Résolution des systèmes linéaires $Ax = b$. Méthodes itératives.

Le 4 novembre 2014

Jean Roux

Mini-cours donné en 3^e année de
Licence de Sciences du département de Géosciences
École normale supérieure, Paris

Résolution des systèmes linéaires $Ax = b$ par des méthodes numériques itératives.

Nous avons déjà signalé que le domaine d'application privilégié de ces méthodes est, essentiellement, celui de la résolution des systèmes linéaires à matrice creuse ou, au moins, possédant une structure (tridiagonale, pentadiagonale, structure bande, tridiagonale par blocs, etc.). Généralement ce type de matrice est obtenu par discrétisation des équations aux dérivées partielles qui fournit usuellement les trois types de matrices suivantes :

- symétriques définies positives,
- strictement diagonalement dominantes,
- irréductibles (voir livre pour cette notion, dans ce mini-cours on fait toujours référence à [1]) et diagonalement dominantes.

1.1 Généralités

Construction et convergence des méthodes itératives Il s'agit toujours de résoudre

$$Ax = b \tag{1.1.1}$$

où $A \in \mathbb{C}^{n,n}$ est inversible et $b \in \mathbb{C}^n$. Résoudre (1.1.1) par une méthode itérative, c'est construire une suite de vecteurs $x^{(m)}$ telle que $x^{(m)} \xrightarrow[m\to\infty]{} x$ où x est solution de (1.1.1).

Dans tout ce cours, la suite $x^{(m)}$ est définie, de façon générale, par le schéma itératif

$$x^{(m+1)} = Bx^{(m)} + c, \tag{1.1.2}$$

où la matrice B est construite à partir de la matrice A, la matrice B et le vecteur c sont *indépendants* de l'indice m de l'itération. La matrice B est appelée *matrice d'itération*. Pour que la limite de la suite soit égale à x il faut que

$$c = (I - B)x = (I - B)A^{-1}b \tag{1.1.3}$$

cette condition doit être vérifiée pour toute construction de la matrice B.

Définition 1.1.1. *La méthode itérative (1.1.2) est dite convergente si et seulement si, pour tout $x^{(0)} \in \mathbb{C}^n$, $x^{(m)} \to x$ lorsque $m \to \infty$.*

Donnons immédiatement des critères de convergence pour une méthode itérative. Ces critères s'appuient sur le théorème ci-après (voir livre) :

[1] M. Ghil, J. Roux Mathématiques appliquées aux Sciences de la Vie et de la Planète, Dunod, Paris, 2010.

Théorème 1.1.1. *Soit A une matrice carrée donnée, les quatre conditions suivantes sont équivalentes :*

- *(a)* $A^k \xrightarrow[k\to\infty]{} 0$,
- *(b)* $A^k x \xrightarrow[k\to\infty]{} 0$, *ceci pour tout* x,
- *(c)* $\rho(A) < 1$, *(on rappelle que* $\rho(A)$ *est le rayon spectral de* A*),*
- *(d) Il existe (au moins) une norme matricielle* $\|.\|$ *telle que* $\|A\| < 1$.

Alors, pour une méthode itérative, on en déduit les critères de convergence suivants :

Théorème 1.1.2. *Les trois propositions suivantes sont équivalentes*

- *(1) La méthode est convergente.*
- *(2)* $\rho(B) < 1$.
- *(3) Il existe une norme matricielle* $\|.\|$ *telle que* $\|B\| < 1$.

Preuve : Soit $x = A^{-1}b$, on pose $\epsilon^{(m)} = x^{(m)} - x$, $m = 0, 1, \cdots$. On remarque, en passant à la limite dans (1.1.2), que x est solution de

$$x = Bx + c.$$

Par soustraction, d'après (1.1.2), il vient donc, pour $m = 0, 1, \cdots$

$$\epsilon^{(m+1)} = B\epsilon^{(m)},$$

soit

$$\epsilon^{(m)} = B^m \epsilon^{(0)} \tag{1.1.4}$$

Pour que $\epsilon^{(m)} \to 0$ lorsque $m \to \infty$, quel que soit $\epsilon^{(0)} \in \mathbb{C}^n$, il faut et il suffit que $\lim_{m\to\infty} B^m = 0$. Les équivalences du Théorème 1.1.2 se déduisent alors du Théorème 1.1.1. □

La question est maintenant de construire simplement la matrice d'itération B et le vecteur c associé donné par (1.1.3).

Nous verrons que toutes les méthodes qui seront exhibées entrent dans le cadre de la définition suivante

Définition 1.1.2. *Soit $A \in \mathbb{C}^{n,n}$, on appelle décomposition de A toute manière d'écrire A sous la forme $A = M - N$, $M \in \mathbb{C}^{n,n}$ et $N \in \mathbb{C}^{n,n}$, où la matrice M est* régulière.

À une décomposition choisie $M-N$ de la matrice A on associe la méthode itérative

$$Mx^{(m+1)} = Nx^{(m)} + b \tag{1.1.5}$$

qui est bien de la forme (1.1.2), (1.1.3) avec $B = M^{-1}N$ et

$$c = M^{-1}b = M^{-1}AA^{-1}b = M^{-1}(M-N)A^{-1}b = (I-B)A^{-1}b.$$

La formule (1.1.5) fournit donc une méthode qui convient.

L'étude des méthodes itératives se ramène à l'étude des deux problèmes suivants

- 1. Étant donné une méthode itérative de matrice d'itération B, déterminer si la méthode est convergente, c'est-à-dire examiner si $\rho(B) < 1$ ou, de façon équivalente, exhiber une norme matricielle $\|.\|$ telle que $\|B\| < 1$.
- 2. Étant donné deux méthodes itératives convergentes, de matrices d'itération B_1 et B_2, les comparer. On admet (voir livre pour les justifications) que la méthode la plus rapide sera celle dont la matrice d'itération aura le plus petit rayon spectral.

