

en calculant d'abord la première colonne de B (correspondant à $i = 1$ dans le système) :

$$\begin{cases} b_{1,1} = \sqrt{a_{1,1}}, \\ b_{2,1} = \dfrac{a_{1,2}}{b_{1,1}}, \\ \ldots \\ b_{n,1} = \dfrac{a_{1,n}}{b_{1,1}}, \end{cases}$$

puis la seconde colonne (en fixant $i = 2$), et ainsi de suite jusqu'à déterminer $b_{n,n}$. Par exemple, pour la colonne i, on a

$$\begin{cases} b_{i,i} = \sqrt{a_{i,i} - \displaystyle\sum_{k=1}^{i-1} b_{i,k}^2}, \\ b_{i+1,i} = \dfrac{1}{b_{i,i}} \left(a_{i,i+1} - \displaystyle\sum_{k=1}^{i-1} b_{i,k} b_{i+1,k} \right), \\ \ldots \end{cases}$$

Au total, la résolution d'un système $Au = b$ par cette méthode nécessite n calculs de racines carrées, $\frac{n(n-1)}{2}$ divisions, $\frac{n(n^2-1)}{6}$ additions et $\frac{n(n^2-1)}{6}$ multiplications (avant la résolution de deux systèmes triangulaires en $2n^2$ opérations) soit de l'ordre de $\frac{n^3}{3}$ opérations et n calculs de racines carrées. Comparée à la méthode de Gauss, la méthode de Cholesky est donc environ deux fois plus rapide.

1.2.4 Méthode de factorisation QR

La méthode QR correspond à une autre factorisation possible d'une matrice $A \in GL_n(\mathbf{R})$: elle consiste à déterminer $Q \in O_n(\mathbf{R})$ et $R \in TS_n(\mathbf{R})$ tels que $A = QR$. Dans ce cas, la résolution du système $Au = b$ peut s'effectuer aisément en se ramenant au système triangulaire $Ru = {}^tQb$. L'existence et l'unicité d'une telle factorisation est assurée par le théorème suivant :

Théorème 7. *Soit $A \in GL_n(\mathbf{R})$. Il existe un unique couple de matrices $(Q,R) \in O_n(\mathbf{R}) \times TS_n(\mathbf{R})$ vérifiant*

$$\forall i \in \{1,...,n\}, \quad r_{i,i} > 0$$

et tel que

$$A = QR.$$

Démonstration. :

L'existence est issue du théorème d'orthonormalisation de Gram Schmidt : on note $(a_1,...,a_n)$ les vecteurs colonnes de A écrits dans la base canonique de $\mathbf{R}^n$. Ils forment une base ($A \in GL_n(\mathbf{R})$) qui peut s'orthonormaliser en une famille $(q_1,...,q_n)$. En notant Q la matrice constituée de ces nouveaux vecteurs colonnes, on montre aisément que $A = QR$ avec les propriétés indiquées dans l'énoncé.

Pour l'unicité, on suppose que $Q_1R_1 = Q_2R_2$ avec les propriétés indiquées. Alors, $T = {}^tQ_2Q_1 = R_2R_1^{-1} \in TS_n(\mathbf{R})$ vérifie en outre ${}^tTT = I_n$ et

$$\forall i \in \{1,...,n\}, \quad t_{i,i} > 0.$$

Par unicité de la décomposition de Cholesky de la matrice identité, on a alors $T = I$ soit $Q_1 = Q_2$ et $R_1 = R_2$ □

Remarque : L'algorithme de construction de Gram Schmidt n'est pas utilisé dans la pratique (car il n'est pas stable). On a plutôt recours à l'algorithme de Householder (voir Exercice 2) nécessitant $\frac{4n^3}{3}$ opérations élémentaires. Cette valeur est deux fois supérieure à celle des autres méthodes directes. Néanmoins, la factorisation QR possède l'avantage de pouvoir s'étendre à des matrices non carrées et permettre ainsi de résoudre efficacement les problèmes de moindre carrés (voir leçon A.VIII.). Elle est aussi utilisée pour calculer les valeurs propres d'une matrice (méthode QR ou de Householder- Givens : voir paragraphe 4.2.1).

1.3 Résolution d'un système linéaire : méthodes itératives

1.3.1 Principe général

Les méthodes itératives de résolution d'un système linéaire du type $Au = b$ consistent à construire une suite de solutions approchées $(u_k)_{k\in\mathbf{N}}$ issues de la résolution d'un système linéaire plus simple.

Définition 2. *Soit $A \in GL_n(\mathbf{R})$. On appelle décomposition régulière (ou "splitting") de A, la donnée d'un couple $(M,N) \in GL_n(\mathbf{R}) \times \mathscr{M}_n(\mathbf{R})$ tel que M soit une matrice facilement inversible (par exemple triangulaire) et*

$$A = M - N.$$

Une méthode itérative basée sur cette décomposition consiste alors à construire la suite $(u_k)_{k\in\mathbf{N}}$ telle que

$$u_0 \in \mathbf{R}^n \quad \text{et} \qquad Mu_{k+1} = Nu_k + b \quad (k \in \mathbf{N}).$$

On dit que cette méthode est convergente si pour toute valeur $u_0 \in \mathbf{R}^n$, la suite $(u_k)_{k\in\mathbf{N}}$ converge (la limite est alors nécessairement u solution de $Au = b$.

Une condition nécessaire et suffisante de convergence d'une méthode itérative est la suivante :

Théorème 8. *La méthode itérative associée à la décomposition (M,N) de A converge si et seulement si*

$$\rho(M^{-1}N) < 1$$

où ρ désigne le rayon spectral d'une matrice :

$$\rho(A) = \max_{\lambda\in\mathbf{C}}\{|\lambda| \ / \ \chi_A(\lambda) = 0\}$$

(et χ_A le polynôme caractéristique de A).

Démonstration. :

On commence par démontrer le lemme suivant :

Lemme 1. *soit $A \in \mathscr{M}_n(\mathbf{R})$ et $\varepsilon > 0$. Il existe une norme subordonée $|||.|||$ sur $\mathscr{M}_n(\mathbf{R})$ telle que*

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

Démonstration. Par le procédé de triangularisation d'une matrice complexe, il existe $U \in GL_n(\mathbf{C})$ tel que

$$T = U^{-1}AU = [t_{i,j}] \in TS_n(\mathbf{C}).$$

Soit $\delta > 0$ choisi tel que

$$\forall i \in \{1, ..., n-1\}, \quad \sum_{j=i+1}^{n} \delta^{j-i} |t_{i,j}| \leq \varepsilon.$$

On note alors

$$\begin{cases} D_\delta = \text{diag}(1, \delta, ..., \delta^{n-1}), \\ T_\delta = D_\delta^{-1} T D_\delta = (UD_\delta)^{-1} A (UD_\delta) \end{cases}$$

et on définit correctement pour tout $B \in \mathscr{M}_n(\mathbf{R})$ la norme suivante :

$$|||B||| = \max_{x \in \mathbf{R}^n,\ ||x||_\infty = 1} \left(||(UD_\delta)^{-1} B (UD_\delta) x||_\infty \right).$$

Par construction, on a $|||A||| \leq \rho(A) + \varepsilon$ et cette norme est subordonnée pour la norme sur $\mathbf{R}^n$ suivante :

$$||x|| = ||(UD_\delta)^{-1} x||_\infty \quad \square$$

Revenant à la démontration du théorème, on note $e_k = u_k - u$ l'erreur d'approximation au rang k ($k \in \mathbf{N}^*$) :

$$e_k = (M^{-1}Nu_{k-1} + M^{-1}b) - (M^{-1}Nu + M^{-1}b) = M^{-1}Ne_{k-1}$$

soit

$$e_k = (M^{-1}N)^k e_0.$$

Si $\rho(M^{-1}N) < 1$, on peut construire grâce au lemme précédent une norme subordonée $|||.|||$ telle que $|||M^{-1}N||| < 1$. En particulier, pour la norme associée sur $\mathbf{R}^n$

$$||e_k|| \leq |||M^{-1}N|||^k ||e_0||$$

et la suite $(e_k)_{k \in \mathbf{N}}$ converge bien vers 0. Réciproquement, si $\rho(M^{-1}N) \geq 1$, soit $\tilde{u} = \tilde{u}_1 + i\tilde{u}_2 \in \mathbf{C}^n$ un vecteur propre complexe de $M^{-1}N$ associé à une valeur propre de module supérieur ou égal à 1. Alors, la méthode itérative ne converge pas en partant de l'une des valeurs $u_0 = u + \tilde{u}_1$ ou $u_0 = u + \tilde{u}_2$ $\square$