1.2 Les méthodes itératives de Jacobi, Gauss-Seidel et de relaxation successive

Il y a deux types de méthodes itératives, les méthodes ponctuelles et les méthodes par blocs inspirées des premières.

Après avoir introduit les deux méthodes de Jacobi et de Gauss-Seidel, on généralisera Gauss-Seidel pour obtenir les méthodes de relaxation.

Ces méthodes itératives ont en commun que chaque itération nécessite un nombre d'op.él. du même ordre de grandeur que celui nécessaire au produit d'un vecteur par une matrice (i.e. environ n^2 op.él. par itération). Il faut ici insister sur le fait que n peut être très grand, quelquefois de l'ordre du million (ou plus), et on a grand intérêt à avoir des méthodes qui convergent vite, ou au moins à utiliser la méthode la plus rapide possible.

Les méthodes ponctuelles Soit $A = (a_{ij})$ une matrice complexe d'ordre n. On pose

$$D = \mathrm{diag}(a_{ii}), E = (e_{ij}) \text{ où } e_{ij} = \begin{cases} 0 & \text{si} \quad j \geq i \\ -a_{ij} & \text{si} \quad j < i \end{cases}, \tag{1.2.1}$$

$$F = (f_{ij}) \text{ où } f_{ij} = \begin{cases} -a_{ij} & \text{si} \quad j > i \\ 0 & \text{si} \quad j \leq i \end{cases}. \tag{1.2.2}$$

ce que l'on peut schématiser ainsi (Figure 1.1).

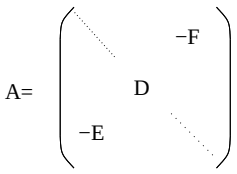

Figure 1.1:

On a donc

$$A = D - E - F.$$

Hypothèse fondamentale On a supposé A inversible. On suppose de plus que $a_{ii} \neq 0$, $1 \leq i \leq n$; de sorte que D est inversible et $D^{-1} = \text{diag}(1/a_{ii})$. □

Dans le cadre de la discrétisation des équations aux dérivées partielles (EDP) par différences finies ou éléments finis, l'hypothèse ci-dessus est vérifiée dans les bons cas où on peut exhiber des théorèmes d'existence et d'unicité de la solution de l'EDP. De ce point de vue ce n'est pas une hypothèse restrictive.

Nous allons considérer trois exemples de décomposition de A faisant intervenir les matrices D, E et F dans la définition des matrices M et N intervenant dans la méthode itérative générale (1.1.5).

La méthode de Jacobi

On utilise la décomposition $M = D$ et $N = E + F = D - A$. On a bien $A = M - N$ avec M inversible. La matrice d'itération associée est alors

$$B = M^{-1}N = D^{-1}(E + F) = D^{-1}(D - A) = I - D^{-1}A,$$

et

$$c = M^{-1}b = D^{-1}b = (D^{-1}A)(A^{-1}b) = (I - B)A^{-1}b;$$

on retrouve la condition nécessaire (1.1.3) sur le vecteur c. Dans toute la suite on notera par B la matrice d'itération de la méthode de Jacobi.

La méthode (1.1.5) est alors appelée la méthode itérative de *Jacobi*, elle s'écrit

$$Dx^{(m+1)} = (E + F)x^{(m)} + b, \quad x^{(0)} \in \mathbb{C}^n \quad \text{donné ;} \tag{1.2.3}$$

ou encore

$$a_{ii}x_i^{(m+1)} = - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}x_j^{(m)} + b_i, \quad 1 \leq i \leq n, \tag{1.2.4}$$

soit, puisque, par hypothèse, $a_{ii} \neq 0$, $1 \leq i \leq n$:

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \left(- \sum_{j=1, j \neq i}^{n} a_{ij}x_j^{(m)} + b_i \right), \quad 1 \leq i \leq n. \tag{1.2.5}$$

La méthode utilise $2n$ mémoires pour stocker toutes les composantes des deux vecteurs itérés successifs $x^{(m)}$ et $x^{(m+1)}$.

La méthode de Gauss-Seidel

Intuitivement, il semble que la convergence de la méthode précédente sera améliorée si pour calculer $x_i^{(m+1)}$ on utilise les $(i-1)$ premières composantes de $x^{(m+1)}$ déjà calculées (ce que l'on appelle les *composantes actualisées*) au lieu d'utiliser les $(i-1)$ premières composantes de $x^{(m)}$. On montre (voir livre pour plus d'informations) que, pour une large classe de matrices, cette amélioration de la convergence est théoriquement justifiée. Par exemple (théorème 1.3.6), lorsque la matrice A est tridiagonale par blocs et définie positive, les méthodes par blocs de Jacobi et de Gauss-Seidel convergent simultanèment et $\rho(\mathcal{L}_1) = (\rho(B))^2$. La méthode de Gauss-Seidel converge plus vite que Jacobi (en un certain sens deux fois plus vite (voir livre)).

Les relations (1.2.4) sont donc modifiées comme suit

$$a_{ii}x_i^{(m+1)} = -\sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(m)} + b_i, \quad 1 \leq i \leq n, \qquad (1.2.6)$$

en faisant la convention $\sum_p^q = 0$ si $p > q$.

Sous forme matricielle cela s'écrit

$$(D-E)x^{(m+1)} = Fx^{(m)} + b, \qquad (1.2.7)$$

ce qui correspond à la décomposition $M = D - E$ et $N = F$ (on a bien $A = D - E - F = M - N$). La matrice M est bien inversible car $(D-E)$ est inversible puisqu'elle est triangulaire inférieure avec des éléments diagonaux $a_{ii} \neq 0$ (D est supposée inversible).

C'est la méthode ponctuelle dite de *Gauss-Seidel* dite aussi des *déplacements successifs*.

La matrice d'itération $\mathcal{L}_1 = (D-E)^{-1}F$ est appelée la matrice de Gauss-Seidel associée à A. En posant $L = D^{-1}E$ et $U = D^{-1}F$ on a

$$\mathcal{L}_1 = (I-L)^{-1}U. \qquad (1.2.8)$$

N.B. : On verra ci-après la raison de la notation $\mathcal{L}_1$. □

Remarque 1.2.1. *Cette méthode n'exige que n mémoires pour conserver les composantes des deux vecteurs itérés successifs, la i^{ieme} composante de $x^{(m+1)}$ venant "écraser", dès qu'elle est calculée, la i^{ieme} composante de $x^{(m)}$ devenue inutile.*

Remarque 1.2.2. *L'examen des formules (1.2.4) et (1.2.6), définissant les méthodes itératives de Jacobi et de Gauss-Seidel, montre que ces méthodes sont très faciles à programmer.*

La méthode de relaxation successive

Supposons qu'à partir du vecteur $x^{(m)}$ on ait calculé, par une méthode qui reste à définir, les $(i-1)$ premières composantes de $x^{(m+1)}$. On calcule alors le nombre $\tilde{x}_i^{(m+1)}$ comme par la méthode de Gauss-Seidel, c'est-à-dire que

$$a_{ii}\tilde{x}_i^{(m+1)} = -\sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(m)} + b_i, \quad 1 \leq i \leq n, \tag{1.2.9}$$

la i^{ieme} composante de $x^{(m+1)}$ est alors donnée par

$$x_i^{(m+1)} = x_i^{(m)} + \omega(\tilde{x}_i^{(m+1)} - x_i^{(m)}) \tag{1.2.10}$$

où $\omega \in \mathbb{R}$, $\omega \neq 0$. Le nombre $x_i^{(m+1)}$ est donc une moyenne pondérée de l'ancienne composante $x_i^{(m)}$ et du nombre auxiliaire $\tilde{x}_i^{(m+1)}$. Si $\omega = 1$, $x_i^{(m+1)} = \tilde{x}_i^{(m+1)}$ avec $\tilde{x}_i^{(m+1)}$ calculé, à partir des composantes de $x^{(m)}$, par Gauss-Seidel. L'idée est donc de faire mieux que Gauss-Seidel, sachant que l'on retrouve cette méthode pour $\omega = 1$. On espère trouver un ω appartenant à un intervalle de confiance, encadrant la valeur 1, où la méthode converge.

En combinant les formules (1.2.9) et (1.2.10) on obtient, en multipliant (1.2.10) par a_{ii} (par hypothèse $a_{ii} \neq 0$), pour $i = 1, 2, \cdots$

$$a_{ii}x_i^{(m+1)} = a_{ii}x_i^{(m)} + \omega \left\{ -\sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(m)} + b_i - a_{ii}x_i^{(m)} \right\}, \tag{1.2.11}$$

relation qui donne directement $x_i^{(m+1)}$ en fonction des $(i-1)$ premières composantes de $x^{(m+1)}$ déjà calculées, et des $(n-i+1)$ dernières composantes de $x^{(m)}$.

Remarque 1.2.3. *La phrase introductive à cette méthode est maintenant éclaircie !*

Remarque 1.2.4. *On fait la même observation que pour Gauss-Seidel pour l'encombrement mémoire.*

En notation matricielle on exprime (1.2.11) par

$$(D - \omega E)x^{(m+1)} = \left((1-\omega)D + \omega F\right)x^{(m)} + \omega b, \tag{1.2.12}$$

la matrice $(D - \omega E)$ est inversible pour tout ω car $a_{ii} \neq 0$ et (1.2.12) correspond à la décomposition

$$M = \frac{1}{\omega}(D - \omega E) \text{ et } N = \frac{1}{\omega}\left((1-\omega)D + \omega F\right). \tag{1.2.13}$$

On vérifie que

$$M - N = \frac{D}{\omega} - E - \frac{1-\omega}{\omega} D - F = D - E - F = A.$$

On a ainsi défini la méthode de relaxation successive associée à A. On a

$$\mathcal{L}_\omega = M^{-1}N = (D - \omega E)^{-1} \left((1-\omega)D + \omega F \right), \tag{1.2.14}$$

en posant $L = D^{-1}E$ et $U = D^{-1}F$, il vient

$$\mathcal{L}_\omega = (I - \omega L)^{-1} \left((1-\omega)I + \omega U \right). \tag{1.2.15}$$

Le paramètre ω est le paramètre de relaxation. Si $\omega > 1$ (resp. $\omega < 1$) on dit qu'il y a sur-relaxation (resp. sous-relaxation). Si $\omega = 1$ on a déjà fait remarquer que l'on retrouve la méthode de Gauss-Seidel, ce qui justifie la notation $\mathcal{L}_1$ utilisée à son sujet.

Pour cette méthode il s'agit de trouver

- Un intervalle de confiance $[\omega_m, \omega_M]$ (contenant le nombre 1) tel que $\rho(\mathcal{L}_\omega) < 1$ pour $\omega_m < \omega < \omega_M$.
- Dans cet intervalle, s'il est non vide, un paramètre optimal ω_{opt}, s'il existe, tel que

  $$\rho(\mathcal{L}_{\omega_{opt}}) = \inf \left\{ \rho(\mathcal{L}_\omega); \quad \omega_m < \omega < \omega_M \right\},$$

  en espérant que $\rho(\mathcal{L}_{\omega_{opt}}) < \rho(\mathcal{L}_1)$.