Lorsque A est une matrice symétrique définie positive, une condition suffisante de convergence peut être explicitée :

Théorème 9. *Soit $A \in SDP_n(\mathbf{R})$. Pour la méthode itérative asociée à la décomposition (M,N) de A, on a ${}^tM+N \in S_n(\mathbf{R})$. Si de plus ${}^tM+N \in SDP_n(\mathbf{R})$, alors la méthode converge.*

Démonstration. On a par définition de M et N :

$$ {}^t({}^tM+N) = M+{}^tN = A+N+{}^tN = {}^tA+{}^tN+N = {}^tM+N $$

On note ensuite correctement $||.||_A$ la norme sur $\mathbf{R}^n$ telle que

$$ ||x||_A = \sqrt{{}^txAx} = \sqrt{<Ax,x>} $$

et $|||.|||_A$ la norme subordonnée associée sur $\mathscr{M}_n(\mathbf{R})$. On montre que si la matrice ${}^tM+N \in SDP_n(\mathbf{R})$, alors

$$ |||M^{-1}N|||_A < 1 $$

(ce qui implique bien que $\rho(M^{-1}N) \leq |||M^{-1}N|||_A < 1$).

Pour cela, soit $v \in \mathbf{R}^n$ tel que $||v||_A = 1$. On calcule

$$\begin{aligned}
||M^{-1}Nv||_A^2 &= <AM^{-1}Nv, M^{-1}Nv> \\
&= <AM^{-1}(M-A)v, M^{-1}(M-A)v> \\
&= <Av - AM^{-1}Av, v - M^{-1}Av> \\
&= <Av,v> - <AM^{-1}Av, v> + <AM^{-1}Av, M^{-1}Av> - <Av, M^{-1}Av> \\
&= 1 - <M^{-1}Av, MM^{-1}Av> + <Aw,w> - <MM^{-1}Av, M^{-1}Av> \\
&= 1 - <w, Mw> + <Aw,w> - <Mw,w> \\
&= 1 - <({}^tM+N)w, w>
\end{aligned}$$

où on a noté $w = M^{-1}Av \neq 0$.

Comme $<({}^tM+N)w,w> \in \mathbf{R}_+^*$ on a bien le résultat annoncé □

Quatre exemples de méthodes itératives (Jacobi, Gauss Seidel, relaxation et gradient) sont à présent construites et comparées entre elles.

1.3.2 Méthode de Jacobi

Définition 3. *On appelle méthode de Jacobi, la méthode itérative associée à la décomposition de A suivante*

$$ \begin{cases} M = \text{diag}(a_{1,1}, ..., a_{n,n}) = D, \\ N = D - A. \end{cases} $$

On note alors $J = M^{-1}N = I - D^{-1}A$.

Pour définir cette méthode, il est nécessaire de supposer que $M \in GL_n(\mathbf{R})$. Ceci est en particulier vérifié lorsque $A \in SDP_n(\mathbf{R})$. Dans ce cas, la méthode de Jacobi est convergente si $2D - A \in SDP_n(\mathbf{R})$ (d'après le Théorème 9). Dans tous les cas, lorsque la méthode est bien définie, la suite des itérées $(u^k \equiv (u_1^k, ..., u_n^k))_{k\in\mathbf{N}}$ est construite à partir des relations

$$u_i^{k+1} = \frac{1}{a_{i,i}}\left(-a_{i,1}u_1^k - ... - a_{i,i-1}u_{i-1}^k - a_{i,i+1}u_{i+1}^k ... - a_{i,n}u_n^k + b_i\right).$$

1.3.3 Méthode de Gauss Seidel

Définition 4. *On appelle méthode de Gauss Seidel, la méthode itérative associée à la décomposition de A suivante*

$$\begin{cases} M = \text{diag}(a_{1,1}, ..., a_{n,n}) + A_{\text{inf}} = D - E, \\ N = -A_{\text{sup}} = F, \end{cases}$$

où A_{inf} (respectivement A_{sup}) désigne la matrice triangulaire inférieure (respectivement supérieure) consituée des éléments de A situés strictement en dessous (respectivement au dessus) de la diagonale. On note alors

$$\mathscr{L}_1 = M^{-1}N = (D - E)^{-1}F.$$

Les même restrictions d'existence que pour la méthode de Jacobi s'appliquent. Par contre, lorsque $A \in SDP_n(\mathbf{R})$, la méthode de Gauss Seidel est bien définie et converge toujours (car ${}^tM + N = D \in SDP_n(\mathbf{R})$).

Dans tous les cas, lorsque la méthode est bien définie, la suite des itérées $(u^k \equiv (u_1^k, ..., u_n^k))_{k\in\mathbf{N}}$ est construite à partir des relations

$$u_i^{k+1} = \frac{1}{a_{i,i}}\left(-a_{i,1}u_1^{k+1} - ... - a_{i,i-1}u_{i-1}^{k+1} - a_{i,i+1}u_{i+1}^k ... - a_{i,n}u_n^k + b_i\right).$$

1.3.4 Comparaison de la convergence des méthodes de Jacobi et Gauss Seidel

Théorème 10. *Soit A une matrice tridiagonale. Si les méthodes de Jacobi et Gauss Seidel sont définies, alors*

$$\rho(\mathscr{L}_1) = \rho(J)^2.$$

En particulier, les deux méthodes convergent simultanément et la méthode de Gauss Seidel converge plus vite (de manière générale) que la méthode de Jacobi.

Démonstration. :

On prouve tout d'abord le lemme suivant :

Lemme 2. *oit $\mu \in \mathbf{C}^*$ et $A(\mu)$ la matrice tridiagonale :*

$$A(\mu) = \begin{pmatrix} b_1 & \frac{c_1}{\mu} & 0 & \dots & 0 \\ \mu a_2 & b_2 & \frac{c_2}{\mu} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \mu a_{n-1} & b_{n-1} & \frac{c_{n-1}}{\mu} \\ 0 & \dots & 0 & \mu a_n & b_n \end{pmatrix}.$$

Alors $\forall \mu > 0, \quad \det A(\mu) = \det A(1)$.

Démonstration. Il suffit de remarquer que $A(\mu) = Q(\mu)A(1)Q(\mu)^{-1}$ avec $Q(\mu) = \mathrm{diag}(\mu, ..., \mu^n)$ □

Revenant à la démonstration du théorème, on suppose que A est tridiagonale et que les méthodes de Jacobi et de Gauss Seidel sont bien définies. On écrit pour tout $\lambda \in \mathbf{C}^*$

$$\chi_J(\lambda) = \det\big((I - D^{-1}(D - E - F)) - \lambda I\big) = \det(-D^{-1}) \det(\lambda D - E - F)$$

et

$$\chi_{\mathscr{L}_1}(\lambda^2) = \det\big((D - E)^{-1}F - \lambda^2 I\big) = \det((D - E)^{-1}) \det(\lambda^2 D - \lambda^2 E - F).$$

En notant alors pour tout $\mu \in \mathbf{C}^*, A(\mu) = \lambda^2 D - \mu\lambda^2 E - \dfrac{1}{\mu}F$, il est possible d'appliquer le Lemme précédent pour déduire que

$$\det A(\frac{1}{\lambda}) = \det A(1)$$

puis ensuite

$$\exists C \in \mathbf{R}^*, \quad \forall \lambda \in \mathbf{C}, \quad \chi_{\mathscr{L}_1}(\lambda^2) = C\lambda^n \chi_J(\lambda).$$