On prouve (voir livre)

Corollaire 1.2.1. *Pour toute matrice A, une condition nécessaire de convergence de la méthode de relaxation successive est que $0 < \omega < 2$.*

Les méthodes par blocs Il est naturel, étant donné la structure usuelle des matrices de discrétisation obtenues par différences finies ou par éléments finis, d'envisager des méthodes par blocs. On considère donc une partition de l'ensemble $\{1, 2, \ldots, n\}$ en s parties par

$$\{1, \ldots, n_1\}, \{n_1+1, \ldots, n_1+n_2\}, \ldots, \{n_1+\cdots+n_{s-1}+1, \ldots, n_1+n_2+\cdots+n_s\}.$$

La partition correspondante en blocs de la matrice A (voir livre pour plus de détails) se présente sous la forme

$$A = \begin{vmatrix} A_{11} & A_{12} & \cdots & A_{1s} \\ A_{21} & A_{22} & \cdots & A_{2s} \\ \vdots & \vdots & & \vdots \\ A_{s1} & A_{s2} & \cdots & A_{ss} \end{vmatrix},$$

où $A_{ij} \in \mathbb{C}^{n_i, n_j}$, pour $1 \leq i,\ j \leq s$.

Les blocs diagonaux sont nécessairement des matrices carrées $A_{ii} \in \mathbb{C}^{n_i, n_i}$ pour $1 \leq i \leq s$ et on peut parler de leur inversibilité. On définit

- la matrice $D \in \mathbb{C}^{n,n}$, matrice bloc-diagonale, $D = \mathrm{diag}(A_{ii})$,
- la matrice $E \in \mathbb{C}^{n,n}$, matrice bloc-triangulaire inférieure, telle que

$$-E = \begin{vmatrix} 0 & & 0 & 0 \\ A_{2,1} & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \\ A_{s,1} & \cdots & A_{s,s-1} & 0 \end{vmatrix},$$

- la matrice $F \in \mathbb{C}^{n,n}$, matrice bloc-triangulaire supérieure, telle que

$$-F = \begin{vmatrix} 0 & A_{1,2} & \cdots & A_{1,s} \\ 0 & \ddots & & \\ \vdots & \ddots & \ddots & A_{s-1,s} \\ 0 & & 0 & 0 \end{vmatrix}.$$

On a bien entendu $A = D - E - F$.

Hypothèse fondamentale : On suppose que la matrice bloc D est régulière, ce qui est équivalent à supposer que toutes les matrices blocs A_{ii}, $1 \leq i \leq s$, sont régulières. □

On définit les méthodes de Jacobi, Gauss-Seidel et de relaxation successive par blocs comme précédemment, les matrices D, E et F n'étant pas, évidemment, les mêmes (elles ne sont identiques que si l'on prend la partition $\{1, 2, \cdots, n\} = \{1\} \cup \{2\} \cup \cdots \cup \{n\}$!).

Par exemple la méthode de relaxation successive par blocs s'écrit

$$\begin{aligned} A_{ii}X_i^{(m+1)} = A_{ii}X_i^{(m)} + \omega \Big\{ &- \textstyle\sum_{j=1}^{i-1} A_{ij}X_j^{(m+1)} - \sum_{j=i+1}^{s} A_{ij}X_j^{(m)} \\ &+ B_i - A_{ii}X_i^{(m)} \Big\}, \quad 1 \leq i \leq s, \end{aligned} \tag{1.2.16}$$

si, avec la même partition, on pose

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_s \end{pmatrix}, \text{ et } B = \begin{pmatrix} B_1 \\ \vdots \\ B_s \end{pmatrix}. \tag{1.2.17}$$

Par l'hypothèse faite, les matrices A_{ii} sont inversibles et la résolution de (1.2.16) est possible : on peut calculer le bloc d'inconnues $X_i^{(m+1)}$ à l'itération $(m+1)$.

Intérêt des méthodes par blocs Prenons le cas de la méthode (1.2.16). À chaque itération nous avons s systèmes linéaires à résoudre, cela paraît être un sérieux handicap ; mais, notons déjà que si chacun de ces systèmes est

très simple à résoudre on peut raisonnablement travailler avec ces méthodes. Cela n'a cependant d'intérêt que si la méthode par blocs est plus rapide que la méthode ponctuelle. Or, sous des hypothèses raisonnables, pour un même ω, la méthode par blocs est plus rapide que la méthode ponctuelle, on a

$$\rho(\mathcal{L}_\omega^B) < \rho(\mathcal{L}_\omega^p) < 1,$$

où on désigne par $\mathcal{L}_\omega^B$ (resp. $\mathcal{L}_\omega^p$) la matrice d'itération de la méthode de relaxation successive par blocs (resp. ponctuelle).

En pratique, souvent les matrices A_{ii} sont des matrices-bandes (par exemple tridiagonales). Les méthodes directes (Gauss,...) sont alors très bien adaptées, car très rapides, à la résolution de (1.2.16). Les formules se simplifient notablement du fait de la structure des A_{ii} ; par exemple, dans le cas des A_{ii} tridiagonales on dispose de formules algèbriques très simples donnant la solution (voir livre).

Dans ce cadre, même en tenant compte de la résolution des s systèmes linéaires à chaque étape, le gain sur la rapidité de convergence des méthodes par blocs compense très souvent cet inconvénient. Dans bien des cas on y gagne !

1.3 Quelques théorèmes de convergence

Généralement on ne sait rien dire de la convergence et de la comparaison des méthodes itératives lorsque la matrice A n'a pas de propriété(s) particulière(s), telle(s) que symétrique définie positive ou diagonalement dominante. On peut exhiber une matrice A_1 telle que la méthode de Jacobi converge alors que Gauss-Seidel diverge ; on peut trouver aussi une autre matrice A_2 telle que, inversement, Gauss-Seidel converge alors que Jacobi diverge.

Cependant il existe grossièrement deux grandes classes de résultats de convergence. L'une relative aux matrices symétriques (hermitiennes) définies positives, l'autre aux matrices strictement diagonalement dominantes. J'omets le cas, fondamental en pratique mais mathématiquement plus difficile, des matrices irréductibles diagonalement dominantes qui sort du cadre restreint de ce mini-cours.

On pose la définition suivante:

Définition 1.3.1. *Une matrice $A \in \mathcal{C}^{n,n}$ est dite à diagonale dominante si et seulement si*

$$|a_{i,i}| \geq \sum_{j=1, j\neq i}^{n} |a_{i,j}|, \qquad i = 1, 2, \cdots, n \tag{1.3.1}$$

avec inégalité stricte pour au moins un indice i.

Elle est dite à diagonale strictement dominante (SDD) s'il y a inégalité stricte dans (1.3.1) pour tout i.

Remarque 1.3.1. *Certains auteurs préfèrent dire fortement dominante au lieu de dominante. Réservant le nom de "dominante" au cas où il n'y a jamais d'inégalité stricte dans (1.3.1).*

On prouve (voir livre) qu'une matrice SDD est régulière. Notre hypothèse fondamentale est satisfaite.

Naturellement ces résultats sont distincts entre eux. Par exemple, une matrice (symétrique) peut-être strictement diagonalement dominante sans être définie positive et inversement. Soit, par exemple,

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix} \tag{1.3.2}$$

Pour $1/2 \leq a < 1$ la matrice A n'est pas SDD et pourtant la matrice A est (symétrique) définie positive. En effet les valeurs propres de A sont données par les racines du polynôme caractéristique $\det(A - \lambda I) = 0$ où la matrice $A - \lambda I$ s'écrit

$$A - \lambda I = \begin{pmatrix} 1-\lambda & a & a \\ a & 1-\lambda & a \\ a & a & 1-\lambda \end{pmatrix}. \tag{1.3.3}$$

On voit que $P(\lambda) = \det(A - \lambda I) = \mu^3 - 3a^2\mu + 2a^3$ où $\mu = 1 - \lambda$. Les racines sont $\mu = a$ (racine double) et $\mu = -2a$, donc les valeurs propres de A sont $\lambda = 1 - a$ (racine double) et $\lambda = 1 + 2a$. Si $1/2 \leq a < 1$ les valeurs propres de A sont toutes strictement positives et cela implique que A est définie positive.

Inversement, soit, par exemple, la matrice

$$A = \begin{pmatrix} -2 & a \\ a & 2 \end{pmatrix};$$

notons qu'il faut choisir, sachant que ses valeurs propres sont réelles puisque la matrice est symétrique, une matrice où au moins un des éléments diagonaux est négatif, de façon à assurer la caractère non positif. On vérifie que les valeurs propres sont données par $\lambda = \pm\sqrt{a^2 + 4}$, une des valeurs propres étant négative la matrice n'est jamais définie positive. Or la matrice est SDD pour $|a| < 2$.

Comme, en pratique, les deux classes de matrices précédentes sont fréquentes (lors de la discrétisation des équations aux dérivées partielles), il faut examiner les deux catégories de résultats.

Cas des matrices symétriques (hermitiennes) définies positives Nous allons d'abord examiner un théorème donnant un critère de convergence des méthodes itératives ponctuelles ou par blocs pour ce type de matrices. *A priori* une des hypothèses semble arbitraire, la suite éclaircira son bien-fondé.

Théorème 1.3.1. *Soit $A \in \mathbb{C}^{n,n}$ une matrice hermitienne et inversible. Soit $A = M - N$ une décomposition de A telle que la matrice $M^* + N$ soit définie positive. Alors $\rho(M^{-1}N) < 1$ si et seulement si A est définie positive.*

Preuve : Admise. □

Nous sommes maintenant en mesure de donner un premier critère de convergence des méthodes de relaxation successive (ponctuelle ou par blocs) incluant, nous le verrons, les méthodes de Gauss-Seidel (ponctuelle ou par blocs). Si A est hermitienne, on constate que $F = E^*$ dans les définitions précédentes donnant A sous la forme $A = D - E - F$.

Théorème 1.3.2. *Soit $A \in \mathbb{C}^{n,n}$ une matrice hermitienne, inversible, D et $E \in \mathbb{C}^{n,n}$ telles que*

$$A = D - E - E^*, \tag{1.3.4}$$

et soit $\omega \in \mathbb{R}$ tel que $0 < \omega < 2$. Supposons que la matrice D est définie positive. Alors $\rho(\mathcal{L}_\omega) < 1$ si et seulement si la matrice A est définie positive.

Preuve : On constate que si A s'écrit sous la forme (1.3.4) alors nécessairement D est hermitienne. Dans tous les cas la méthode de relaxation successive correspond à la forme $A = M - N$ avec (voir (1.2.13)) $M = (1/\omega)(D - \omega E)$ et $N = (1/\omega)\left((1-\omega)D + \omega E^*\right)$. On vérifie, puisque $D = D^*$, que $M^* + N = (1/\omega)D - E^* + ((1-\omega)/\omega)D + E^* = ((2-\omega)/\omega)D$.

Or, par hypothèse, $0 < \omega < 2$ et D est définie positive, la matrice $M^* + N$ est donc hermitienne et définie positive.

Le Théorème 1.3.1 permet de conclure. □

Remarque 1.3.2. *La condition suffisante n'exige pas que la matrice diagonale D soit définie positive. En effet cette hypothèse est une conséquence du fait que A soit définie positive.*

Remarque 1.3.3. *Indépendamment de l'équivalence qui conclut l'énoncé du Théorème 1.3.2, l'ensemble des conditions de l'énoncé forme "seulement" un ensemble de conditions suffisantes. Par exemple si D n'est pas définie positive, la méthode de relaxation successive peut converger sans que A soit définie positive. Il suffit de choisir le cas trivial suivant d'une matrice A diagonale possédant des 1 et des -1 sur la diagonale*

$$A = \begin{pmatrix} 1 & & 0 & \\ & \ddots & & \\ 0 & & 1 & \\ & & & -1 \end{pmatrix}. \tag{1.3.5}$$

La matrice A est inversible, la matrice D n'est pas définie positive. Dans ce cas on a évidemment (voir (1.2.14))

$$\mathcal{L}_\omega = D^{-1}(1-\omega)D = 1 - \omega,$$

et donc

$$\rho(\mathcal{L}_\omega) < 1, \qquad \textit{pour } 0 < \omega < 2,$$

alors que la matrice A n'est évidemment pas définie positive.