En particulier, on a bien $\rho(\mathscr{L}_1) = \rho(J)^2$ □

1.3.5 Méthode de relaxation

Définition 5. *On appelle méthode de relaxation (ou SOR pour successive over relaxation) la méthode itérative qui généralise la méthode de Gauss Seidel en introduisant un paramètre de relaxation $\omega > 0$ dans la décomposition de A :*

$$\begin{cases} M = \dfrac{1}{\omega}\mathrm{diag}(a_{1,1},...,a_{n,n}) + A_{\mathrm{inf}} = \dfrac{D}{\omega} - E, \\ N = \dfrac{(1-\omega)}{\omega}D - A_{\mathrm{sup}} = \dfrac{(1-\omega)}{\omega}D + F. \end{cases}$$

On note alors $\mathscr{L}_\omega = M^{-1}N$

Lorsque $\omega = 1$, on retrouve la méthode de Gauss Seidel. Lorsque $\omega > 1$ (respectivement $\omega < 1$) on parle de sur (respectivement sous) relaxation. On démontre le théorème de convergence suivant :

Théorème 11. *La méthode de relaxation ne peut converger que si $\omega \in]0,2[$. Réciproquement, si $A \in SDP_n(\mathbf{R})$, la méthode de relaxation est bien définie et converge pour tout $\omega \in]0,2[$.*

Démonstration. Pour tout $\omega > 0$ pour lequel la méthode est définie, on a

$$\begin{aligned} |\det(\mathscr{L}_\omega)| &= \left| \frac{\det(\frac{(1-\omega)}{\omega}D + F)}{\det(\frac{D}{\omega} - E)} \right| = |1-\omega|^n \\ &= |\prod_{i=1}^{n} \lambda_i(\mathscr{L}_\omega)| \leq \rho(\mathscr{L}_\omega)^n \end{aligned}$$

où $(\lambda_i(\mathscr{L}_\omega))_{1\leq i\leq n}$ désigne la famille des racines complexes comptées avec leur multiplicité du polynôme caractéristique de $\mathscr{L}_\omega$.

En particulier, lorsque $\omega \geq 2$, on a $\rho(\mathscr{L}_\omega) \geq |\omega - 1| \geq 1$, et la méthode ne peut converger suivant le Théorème 8. Réciproquement, si $A \in SDP_n(\mathbf{R})$ et $\omega \in]0,2[$, la méthode est bien définie et converge car

$${}^tM + N = \frac{(2-\omega)}{\omega}D \in SDP_n(\mathbf{R}).$$

Le Théorème suivant (dont la démonstration peut être trouvée dans [Ci]) justifie l'introduction du paramètre de relaxation ω afin d'optimiser la convergence de la méthode de Gauss Seidel :

Théorème 12. *Soit A une matrice tridiagonale, symétrique, définie positive. Les trois méthodes présentées (Jacobi, Gauss Seidel, relaxation) convergent et*

$$\exists!\, \omega_{\text{opt}} \in]0,2[, \quad \rho(\mathscr{L}_{\omega_{\text{opt}}}) = \inf_{\omega\in]0,2[} \rho(\mathscr{L}_\omega).$$

De plus, $\omega_{\text{opt}} = \dfrac{2}{1+\sqrt{1-\rho(J)^2}} \geq 1$ *et* $\rho(\mathscr{L}_{\omega_{\text{opt}}}) = \omega_{\text{opt}} - 1$.

1.3.6 Méthode du gradient

Définition 6. *On appelle méthode du gradient, la méthode itérative associée à la décomposition de A suivante :*

$$\begin{cases} M = \dfrac{1}{\alpha} I, \\ N = \dfrac{1}{\alpha} I - A, \end{cases}$$

où $\alpha \in \mathbf{R}^*$ *est fixé. On note alors* $\mathscr{G}_\alpha = M^{-1}N$.

La méthode du gradient est toujours définie et la suite des itérées est donnée par la relation de récurrence

$$u_{k+1} = u_k + \alpha(b - Au_k) \quad (k \geq 0).$$

Le théorème suivant précise certains cas de convergence de la méthode :

Théorème 13. *Soit* $A \in GL_n(\mathbf{R})$ *une matrice diagonalisable et de valeurs propres*

$$\lambda_1 \leq \ldots \leq \lambda_n.$$

(i) Si $\lambda_1 \leq 0 \leq \lambda_n$, *la méthode du gradient ne converge pour aucune valeur de* α.

(ii) Si $0 < \lambda_1 \leq \ldots \leq \lambda_n$, *la méthode du gradient converge si et seulement si* $\alpha \in]0, \dfrac{2}{\lambda_n}[$

et le paramètre optimal qui minimise $\rho(\mathscr{G}_\alpha)$ *est* $\alpha_{opt} = \dfrac{2}{\lambda_1 + \lambda_n}$ *pour lequel*

$$\rho(\mathscr{G}_{\alpha_{opt}}) = \frac{\text{cond}_2 A - 1}{\text{cond}_2 A + 1}.$$

Démonstration. On calcule pour tout $\alpha \in \mathbf{R}^*$,

$$\rho(\mathscr{G}_\alpha) = \rho(I - \alpha A) = \max_{1\leq i\leq n} |1 - \alpha\lambda_i|.$$

Pour que $\rho(\mathscr{G}_\alpha) < 1$, il faut et il suffit que

$$\forall i \in \{1, ..., n\}, \quad 0 < \alpha\lambda_i < 2.$$

La première inégalité implique que toutes les valeurs propres de A doivent être non nulles et du même signe. Dans ce cas, en supposant celles-ci toutes positives, la condition nécessaire et suffisante de convergence se traduit par

$$0 < \alpha < \frac{2}{\lambda_n}.$$

On observe ensuite que $\rho(\mathscr{G}_\alpha) = \max(|1-\alpha\lambda_1|, |1-\alpha\lambda_n|)$. Cette fonction de la variable α est continue sur $[0, \frac{2}{\lambda_n}]$ et atteint donc son minimum en α_0 à l'intérieur de l'intervalle (aux deux bornes, $\rho(\mathscr{G}_\alpha)$ est en effet égal à sa valeur maximale, à savoir 1).
Nécessairement, on a $|1-\alpha_0\lambda_1| = |1-\alpha_0\lambda_n|$, soit $\alpha_0 = \alpha_{opt} = \frac{2}{\lambda_1+\lambda_n}$ pour lequel

$$\rho(\mathscr{G}_{\alpha_{opt}}) = 1 - \frac{2\lambda_1}{\lambda_1+\lambda_n} = \frac{\frac{\lambda_n}{\lambda_1}-1}{\frac{\lambda_n}{\lambda_1}+1} = \frac{\text{cond}_2 A - 1}{\text{cond}_2 A + 1}.$$

En effet, en supposant par exemple

$$|1-\alpha_0\lambda_1| < |1-\alpha_0\lambda_n|$$

on aurait dans un voisinage de α_0,

$$\rho(\mathscr{G}_\alpha) = |1-\alpha\lambda_n|$$

et nécessairement, $\alpha_0 = \frac{1}{\lambda_n}$ pour lequel

$$\rho(\mathscr{G}_{\alpha_0}) = 1 - \frac{\lambda_1}{\lambda_n} > \rho(\mathscr{G}_{\alpha_{opt}}).$$

□

Remarque : La méthode du gradient peut aussi être interprétée comme une méthode de recherche du minimum d'une fonctionnelle quadratique par un procédé de descente. Pour plus de détails, le lecteur est appelé à se reporter à la leçon A.VIII. où d'autres méthodes de gradient (à pas variable, conjugué) sont construites dans ce contexte mais pouvent aussi être utilisées dans la résolution de systèmes linéaires.

1.4 Recherche d'éléments propres d'une matrice

Le problème de la recherche des valeurs propres et des vecteurs propres d'une matrice est un sujet particulièrement important et difficile (voir exemples en mécanique dans [Ci] et [Sch]). Outre les problèmes de conditionnement évoqués au paragraphe 1, il n'existe pas de méthode directe de recherche de valeurs propres au delà de $n = 5$ (puisque celle-ci résoudrait en particulier le problème de la recherche des racines d'un polynôme de degré n). On présente ici en détail la méthode de Jacobi pour les matrices symétriques et plus succintement trois autres méthodes de calcul approché de valeurs propres et/ou de vecteurs propres.

1.4.1 Méthode de Jacobi

On démontre en préambule de la construction de la méthode le lemme suivant :

Lemme 3. *Soit la matrice $Q_{p,q}(\theta)$ représentant dans la base canonique $(e_1, ..., e_n)$ la rotation d'angle $\theta \in \mathbf{R}$ dans le plan (e_p, e_q) $(p \neq q)$:*

$$Q_{p,q}(\theta)(e_i) = \begin{cases} e_i \quad \text{si} \quad i \neq p, q, \\ \cos\theta e_p - \sin\theta e_q \quad \text{si} \quad i = p, \\ \sin\theta e_p + \cos\theta e_q \quad \text{si} \quad i = q. \end{cases}$$

Alors, $Q_{p,q}(\theta) \in O_n(\mathbf{R})$ et pour tout $A = [a_{i,j}] \in S_n(\mathbf{R})$, la matrice

$$B = {}^tQ_{p,q}(\theta) A Q_{p,q}(\theta) = [b_{i,j}]$$

vérifie

$$\sum_{i,j=1}^{n} b_{i,j}^2 = \sum_{i,j=1}^{n} a_{i,j}^2.$$

En outre, pour tout couple $(p, q) \in \{1, ..., n\}^2$ avec $p \neq q$ et $a_{p,q} \neq 0$, il existe un réel $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}] \setminus \{0\}$ tel que la matrice $B = {}^tQ_{p,q}(\theta) A Q_{p,q}(\theta)$ vérifie

$$\begin{cases} b_{p,q} = 0, \\ \displaystyle\sum_{i=1}^{n} b_{i,i}^2 = \sum_{i=1}^{n} a_{i,i}^2 + 2a_{p,q}^2. \end{cases}$$

Démonstration. : La première partie du lemme se démontre à l'aide des propriétés

de la trace :

$$\begin{aligned}\mathrm{Tr}({}^tBB) &= \sum_{i,j=1}^{n} b_{i,j}^2 \\ &= \mathrm{Tr}({}^tQ_{p,q}(\theta){}^tAQ_{p,q}(\theta){}^tQ_{p,q}(\theta)AQ_{p,q}(\theta)) = \mathrm{Tr}({}^tAA) = \sum_{i,j=1}^{n} a_{i,j}^2\end{aligned}$$

Dans la deuxième partie, on traite d'abord le cas $n = 2$ en calculant explicitement B en fonction de A :

$$\begin{cases} b_{1,1} = -2a_{1,2}\sin\theta\cos\theta + a_{1,1}\cos^2\theta + a_{2,2}\sin^2\theta, \\ b_{1,2} = b_{2,1} = a_{2,1}(\cos^2\theta - \sin^2\theta) + (a_{1,1} - a_{2,2})\sin\theta\cos\theta, \\ b_{2,2} = 2a_{1,2}\sin\theta\cos\theta + a_{1,1}\sin^2\theta + a_{2,2}\cos^2\theta. \end{cases}$$

On peut alors choisir $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}] \setminus \{0\}$ tel que

$$b_{1,2} = a_{2,1}\cos 2\theta + \frac{a_{1,1} - a_{2,2}}{2}\sin 2\theta = 0$$

car $a_{2,1}$ est non nul et la fonction cotangente est surjective de $[-\frac{\pi}{2}, \frac{\pi}{2}] \setminus \{0\}$ sur $\mathbf{R}$. Dans ce cas, l'égalité obtenue dans la première partie se réduit à :

$$b_{1,1}^2 + b_{2,2}^2 = a_{1,1}^2 + a_{2,2}^2 + 2a_{1,2}^2.$$

Le cas $n > 2$ est analogue car seules les lignes et colonnes p et q de A sont modifiées par l'action de $Q_{p,q}$. Il suffit donc de reprendre le cas $n = 2$ en considérant la matrice extraite

$$\begin{pmatrix} a_{p,p} & a_{p,q} \\ a_{p,q} & a_{q,q} \end{pmatrix}$$

□

On démontre à présent le résultat suivant servant de définition à la méthode de Jacobi :

Théorème 14. *Soit $A \in S_n(\mathbf{R})$. On peut construire la suite des matrices $(A_k \equiv [a_{i,j}^k])_{k\in\mathbf{N}^*}$ telle que*

$$\begin{cases} A_1 = A, \\ A_{k+1} = {}^tQ_{p_k,q_k}(\theta_k)A_kQ_{p_k,q_k}(\theta_k) \quad (k \in \mathbf{N}^*), \end{cases}$$

où (p_k, q_k) représente un couple d'indices tel que $|a_{p_k,q_k}^k| = \max_{i\neq j} |a_{i,j}^k|$ et θ_k est un angle tel que $a_{p_k,q_k}^{k+1} = 0$ (voir Lemme précédent).

Une telle suite converge et il existe une permutation σ de $\{1, ..., n\}$ telle que

$$\lim_{k\to+\infty} A_k = \mathrm{diag}(\lambda_{\sigma(1)}, ..., \lambda_{\sigma(n)})$$

où $(\lambda_i)_{1\leq i\leq n}$ désigne l'ensemble des valeurs propres de A.

Démonstration. : On démontre tout d'abord le lemme technique suivant :

Lemme 4. *Soit $(M_k)_{k\in\mathbf{N}}$ une suite dans $\mathbf{R}^n$ telle que (i)* $\lim_{k\to+\infty}(M_{k+1}-M_k)=0$.

(ii) la suite $(M_k)_{k\in\mathbf{N}}$ est bornée.

(iii) la suite $(M_k)_{k\in\mathbf{N}}$ a un nombre fini de points d'accumulation.

Sous ces hypothèses, la suite $(M_k)_{k\in\mathbf{N}}$ converge.

Démonstration. : On note A_1, ...,A_p les points d'accumulation de la suite $(M_k)_{k\in\mathbf{N}}$.

On remarque tout d'abord en raisonnant par l'absurde, que pour une norme quelconque $||.||$ sur $\mathbf{R}^n$, on a

$$\forall\varepsilon>0,\quad \exists K\in\mathbf{N}^*,\quad \forall k\geq K,\quad ||M_{k+1}-M_k||\leq\varepsilon\quad\text{et}\quad M_k\in\bigcup_{i=1}^{p}B(A_i,\varepsilon)$$

en notant $B(A_i,\varepsilon)$ la boule ouverte de $\mathbf{R}^n$ de centre A_i et de rayon ε (sinon, il existerait une suite extraite de $(M_k)_{k\in\mathbf{N}}$ en dehors de la réunion des boules précédentes. La suite étant bornée, elle possèderait un point d'accumulation qui ne serait égal à aucun des $(A_i)_{1\leq i\leq p}$).

En raisonnant par l'absurde (c'est à dire en supposant simplement au vu de l'hypothèse (ii) que $p\geq 2$), on choisit

$$\varepsilon_0=\frac{1}{4}\min_{i\neq j}||A_i-A_j||$$

et on note K_0 l'entier associé à ε_0 dans la relation précédente. On montre que

$$\exists i_0\in\{1,...,p\},\quad \forall k\geq K_0,\quad M_k\in B(A_{i_0},\varepsilon_0)$$

(ce qui contredit l'hypothèse $p\geq 2$). En effet, en raisonnant à nouveau par l'absurde, on aboutit à l'existence de $k_0\geq K_0$ tel que M_{k_0} et M_{k_0+1} soient dans deux boules différentes. En particulier

$$||M_{k_0+1}-M_{k_0}||\geq\min_{i\neq j}||A_i-A_j||-2\varepsilon_0=2\varepsilon_0>\varepsilon_0$$

ce qui est absurde □

La démonstation du Théorème 14 peut être à présent entreprise : on écrit

$$A_k=D_k+B_k$$

avec $D_k=\mathrm{diag}(a^k_{1,1},...,a^k_{n,n})$ et on note $\varepsilon_k=\mathrm{Tr}({}^tB_kB_k)$ $(=||B_k||^2)$.