Corollaire 1.3.1. *Soit $A \in \mathbb{C}^{n,n}$ une matrice hermitienne inversible, D et $E \in \mathbb{C}^{n,n}$ telles que*

$$A = D - E - E^*.$$

On suppose que la matrice D est définie positive. Alors la méthode de Gauss-Seidel converge si et seulement si A est définie positive.

Preuve : Il suffit de faire $\omega = 1$ dans le Théorème 1.3.2. □

Le Corollaire 1.3.1 s'énonce plus simplement dans le cas ponctuel, c'est le

Corollaire 1.3.2. *Soit $A \in \mathbb{C}^{n,n}$ une matrice hermitienne inversible, D et $E \in \mathbb{C}^{n,n}$ telles que*

$$A = D - E - E^*.$$

On suppose que $a_{i,i} > 0$, $1 \leq i \leq n$. La méthode de Gauss-Seidel ponctuelle converge si et seulement si A est définie positive.

Preuve : Elle est évidente. □

Le Théorème général 1.3.1 permet aussi de donner un critère de convergence pour la méthode de Jacobi (ponctuelle ou par blocs) dans le cas d'une matrice A hermitienne définie positive

Corollaire 1.3.3. *Soit $A \in \mathbb{C}^{n,n}$ une matrice hermitienne inversible, D et $E \in \mathbb{C}^{n,n}$ telles que*

$$A = D - E - E^*.$$

On suppose que la matrice $2D - A$ est définie positive. La méthode de Jacobi converge si et seulement si A est définie positive.

Preuve : Dans la méthode de Jacobi on a $M = D$ et $N = D - A$ et donc, puisque $D = D^*$ car la matrice A est hermitienne,

$$M^* + N = 2D - A.$$

Comme la matrice $2D-A$ est définie positive par hypothèse, la conclusion suit du Théorème 1.3.1. □

Il s'agit maintenant de déterminer *l'intervalle de confiance* de convergence de la méthode de relaxation successive. Cet intervalle a été présupposé dans l'énoncé du Théorème 1.3.2. On va voir que $]0, 2[$ est effectivement cet intervalle.

La démonstration passe par celle du théorème suivant

Théorème 1.3.3. *Soit $A \in \mathbb{C}^{n,n}$ une matrice inversible, soient $E, F \in \mathbb{C}^{n,n}$ deux matrices respectivement strictement triangulaires inférieure et supérieure, $D \in \mathbb{C}^{n,n}$ une matrice inversible telles que $A = D - E - F$.*

Alors $\forall \omega \in \mathbb{R}$, $\rho(\mathcal{L}_\omega) \geq |\omega - 1|$ avec égalité si et seulement si toutes les valeurs propres de $\mathcal{L}_\omega$ sont en module égales à $|\omega - 1|$.

Preuve : Admise. □

Une conséquence immédiate de ce théorème est le

Corollaire 1.3.4. *Pour toute matrice A satisfaisant aux hypothèses du théorème 1.3.3, une condition nécessaire de convergence de la méthode de relaxation successive est que $0 < \omega < 2$.*

Preuve : En effet pour que $\rho(\mathcal{L}_\omega) < 1$, par le théorème précédent il faut que $|\omega - 1| < 1$ ce qui implique que $0 < \omega < 2$. □

Par le Théorème 1.3.3, nous allons aussi déduire le résultat déjà annoncé de façon informelle,

Théorème 1.3.4. *Soit A une matrice hermitienne (dans ce cas $F = E^*$) définie positive, alors la méthode de relaxation successive (ponctuelle ou par blocs) converge si et seulement si $0 < \omega < 2$. Autrement dit*

$$\rho(\mathcal{L}_\omega) < 1 \Leftrightarrow 0 < \omega < 2.$$

Preuve : Prouvons d'abord que $0 < \omega < 2$ entraîne que $\rho(\mathcal{L}_\omega) < 1$. Il suffit d'appliquer le Théorème 1.3.2 en remarquant que, si A est hermitienne définie positive alors elle est inversible et sa matrice diagonale D possède les mêmes propriétés. Si $0 < \omega < 2$ toutes les hypothèses du Théorème 1.3.2 sont satisfaites, comme A est définie positive alors $\rho(\mathcal{L}_\omega) < 1$.

Réciproquement si $\rho(\mathcal{L}_\omega) < 1$ on sait, grâce au Corollaire 1.3.4, que $0 < \omega < 2$. □

La question de l'intervalle de confiance de la méthode de relaxation successive étant résolue, il reste à résoudre le problème du ω_{opt}, c'est-à-dire du choix du ω tel que

$$\rho(\mathcal{L}_{\omega_{opt}}) = \min_{0<\omega<2} \rho(\mathcal{L}_\omega),$$

assurant la plus grande vitesse de convergence de la méthode de relaxation successive.

On ne va donner une réponse au choix de ω_{opt} que dans le cas où la matrice A est *tridiagonale par blocs*, c'est-à-dire que l'on se place *a priori* dans le cadre de la méthode de relaxation successive par blocs. Dans les applications cela recouvre une importante partie des besoins.

Les résultats présentés ici sont énoncés sans démonstration.

Théorème 1.3.5. *Soit $A \in \mathbb{C}^{n,n}$ une matrice tridiagonale par blocs. Si toutes les valeurs propres de la matrice d'itération de Jacobi correspondante*

sont réelles, alors les méthodes par blocs de Jacobi et de relaxation successive pour $0 < \omega < 2$ convergent ou divergent simultanèment.