Par construction et grâce au Lemme démontré précédemment

$$\varepsilon_{k+1} = \varepsilon_k - 2(a^k_{p_k,q_k})^2.$$

Comme $a^k_{p_k,q_k}$ est la plus grande valeur en module de B_k, on peut écrire que $\varepsilon_k \leq (n^2-n)|a^k_{p_k,q_k}|^2$ ce qui donne alors

$$0 \leq \varepsilon_{k+1} \leq (1 - \frac{2}{n(n-1)})\varepsilon_k$$

puis facilement $\lim_{k\to+\infty} \varepsilon_k = 0$. Il reste donc seulement à montrer que

$$\lim_{k\to+\infty} D_k = \mathrm{diag}(\lambda_{\sigma(1)}, ..., \lambda_{\sigma(n)}).$$

Pour cela, on montre que la suite $(D_k)_{k\in\mathbf{N}^*}$ dans $\mathbf{R}^{n^2}$ vérifie les hypothèses du Lemme précédemment démontré en utilisant toujours la norme $||D_k|| = \sqrt{\mathrm{Tr}({}^tD_kD_k)}$.

La suite $(D_k)_{k\in\mathbf{N}^*}$ est tout d'abord bornée grâce au Lemme . Ensuite, si D est un point d'accumulation de $(D_k)_{k\in\mathbf{N}^*}$, donc de $(A_k)_{k\in\mathbf{N}^*}$, on a en particulier

$$\forall \lambda \in \mathbf{R}, \quad \det(D-\lambda I) = \lim_{k\to+\infty} \det(A_k - \lambda I) = \det(A-\lambda I)$$

par construction et D ne peut être que de la forme $D = \mathrm{diag}(\lambda_{\sigma(1)}, ..., \lambda_{\sigma(n)})$.

Il suffit donc pour conclure la démonstration du Théorème 14 de montrer que

$$\lim_{k\to+\infty} ||D_{k+1} - D_k|| = 0.$$

Pour cela, on remarque tout d'abord que si $k \neq p_k$ ou q_k, alors $a^{k+1}_{k,k} = a^k_{k,k}$. Il suffit donc de déterminer $a^{k+1}_{p_k,p_k} - a^k_{p_k,p_k}$. En reprenant le calcul du Lemme précédent, il vient

$$\begin{aligned} a^{k+1}_{p_k,p_k} - a^k_{p_k,p_k} &= -2a^k_{p_k,q_k}\sin\theta_k\cos\theta_k + (a^k_{q_k,q_k} - a^k_{p_k,p_k})\sin^2\theta_k \\ &= -2a^k_{p_k,q_k}\sin\theta_k\cos\theta_k + a^k_{p_k,q_k}\frac{\sin^2\theta_k}{\sin\theta_k\cos\theta_k}(1-2\sin^2\theta_k) \\ &= -a^k_{p_k,q_k}\tan\theta_k \end{aligned}$$

grâce à la définition de θ_k.

Comme de plus $\theta_k \in [-\frac{\pi}{4}, \frac{\pi}{4}]$, on a alors $||D_{k+1} - D_k||^2 \leq 2(a^k_{p_k,q_k})^2$ et finalement

$$\lim_{k\to+\infty} ||D_{k+1} - D_k|| = 0$$

□

On peut également dans certains cas à partir de la méthode de Jacobi obtenir une approximation des vecteurs propres de A :

Théorème 15. *Les notations du Théorème 14 sont conservées. Si les valeurs propres de A sont simples, alors la suite de matrices orthogonales*

$$Q_k = \prod_{i=1}^{k} Q_{p_i,q_i}(\theta_i)$$

converge vers une matrice orthogonale dont les vecteurs colonnes sont les vecteurs propres de A rangés dans le même ordre que les valeurs propres dans la matrice limite de la suite $(A_k)_{k\in\mathbf{N}^*}$.

Démonstration. : On applique à nouveau le Lemme précédemment démontré pour la suite $(Q_k)_{k\in\mathbf{N}^*}$: cette suite est bornée ($\mathrm{tr}({}^tQ_kQ_k)$ vaut n). D'autre part, si Q est une valeur d'adhérence, on a

$$\lim_{k\to+\infty} A_{\alpha(k)} = \mathrm{diag}(\lambda_{\sigma(1)},...,\lambda_{\sigma(n)}) = \lim_{k\to+\infty} {}^tQ_{\alpha(k)}AQ_{\alpha(k)} = {}^tQAQ$$

et Q est bien formé de vecteurs propres de A. Enfin,

$$Q_{k+1} - Q_k = (Q_{p_{k+1},q_{k+1}}(\theta_{k+1}) - I)Q_k$$

et il suffit de montrer que $\lim_{k\to+\infty} \theta_k = 0$. Ceci est vérifié car pour pour tout k assez grand,

$$|a^k_{q_k,q_k} - a^k_{p_k,p_k}| \geq \frac{1}{2}\min_{i\neq j}|\lambda_i - \lambda_j| = C$$

puisque les valeurs propres de A sont distinctes et

$$|\tan 2\theta_k| = |\frac{2a^k_{p_k,q_k}}{a^k_{q_k,q_k} - a^k_{p_k,p_k}}| \leq \frac{2a^k_{p_k,q_k}}{C}.$$

La convergence vers zéro de la suite $(a^k_{p_k,q_k})_{k\in\mathbf{N}^*}$ démontrée dans le Théorème 14 permet alors de conclure □

1.4.2 Autres méthodes

1.4.2.1 Méthode QR

La méthode QR est une méthode de recherche des valeurs propres d'une matrice basée sur la factorisation QR (voir paragraphe 2.4) et valable pour une large classe de matrices. Le théorème suivant peut en particulier être démontré (voir [Ci] ou [BuSt]) :

Théorème 16. *Soit $A \in GL_n(\mathbf{C})$ une matrice diagonalisable dont les valeurs propres sont toutes de module différent. Soit $P \in GL_n(\mathbf{C})$ tel que $A = P\Lambda P^{-1}$ avec*

$$\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n), \quad |\lambda_1| > ... > |\lambda_n| > 0.$$

On suppose que la matrice P^{-1} admet une factorisation LU. Alors, la suite de matrices $(A_k)_{k\in\mathbf{N}^*}$ construite à partir des factorisations QR successives

$$\begin{cases} A_1 = A = Q_1 R_1, \\ A_2 = R_1 Q_1 = Q_2 R_2, \\ ... \\ A_n = R_{n-1} Q_{n-1} = Q_n R_n, \end{cases}$$

est telle que

$$\begin{cases} \lim\limits_{k\to+\infty} (A_k)_{i,i} = \lambda_i, \quad 1 \le i \le n, \\ \lim\limits_{k\to+\infty} (A_k)_{i,j} = 0, \quad 1 \le j < i \le n. \end{cases}$$

1.4.2.2 Méthode de Givens-Householder

La méthode de Givens-Householder permet d'approcher toutes (ou une partie) des valeurs propres d'une matrice symétrique en se ramenant tout d'abord au cas d'une matrice tridiagonale grâce aux matrices de Householder (voir Exercice 2). Les valeurs propres dans ce cas sont ensuite approchées avec la précision souhaitée grâce à une suite de Sturm de polynômes (voir [Ci] ou [BuSt] pour une construction complète).

1.4.2.3 Méthode de la puissance inverse

Le théorème suivant (démontré dans [Ci]) propose une méthode d'approximation d'un vecteur propre associé à une valeur propre préalablement estimée :

Théorème 17. *Soit $A \in \mathscr{M}_n(\mathbf{C})$ une matrice diagonalisable et λ une valeur propre de A ($\lambda \in \mathrm{Sp}(A)$). Soit $\tilde{\lambda}$ une valeur approchée de λ telle que*

$$\tilde{\lambda} \neq \lambda \quad \text{et} \quad \forall \mu \in Sp(A) \setminus \{\lambda\}, \quad |\tilde{\lambda} - \lambda| < |\tilde{\lambda} - \mu|.$$

Soit $u_0 \in \mathbf{C}^n$ non contenu dans le sous-espace engendré par les valeurs propres autres que λ. Alors la suite $(u_k)_{k\in\mathbf{N}}$ telle que

$$\forall n \in \mathbf{N}, \quad (A - \tilde{\lambda} I) u_{n+1} = u_n$$

est telle que

$$\lim_{k\to+\infty} \left(\frac{(\lambda - \tilde{\lambda})^k}{|\lambda - \tilde{\lambda}|^k} \frac{u_k}{||u_k||} \right) = q$$

où q est un vecteur propre de A associé à la valeur propre λ.

1.5 Fonctionnelles quadratiques sur $\mathbf{R^n}$

1.5.1 Définition

Définition 7. *On appelle fonctionnelle quadratique sur* $\mathbf{R^n}$ *toute application* $J : \mathbf{R}^n \to \mathbf{R}$ *telle que*

$$J(x) = \frac{1}{2} < Ax, x > - < b, x > + K$$

où $A \in SDP_n(\mathbf{R})$ *(matrice symétrique définie positive),* $b \in \mathbf{R}^n$ *et* $K \in \mathbf{R}$.

1.5.2 Minimisation d'une fonctionnelle quadratique sur $\mathbf{R}^n$

Théorème 18. *Soit* $A \in SDP_n(\mathbf{R})$ *et* $b \in \mathbf{R}^n$. *On considère la fonctionnelle quadratique* J *de* $\mathbf{R}^n$ *dans* $\mathbf{R}$ *suivante :*

$$J(x) = \frac{1}{2} < Ax, x > - < b, x > + K = \frac{1}{2} \sum_{i,j=1}^{n} a_{i,j} x_i x_j - \sum_{i=1}^{n} b_i x_i + K.$$

J est une fonction strictement convexe et atteint son minimum en un unique point. Celui-ci, noté u, est caractérisé par l'annulation du gradient de J :

$$\nabla J(u) = Au - b = 0.$$

De plus, pour tout $x \in \mathbf{R}^n \setminus \{u\}$, *on a*

$$\forall \alpha \in]0, \frac{2}{\rho(A)}[, \quad J(x - \alpha \nabla J(x)) < J(x).$$

Démonstration. : On montre que J est strictement convexe en vérifiant que si $(\theta, x, y) \in [0,1] \times (\mathbf{R}^n)^2$, on a

$$J(\theta x + (1-\theta)y) = \theta J(x) + (1-\theta)J(y) - \theta(1-\theta)\overbrace{< A(x-y), x-y >}^{\geq 0}$$

On note ensuite $(e_i)_{1 \leq i \leq n}$ une base orthonormée de vecteurs propres de A associée aux valeurs propres strictement positives $(\lambda_i)_{1 \leq i \leq n}$. Si on décompose x et b dans cette base :

$$\begin{cases} x = \sum_{i=1}^{n} x_i e_i, \\ b = \sum_{i=1}^{n} b_i e_i, \end{cases}$$

on remarque que

$$J(x) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i x_i^2 - \sum_{i=1}^{n} b_i x_i + K = \frac{1}{2}\sum_{i=1}^{n} \left(\lambda_i (x_i - \frac{b_i}{\lambda_i})^2 - \frac{b_i^2}{\lambda_i} \right) + K$$

et le minimum de J est bien uniquement atteint en $u = \sum_{i=1}^{n} \frac{b_i}{\lambda_i} e_i$ tel que

$$\nabla J(u) = Au - b = 0.$$

Pour finir, soit $x \in \mathbf{R}^n \setminus \{u\}$, on note $\delta = -\alpha \nabla J(x)$ ($\alpha \in \mathbf{R}$). Alors

$$\begin{aligned} J(x+\delta) &= J(x) + \frac{1}{2} < A\delta, \delta > + < Ax - b, \delta > \\ &= J(x) + \frac{1}{2}\alpha^2 < A\nabla J(x), \nabla J(x) > - \alpha < \nabla J(x), \nabla J(x) > \\ &\leq J(x) + (\frac{1}{2}\alpha^2 \rho(A) - \alpha) < \nabla J(x), \nabla J(x) > \end{aligned}$$

et $J(x+\delta) < J(x)$ si $\alpha \in]0, \frac{2}{\rho(A)}[$ □

1.5.3 Minimisation d'une fonctionnelle quadratique avec contrainte linéaire

En conservant les notations du sous-paragraphe précédent, on cherche à présent à minimiser la fonctionnelle quadratique J sur l'ensemble

$$V = \{v \in \mathbf{R}^n \ / \ Cv = d\}$$

avec $C \in \mathscr{M}_{p,n}(\mathbf{R})$ et $d \in \mathbf{R}^p$ ($p < n$).

On montre aisément que si le rang de C est égal à p, une condition nécessaire pour que J soit extrémale sur V en $u \in V$ est qu'il existe $(u, \lambda) \in \mathbf{R}^n \times \mathbf{R}^p$ solution du système linéaire de taille $(n+p)$:

$$\begin{cases} Au + {}^tC\lambda = b, \\ \quad Cu + 0 = d. \end{cases}$$

1.5.4 Problème aux moindres carrés

Le problème suivant, souvent rencontré dans la pratique, se ramène dans certains cas à la minimisation d'une fonctionnelle quadratique : Définition : Soit $B \in \mathscr{M}_{m,n}(\mathbf{R})$

et $c \in \mathbf{R}^m$ avec $n \leq m$. On appelle problème aux moindres carrés (sans contrainte) la recherche des vecteurs $u \in \mathbf{R}^n$ minimisant la fonctionnelle $J : \mathbf{R}^n \to \mathbf{R}$ telle que

$$J(x) = ||c - Bx||^2_{\mathbf{R}^m} =< c - Bx, c - Bx >_{\mathbf{R}^m} .$$

Si $\mathrm{Ker}(B) = \{0\}$, le problème aux moindres carrés peut se ramener à la minimisation

d'une fonctionnelle quadratique sur $\mathbf{R}^n$ car

$$\frac{1}{2}J(x) = \frac{1}{2} < Ax, x >_{\mathbf{R}^n} - < b, x >_{\mathbf{R}^n} + K$$

en notant $A = {}^tBB \in SDP_n(\mathbf{R})$, $b = {}^tBc$ et $K = \dfrac{1}{2} < c, c >_{\mathbf{R}^m}$.

Remarque : On a montré au cours de la leçon A.III. que pour toute famille de $(m+1)$ points de $\mathbf{R}^2$ notés (x_i, y_i) $(i \in \{0, ..., m\})$ dont les abscisses sont deux à deux distinctes, il existait pour tout $n \leq m$ un unique polynôme $P \in \mathbf{P}_n$ tel que

$$\sum_{i=0}^{m} |y_i - P(x_i)|^2 = \inf_{Q \in \mathbf{P}_n} (\sum_{i=0}^{m} |y_i - Q(x_i)|^2).$$

Il s'agit en fait d'un problème aux moindres carrés sans contrainte sur $\mathbf{R}^{n+1}$ (identifié à $\mathbf{P}_n$) avec

$$B = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_1^2 & \dots & x_m^n \end{pmatrix} \quad \text{et} \quad c = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}.$$

1.5.5 Résolution d'un problème aux moindres carrés

Théorème 19. *Une solution $u \in \mathbf{R}^n$ d'un problème aux moindres carrés sans contrainte est exactement une solution du système linéaire $n \times n$ (appelé équation normale) :*

$${}^tBBu = {}^tBc.$$

Une telle solution existe toujours. Elle est unique si et seulement si $\mathrm{Ker}(B) = \{0\}$.

Démonstration. : On remarque par un calcul simple que la propriété

$$\forall x \in \mathbf{R}^n, \quad ||c - Bu||^2_{\mathbf{R}^m} \leq ||c - Bx||^2_{\mathbf{R}^m}$$

est équivalente à

$$\forall z \in \mathbf{R}^n, \quad < c - Bu, Bz >_{\mathbf{R}^m} = 0$$

(en prenant $x = u + tz$) soit exactement :

$${}^tBc - {}^tBBu = 0.$$

Cette équation a une solution si en particulier

$$\mathrm{Im}({}^tB) \subset \mathrm{Im}({}^tBB).$$

Pour démontrer l'inclusion (et même l'égalité) de ces deux ensembles, on remarque tout d'abord que

$$\mathrm{Ker}(B) = \mathrm{Ker}({}^tBB)$$

(une inclusion est claire et ${}^tBBx = 0$ implique ${}^tx{}^tBBx = {}^t(Bx)Bx = ||Bx||^2 = 0$).