De plus si $\rho(B) < 1$ (i.e. si Jacobi par blocs converge), il existe une valeur de ω et une seule, soit ω_{opt}, rendant $\rho(\mathcal{L}_\omega)$ minimum, on a

$$\omega_{opt} = \frac{2}{1+\sqrt{1-(\rho(B))^2}} \quad et \quad \rho(\mathcal{L}_{\omega_{opt}}) = \omega_{opt} - 1 = \frac{1-\sqrt{1-(\rho(B))^2}}{1+\sqrt{1-(\rho(B))^2}}. \tag{1.3.6}$$

Remarque 1.3.4. *Il faut noter que la formule (1.3.6) n'est pas d'un usage pratique immédiat, elle nécessite de calculer numériquement le rayon spectral de la matrice B. L'emploi d'une méthode de calcul de la plus grande (en module) valeur propre d'une matrice est nécessaire ; la méthode dite de la puissance répond à cet objectif.*

Remarque 1.3.5. *Pratiquement on a intérêt à surestimer ω_{opt} plutôt qu'à le sous-estimer lorsqu'on ne le connaît qu'approximativement. En effet, la courbe donnant $\rho(\mathcal{L}_\omega)$ en fonction de ω a l'allure suivante (Figure 1.2).*

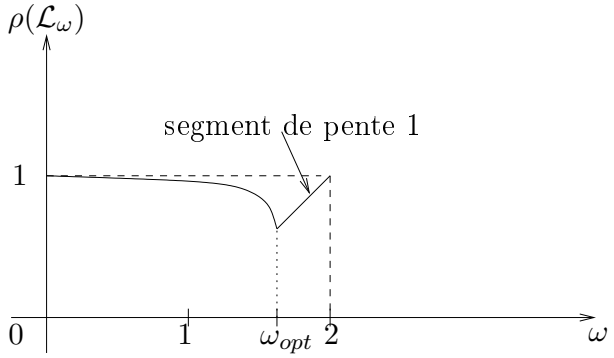

Figure 1.2:

Cette courbe possède une pente infinie à gauche du point ω_{opt} ; c'est-à-dire qu'une légère variation à gauche dans la connaissance de ω_{opt} entraine une grande variation dans celle de $\rho(\mathcal{L}_\omega)$. Ce qui justifie la sur-estimation de ω_{opt}. On remarque aussi sur la figure, que la méthode de relaxation successive optimale est plus rapide que Gauss-Seidel (associée à $\omega = 1$), on a $\rho(\mathcal{L}_{\omega_{opt}}) < \rho(\mathcal{L}_1)$. On voit aussi que, si ω est mal choisi, la méthode de Gauss-Seidel peut-être plus rapide que la méthode de relaxation successive.

Nous allons conclure ce sous-paragraphe relatif au cas des matrices hermitiennes définies positives par un résultat, très important dans les applications, donnant une condition suffisante de convergence pour les trois méthodes ainsi qu'une comparaison de celles-ci. On note toujours par B (resp. $\mathcal{L}_1$ et $\mathcal{L}_\omega$) la matrice itérative par blocs de Jacobi (resp. Gauss-Seidel et relaxation successive).

Théorème 1.3.6. *Soit A une matrice hermitienne définie positive et tridiagonale par blocs, alors les méthodes par blocs de Jacobi, Gauss-Seidel et de relaxation successive pour $0 < \omega < 2$ sont convergentes simultanément.*

De plus il existe un paramètre optimal ω_{opt}, $1 < \omega_{opt} < 2$, pour la méthode de relaxation successive et l'on a

$$\rho(\mathcal{L}_{\omega_{opt}}) < \rho(\mathcal{L}_1) < \rho(B).$$

Plus précisèment, nous avons

$$\rho(\mathcal{L}_1) = (\rho(B))^2 \tag{1.3.7}$$

$$\rho(\mathcal{L}_{\omega_{opt}}) = \omega_{opt} - 1 = \frac{1 - \sqrt{1 - (\rho(B))^2}}{1 + \sqrt{1 - (\rho(B))^2}} \tag{1.3.8}$$

Remarque 1.3.6. *Par la définition (voir livre) du taux asymptotique de convergence on voit que $R_\infty(\rho(\mathcal{L}_1)) = 2R_\infty(B)$. la méthode de Gauss-Seidel par blocs est deux fois plus rapide que la méthode de Jacobi par blocs.*

Remarque 1.3.7. *Ce théorème reste vrai lorsque A est une matrice tridiagonale par points : les "blocs" diagonaux sont alors d'ordre 1.*

Remarque 1.3.8. *Dans le cas ponctuel, ce théorème ne s'applique que si A est tridiagonale par points. Si A n'est pas, dans le cas ponctuel, tridiagonale par points, même si A est symétrique définie positive, la méthode de Jacobi ponctuelle peut diverger. Prenons pour s'en convaincre une matrice A du type*

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix} \tag{1.3.9}$$

Pour $1/2 \leq a < 1$ on vérifie que $\rho(B) \geq 1$ et pourtant la matrice A est (symétrique) définie positive - voir les calculs suivants (1.3.3).

Par ailleurs la matrice d'itération B de Jacobi a la forme suivante

$$B = \begin{pmatrix} 0 & -a & -a \\ -a & 0 & -a \\ -a & -a & 0 \end{pmatrix}. \tag{1.3.10}$$

Ses valeurs propres sont les racines du déterminant de la matrice

$$B - \lambda I = \begin{pmatrix} -\lambda & -a & -a \\ -a & -\lambda & -a \\ -a & -a & -\lambda \end{pmatrix}. \tag{1.3.11}$$

c'est, au signe près, la matrice (1.3.3) avec λ au lieu de $1 - \lambda$. Les valeurs propres de B sont donc, au signe près, $\lambda = a$ (racine double) et $\lambda = -2a$. Le rayon spectral de B est la plus grande valeur propre en module, donc $\rho(B) = 2a$; comme $1/2 \leq a < 1$, $\rho(B) \geq 1$ et Jacobi ponctuelle diverge.

On vient de voir que la convergence de la méthode de Jacobi ponctuelle n'est pas assurée par l'hypothèse symétrique définie positive, cependant elle l'est pour une autre classe de matrices qui est celle des matrices strictement diagonalement dominante (voir définition 1.3.1).