Comme simultanément on a toujours $\mathrm{Im}({}^tBB) \subset \mathrm{Im}({}^tB) = \mathrm{Ker}(B)^\perp$ (orthogonal de $\mathrm{Ker}(B)$), on conclut pour des raisons de dimension que $\mathrm{Im}({}^tBB) = \mathrm{Im}({}^tB)$. Enfin, l'existence et l'unicité d'une solution si $\mathrm{Ker}(B) = \{0\}$ a déjà été observée □

Remarque : On peut aussi utiliser le théorème de projection de Riesz pour démontrer le Théorème 19 : en effet, le problème aux moindres carrés revient exactement à rechercher la projection de c sur $\mathrm{Im}(B) \subset R^m$. L'équation normale correspond alors à la caractérisation de celle-ci (noté Bu) et l'unicité d'une solution du problème aux moindres carrés est bien équivalente à l'injectivité de B.

1.5.6 Méthodes directes de minimisation

Toutes les méthodes directes définies dans la leçon A.VII. pour la résolution de systèmes linéaires du type $Au = b$ avec $A \in SDP_n(\mathbf{R})$ peuvent être réutilisées ici pour déterminer l'unique minimum d'une fonctionnelle quadratique sur $\mathbf{R}^n$.

À noter que dans le cas d'un problème aux moindres carrés, il est intéressant de pouvoir se dispenser du calcul de $A = {}^tBB$. En effet, celui-ci est non seulement coûteux (surtout si m est très supérieur à n) mais aussi dangereux numériquement comme le montre clairement l'exemple (issu de [LaTh1]) de la matrice

$$B = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix}$$

pour laquelle

$$A = {}^tBB = \begin{pmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1+\varepsilon^2 \end{pmatrix}.$$

Plutôt qu'une factorisation de A, on peut utiliser la factorisation QR de B (obtenue par l'algorithme de Householder : voir leçon A.VII.) :

$$B = QR$$

où $Q \in O_m(\mathbf{R})$ et $R \in \mathscr{M}_{m,n}(\mathbf{R})$ s'écrit $R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$ avec $R_1 \in TS_n(\mathbf{R})$.

On observe en effet qu'on peut alors écrire

$$J(x) = ||c - Bx||^2_{\mathbf{R}^m} = ||{}^tQc - {}^tQBx||^2_{\mathbf{R}^m} = ||({}^tQc)_n - R_1x||^2_{\mathbf{R}^n} + ||({}^tQc)_{m-n}||^2_{\mathbf{R}^{m-n}}$$

où on a décomposé

$${}^tQc = (({}^tQc)_n, ({}^tQc)_{m-n}) \in \mathbf{R}^n \times \mathbf{R}^{m-n}.$$

L'unique solution du problème aux moindres carrés (si B est injective) est alors donnée par l'inversion d'un système triangulaire :

$$u = R_1^{-1}({}^tQc)_n.$$

1.5.7 Méthodes itératives. Algorithmes de gradient

Toutes les méthodes itératives définies dans la leçon A.VII. pour la résolution de systèmes linéaires du type $Au = b$ avec $A \in SDP_n(\mathbf{R})$ peuvent également être réutilisées ici. Parmi celles-ci, on rappelle la méthode du gradient consistant à construite la suite $(u_k)_{k\in\mathbf{N}}$ telle que

$$u_0 \in \mathbf{R}^n \quad \text{et} \quad u_{k+1} = u_k + \alpha(b - Au_k) = u_k - \alpha\nabla J(u_k) \quad (k \in \mathbf{N})$$

(avec $\alpha \in]0, \frac{2}{\rho(A)}[$).

Cette méthode peut être réinterprétée (et sa convergence à nouveau justifiée) comme une méthode de recherche du minimum de la fonctionnelle quadratique

$$J(x) = \frac{1}{2} < Ax, x > - < b, x >$$

par un procédé de descente suivant la direction du gradient en chaque point de la suite.

Il est alors possible de proposer plusieurs variantes (ou améliorations) à cette méthode. La première consiste à choisir un pas de descente α_k variable dans $]0, \frac{2}{\rho(A)}[$ et la deuxième à rendre celui-ci optimal à chaque étape :

$$J(x-\rho_k \nabla J(x)) = \min_{\alpha \in \mathbf{R}} J(x-\alpha \nabla J(x)).$$

On peut montrer que dans le cas d'une fonctionnelle quadratique

$$\rho_k = \frac{||r_k||^2}{< Ar_k, r_k >}$$

avec $r_k = b_k - Au_k$.

1.5.8 Algorithme du gradient conjugué

L'algorithme du gradient conjugué reprend le principe des algorithmes précédents de gradient mais détermine une meilleure direction de descente en un point que celle du gradient, rendant cette méthode convergente en un nombre fini d'itérations.

1.5.8.1 Théorie

Théorème 20. *Soit $A \in SDP_n(\mathbf{R})$ et $x_0 \in \mathbf{R}^n$. On note $r_0 = b - Ax_0$ et pour tout $k \in \mathbf{N}$,*

$$K_k(r_0) = \text{Vect}(r_0, ..., A^k(r_0))$$

(espaces de Krilov associés à r_0).
Il existe une unique suite $(x_k)_{k \in \mathbf{N}}$ définie au choix par l'une des deux propriétés : (i) $r_{k+1} = b - Ax_{k+1}$ est orthogonal à $K_k(r_0)$.

(ii) x_{k+1} minimise la fonctionnelle $J(x) = \frac{1}{2} < Ax, x > - < b, x >$ dans l'espace affine $x_0 + K_k(r_0)$. De plus, la suite converge en au plus n itérations vers u solution de $Au = b$.

Démonstration. : On montre tout d'abord que les propriétés (i) et (ii) sont équivalentes et définissent bien un unique vecteur x_{k+1}. Pour cela, on définit pour tout $y \in K_k(r_0)$ la fonction g par

$$g(y) = J(x_0 + y) = \frac{1}{2} < \tilde{A}y, y > - < r_0, y > + \frac{1}{2} < Ax_0, x_0 > - < b, x_0 >$$

où $\tilde{A}$ désigne la restriction de A à $K_k(r_0)$. $\tilde{A}$ est encore une matrice symétrique et définie positive. En vertu du Théorème 18, g admet donc un unique minimum dans $K_k(r_0)$ pouvant être noté $x_{k+1} - x_0$. Ainsi,

$$\forall y \in K_k(r_0), \quad g(x_{k+1} - x_0 + y) \geq g(x_{k+1} - x_0)$$

ce qui donne

$$\forall y \in K_k(r_0), \quad \frac{1}{2} < Ay, y > + < Ax_{k+1} - b, y > \geq 0.$$

Remplaçant y par ty ($t \in \mathbf{R}$), on trouve aisément que

$$\forall y \in K_k(r_0), \quad < Ax_{k+1} - b, y > = 0$$

soit exactement la propriété (i).

Pour démontrer la convergence de la suite $(x_k)_{k\in\mathbf{N}}$, on remarque que la famille des sous-espaces vectoriels $(K_k(r_0))_{k\in\mathbf{N}}$ forme une suite strictement croissante pour l'inclusion jusqu'à un certain rang $k_0 \leq n-1$ (appelé indice de Krylov de r_0) au delà duquel elle stationne.

Si $k_0 = n-1$, alors $K_{k_0}(r_0) = \mathbf{R}^n$ et la propriété (ii) indique bien que $x_{k_0+1} = u$, solution de $Au = b$.

Si $k_0 < n-1$, on décompose $A^{k_0+1}(r_0)$ dans $K_{k_0}(r_0)$:

$$A^{k_0+1}(r_0) = \sum_{i=0}^{k_0} \alpha_i A^i(r_0)$$

Nécessairement, $\alpha_0 \neq 0$ (car sinon $K_{k_0-1}(r_0) = K_{k_0}(r_0)$) et alors

$$A\left(\frac{1}{\alpha_0} A^{k_0}(r_0) - \sum_{i=1}^{k_0} \frac{\alpha_i}{\alpha_0} A^{i-1}(r_0) + x_0\right) = b$$

et la solution u de $Au = b$ appartient à l'espace affine $x_0 + K_{k_0}(r_0)$, ce qui implique également que $x_{k_0+1} = u$ grâce à la propriété (ii) □

1.5.8.2 Construction de l'algorithme

Avant de construire un algorithme plus efficace de calcul de la suite des itérations du gradient conjugué, on montre la proposition suivante :

Proposition 1. *Les notations du Théorème précédent sont conservées. On a*

$$\forall k \in \mathbf{N}, \quad K_k(r_0) = \text{Vect}(r_0, ..., r_k) = \text{Vect}(d_0, ..., d_k)$$

où on a noté $d_k = x_{k+1} - x_k$. De plus les familles $(r_i)_{i\in\mathbf{N}}$ et $(d_i)_{i\in\mathbf{N}}$ sont orthogonales pour le produit scalaire $< ., . >$, respectivement $< A., . >$.