Cas des matrices SDD Ce paragraphe ne concerne que les méthodes ponctuelles. On a déjà dit que les matrices SDD sont régulières.

Pour ces matrices on a le

Théorème 1.3.7. *Soit A une matrice strictement diagonalement dominante, alors la méthode ponctuelle de Jacobi converge.*

Preuve : Notons d'abord, qu'avec cette hypothèse, l'écriture de Jacobi, $B = M^{-1}N$ avec $M = D$ et $N = E + F$, a un sens car la matrice diagonale est régulière car nécessairement $|a_{i,i}| > 0$ lorsque A est SDD.

Une condition suffisante de convergence est $\rho(B) < 1$. Prouvons-le en raisonnant par l'absurde.

Soit λ une valeur propre de B, par définition nous avons

$$\det(M^{-1}N - \lambda I) = 0.$$

Comme la matrice M est régulière, de façon équivalente, il vient

$$\det(N - \lambda M) = \det(M)\det(M^{-1}N - \lambda I) = 0. \tag{1.3.12}$$

Posons $C = N - \lambda M$, $C = (c_{i,j})$. Supposons que $|\lambda| \geq 1$. On a, d'une part, puisque M est diagonale

$$\sum_{j=1, j\neq i}^{n} |c_{i,j}| = \sum_{j=1, j\neq i}^{n} |a_{i,j}|, \quad \text{et } |c_{i,i}| = |\lambda||a_{i,i}| \; \forall i \in [1, n]; \tag{1.3.13}$$

d'autre part, puisque A est SDD et $|\lambda| \geq 1$,

$$\sum_{j=1, j\neq i}^{n} |a_{i,j}| < |a_{i,i}| \leq |\lambda||a_{i,i}| = |c_{i,i}|, \;\; \forall i \in [1, n]. \tag{1.3.14}$$

Les relations (1.3.13) et (1.3.14) impliquent que la matrice C est aussi SDD si $|\lambda| \geq 1$. La matrice C est alors régulière et donc $\det(N - \lambda M) \neq 0$, ce qui est absurde d'après (1.3.12). Donc λ telle que $|\lambda| \geq 1$ ne peut pas être valeur propre de B et nécessairement $\rho(B) < 1$. La méthode ponctuelle de Jacobi converge si A est SDD. □

Ce théorème n'énonce qu'une condition suffisante de convergence. Il se peut que Jacobi converge sans qu'elle soit SDD !

Montrons, par la même technique, que la méthode ponctuelle de Gauss-Seidel converge si A est SDD.

Théorème 1.3.8. *Lorsque A est SDD, la méthode ponctuelle de Gauss-Seidel converge.*

Preuve : La matrice d'itération s'écrit ici $\mathcal{L}_1 = (D-E)^{-1}F$, la matrice $(D-E)$ étant clairement régulière si A est SDD. Soit λ une valeur propre de $\mathcal{L}_1$, alors

$$\det((D-E)^{-1}F - \lambda I) = 0,$$

de façon équivalente, on peut encore écrire

$$\det(F - \lambda(D-E)) = 0. \tag{1.3.15}$$

Raisonnons encore par l'absurde en supposant $|\lambda| \geq 1$.
Nous avons, puisque A est SDD,

$$|a_{i,i}| > \sum_{j<i} |a_{i,j}| + \sum_{j>i} |a_{i,j}|, \quad \forall i \in [1,n].$$

soit

$$|\lambda||a_{i,i}| > |\lambda| \sum_{j<i} |a_{i,j}| + |\lambda| \sum_{j>i} |a_{i,j}| \geq |\lambda| \sum_{j<i} |a_{i,j}| + \sum_{j>i} |a_{i,j}|, \quad \forall i \in [1,n].$$

Par définition des matrices D, E et F, nous avons donc que la matrice $(F + \lambda E - \lambda D)$ est aussi SDD et donc régulière. Ce qui est absurde (voir (1.3.15)). Donc λ telle que $|\lambda| \geq 1$ ne peut pas être une valeur propre de $\mathcal{L}_1$ et nécessairement $\rho(\mathcal{L}_1) < 1$. □

Remarque 1.3.9. *Les Théorèmes 1.3.7 et 1.3.8 restent vrais lorsque A est irréductible et diagonalement dominante au lieu d'être SDD.*

Remarque 1.3.10. *Les Théorèmes 1.3.7 et 1.3.8, ainsi que la Remarque 1.3.9 n'énoncent que des conditions suffisantes de convergence.*

Finalement on démontre (voir livre) que la méthode de relaxation successive pour $0 < \omega \leq 1$ converge si A est SDD.

Théorème 1.3.9. *Si A est SDD, la méthode de relaxation successive $\mathcal{L}_\omega$ est convergente pour $0 < \omega \leq 1$.*

Preuve : Admise. □

Remarque 1.3.11. *Évidemment ce théorème recouvre le cas de Gauss-Seidel puisqu'il est vrai pour $0 < \omega \leq 1$ (la valeur 1 incluse).*

Remarque 1.3.12. *L'intervalle de confiance de la méthode de relaxation successive étant l'intervalle $]0,2[$, on aimerait pouvoir étendre la preuve précédente de la convergence à cet intervalle. Malheureusement le raisonnement du Théorème 1.3.9 n'aboutit pas dans ce cas.*

Remarque 1.3.13. *Ce théorème n'est qu'une condition suffisante de convergence. Il ne dit pas que si $\omega > 1$ la méthode de relaxation diverge. Soit en effet la matrice*

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

Pour $0 < a < 1$ cette matrice est SDD. Il est facile de voir que la matrice $\mathcal{L}_\omega$ s'écrit

$$\mathcal{L}_\omega = \begin{pmatrix} 1-\omega & -\omega a \\ 0 & 1-\omega \end{pmatrix},$$

sa valeur propre (double) est $1-\omega$. La méthode de relaxation converge si $|1-\omega| < 1$, c'est-à-dire pour $0 < \omega < 2$.