Démonstration. : Par construction, r_{k+1} est orthogonal à $K_k(r_0)$ ($k \in \mathbf{N}$) et

$$r_{k+1} = b - Ax_{k+1} = b - A(x_0 + \underbrace{x_{k+1} - x_0}_{\in K_k(r_0)}) \in K_{k+1}(r_0).$$

D'autre part, $d_k = (x_{k+1} - x_0) - (x_k - x_0) \in K_k(r_0)$ et si $0 \leq l < k$,

$$< Ad_k, d_l > = < Ax_{k+1} - Ax_k, d_l > = < r_k - r_{k+1}, d_l > = 0 - 0 = 0$$

car r_k et r_{k+1} sont orthogonaux à $K_l(r_0)$ □

Remarque : On dit que la famille $(d_i)_{i\in\mathbf{N}}$ est conjuguée par rapport à A. Cette famille correspondant aux directions successives de descente de la suite du gradient conjugué, le nom donné à la méthode est ainsi justifié.

Théorème 21. *Soit $A \in SDP_n(\mathbf{R})$ et $x_0 \in \mathbf{R}^n$. On définit les suites $(x_k)_{0\leq k\leq k_0+1}$ et $(r_k)_{0\leq k\leq k_0+1}$ comme dans le Théorème 20. Il existe une suite $(p_i)_{0\leq i\leq k_0}$ conjuguée par rapport à A telle que*

$$p_0 = r_0 = b - Ax_0 \quad \text{et} \quad \begin{cases} r_k = r_{k-1} - \alpha_{k-1} A p_{k-1}, \\ p_k = r_k + \beta_{k-1} p_{k-1}, \\ x_{k+1} = x_k + \alpha_k p_k \end{cases} \qquad (1 \leq k \leq k_0), \qquad (1)$$

avec $\alpha_{k-1} = \dfrac{< r_{k-1}, r_{k-1} >}{< Ap_{k-1}, p_{k-1} >}$ *et* $\beta_{k-1} = \dfrac{< r_k, r_k >}{< r_{k-1}, r_{k-1} >}$. *Réciproquement, il est possible de définir les trois suite (x_k, r_k, p_k) par la relation (1) jusquà l'indice k_0. La suite $(x_k)_{0\leq k\leq k_0+1}$ correspond alors à la suite des itérées du gradient conjugué.*

Démonstration. :

On va identifier la famille $(p_k)_{0\leq k\leq k_0}$ à une certaine famille orthogonalisée de Gram Schmidt de la famille libre $(r_k)_{0\leq k\leq k_0}$ pour le produit scalaire $< A., . >$. Soit donc la famille $(p'_k)_{0\leq k\leq k_0}$ telle que

$$p_0 = r_0 \quad \text{et} \quad p'_k = r_k + \sum_{j=0}^{k-1} \beta_{j,k} p'_j \quad (1 \leq k \leq k_0)$$

avec $\beta_{j,k} = -\dfrac{< Ar_k, p'_j >}{< Ap'_j, p'_j >}$. On montre que $p'_k = p_k$. Par la Proposition précédente, la famille $(d_k)_{0\leq k\leq k_0}$ est aussi une famille orthogonalisée de Gram Schmidt de la famille $(r_k)_{0\leq k\leq k_0}$ pour le produit scalaire $< A., . >$. Par unicité, on a donc

$$\forall k \in \{0, ..., k_0\}, \quad \exists \alpha_k \in \mathbf{R} \ / \ d_k = x_{k+1} - x_k = \alpha'_k p'_k$$

ce qui implique aussitôt :

$$\forall k \in \{0, ..., k_0\}, \quad r_{k+1} - r_k = -\alpha'_k A p'_k$$

En particulier, pour tout $k \in \{1, ..., k_0\}$ et $j \in \{0, ..., k-1\}$:

$$< Ar_k, p'_j > = < r_k, Ap'_j > = \frac{1}{\alpha'_j} < r_k, r_j - r_{j+1} >$$

et

$$\beta_{j,k} = \begin{cases} 0 \quad \text{si} \quad 0 \leq j \leq k-2, \\ \dfrac{< r_k, r_k >}{\alpha'_{k-1} < Ap'_{k-1}, p'_{k-1} >} \quad \text{si} \quad j = k-1. \end{cases}$$

Pour établir que $p'_k = p_k$, il reste simplement à obtenir les expressions de α'_k et $\beta_{k-1,k}$ données dans l'énoncé du théorème. Pour cela, on écrit que r_{k+1} est orthogonal à r_k :

$$\begin{aligned} 0 = < r_{k+1}, r_k > &= < r_k, r_k > - \alpha'_k < Ap_k, r_k > \\ &= < r_k, r_k > - \alpha'_k < Ap'_k, p'_k - \beta_{k-1,k} p'_{k-1} > \\ &= < r_k, r_k > - \alpha'_k < Ap'_k, p'_k > \end{aligned}$$

et

$$\beta_{k-1,k} = \frac{< r_k, r_k >}{\alpha'_{k-1} < Ap'_{k-1}, p'_{k-1} >} = \frac{< r_k, r_k >}{< r_{k-1}, r_{k-1} >} = \beta_{k-1}.$$

ce qui clôt la démonstration de la partie directe. Réciproquement, en définissant les suites (x_k, r_k, p_k) par la relation (1) jusqu'à un indice $k'_0 \in \mathbf{N}^*$, on montre facilement par récurrence que $r_k = b - Ax_k$. De la même façon, p_k et r_k appartiennent à $K_k(r_0)$ et la famille $(r_k)_{0 \leq k \leq k'_0+1}$ est orthogonale pour le produit scalaire usuel.

D'après la première définition du Théorème précédent, ces propriétés suffisent pour conclure que k'_0 est l'indice de Krylov de r_0 et que la suite $(x_k)_{0 \leq k \leq k_0+1}$ correspond à la suite des itérées du gradient conjugué □

Remarque : Un intérêt de la méthode du gradient conjugué est de ne nécessiter que le stockage de Au pour tout $u \in \mathbf{R}^n$ et non de la matrice A entière, ce qui s'avère économique lorsque A est une matrice creuse.

1.5.8.3 convergence et conditionnement

La méthode du gradient conjugué est construite dans la pratique à partir de la relation (1). Lorsqu'elle est utilisée comme méthode directe, il est nécessaire dans le cas le plus défavorable d'effectuer environ $2n^3$ opérations avant d'obtenir la solution exacte, soit plus qu'avec les méthodes de Gauss ($\frac{2n^3}{3}$) ou de Cholesky ($\frac{n^3}{3}$).

En fait, elle est davantage utilisée comme méthode itérative (les erreurs numériques empêchent de toute manière d'obtenir exactement $r_{k_0+1} = 0$). Par exemple, un test est effectué sur $\frac{<r_k,r_k>}{<r_0,r_0>}$ pour décider d'arrêter le calcul avant le rang k_0+1. Il existe en effet une estimation de l'erreur commise (dont la démonstration est assez technique : voir [LaTh2]) indiquant que :

$$||x_k - u||_2 \leq 2\sqrt{\text{cond}_2 A}\left(\frac{\sqrt{\text{cond}_2 A}-1}{\sqrt{\text{cond}_2 A}+1}\right)^k ||x_0 - u||_2$$

où $\text{cond}_2 A$ désigne le conditionnement de A suivant la norme euclidienne (voir leçon A.VII.). On s'aperçoit en particulier qu'un préconditionnement de A peut accélérer la convergence (voir [LuPi] et [LaTh2] pour des méthodes de gradient conjugué préconditionné).