Méthodes itératives pour des systèmes linéaires

On sait aujourd'hui résoudre numériquement des systèmes linéaires de l'ordre du million d'inconnues (et d'équations). Pour des **systèmes creux**, c'est-à-dire lorsque la matrice du système possède beaucoup de coefficients nuls, on arrive à une centaine de millions d'inconnues. Les **systèmes pleins** font appel à des **méthodes directes**, qui donnent la solution exacte (aux erreurs d'arrondi près) en un nombre fini d'itérations, et seront décrites dans un chapitre ultérieur.

Pour les **très grands systèmes creux**[1] on utilise des **méthodes itératives**, où on construit une suite de vecteurs qui convergent vers la solution.

L'intérêt est que **ces méthodes ne manipulent pas la matrice**, mais seulement une fonction qui définit une suite par récurrence.

Définition 1 Soit $A \in M_n(\mathbb{R})$ inversible et $b \in \mathbb{R}^n$. On appelle méthode itérative de résolution du système linéaire $Ax = b$, $(x \in \mathbb{R}^n)$ une méthode qui construit une suite récurrente $(x_k)_{k\geq 0}$ telle que

$$(x_k \underset{k\to+\infty}{\longrightarrow} x) \Rightarrow Ax = b$$

Une méthode itérative est convergente si $x_k \underset{k\to+\infty}{\longrightarrow} x$ pour toute condition initiale $x_0 \in \mathbb{R}^n$

Tests d'arrêt typiques :

- $\frac{\|Ax_k-b\|}{\|b\|} < \varepsilon$ (norme du "résidu" / norme de b).
  Noter que :

$$\begin{aligned}\frac{\|x_k - x\|}{\|x\|} = \frac{\|A^{-1}(Ax_k - b)\|}{\|x\|} &\leq \|A^{-1}\| \frac{\|b\|}{\|x\|}\varepsilon \\ &\leq \|A^{-1}\|\|A\|\varepsilon, \text{ peut-être grand!}\end{aligned}$$

Nous allons décrire ici des méthodes itératives avec "splitting" de A.

1. Exemple : discrétisation par différences finies de problèmes aux limites pour des équations aux dérivées partielles ... + schéma

1 Description générale :

On considère ici le système

$$Ax = b \tag{1}$$

où $A \in M_n(\mathbb{R})$, $x \in \mathbb{R}^n$ et $b \in \mathbb{R}^n$. On suppose que la matrice A est inversible.

On considère une décomposition de A ("splitting") $A = M - N$ avec M inversible et on considère l'itération :

$$\left\{ \begin{array}{lcl} Mx_{k+1} & = & Nx_k + b \\ \\ x_0 \in \mathbb{R}^n & & \end{array} \right. \tag{2}$$

Si $x_k \to x$ quand $k \to +\infty$ alors $Mx = Nx + b$, càd x est solution de (1).

Le choix du splitting est très important pour la performance de la méthode :

- Bien sûr la méthode doit être convergente (voir plus loin)
- On doit choisir M de telle sorte que le système (2) soit beaucoup plus facile à résoudre que (1) (il faut résoudre (2) à chaque étape de l'itération).

  Exemples : M diagonale ou triangulaire, diagonale ou triangulaire par blocs.

Étudions les conditions de convergence de (2).

Définition 2 Étant donné $A \in M_n(\mathbb{C})$, on note $S_p(A)$ l'ensemble des valeurs propres de A (ou "spectre de A"). On appelle rayon spectral de A et on note $\rho(A)$:

$$\rho(A) = \operatorname*{Max}_{\lambda \in S_p(A)} |\lambda|$$

Theoreme 1 *La méthode (2) converge si et seulement si*

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

La preuve complète de ce résultat sera étudiée en TD. Ici nous allons simplement montrer que $\rho(M^{-1}N) < 1 \Rightarrow$ convergence de (2), en admettant pour cela deux résultats.

Theoreme 2 (de l'application contractante (dans $\mathbb{R}^n$)) *Soit E un sous-ensemble de $\mathbb{R}^n$ fermé (et non vide). On considère une norme $\| \, \|$ sur $\mathbb{R}^n$.*

Soit $F : E \to E$ une application contractante, càd pour laquelle il existe $\alpha \in [0,1[$ tel que :

$$\|F(x) - F(y)\| \leq \alpha \|x - y\|, \ \forall x, y \in E$$

Alors il existe un unique $x^* \in E$ tel que $F(x^*) = x^*$ (càd F admet un unique point fixe dans E). De plus, pour tout $x_0 \in E$, la suite définie par :

$$x_{k+1} = F(x_k)$$

converge vers x^*, avec

$$\|x^* - x_k\| \leq \frac{\alpha^k}{1-\alpha} \|x_1 - x_0\| \tag{3}$$

Le système (2) s'écrit :

$$\begin{cases} \boxed{x_{k+1} = M^{-1}N x_k + M^{-1}b} \\ \\ x_0 \in \mathbb{R}^n \end{cases} \tag{4}$$

Remarque 1 Théorème encore appelé “Théorème du point fixe de Banach”. Le théorème reste vrai lorsque E est un espace métrique complet.

Theoreme 3 (cf TD pour la démonstration) *Soit $A \in M_n(\mathbb{C})$ et $\varepsilon > 0$. Il existe une norme $\| \ \|$ sur $\mathbb{C}^n$ telle que :*

$$\underbrace{\|A\|}_{\substack{\textit{norme sur} \\ M_n(\mathbb{C}) \textit{ induite} \\ \textit{par la norme} \\ \| \ \| \textit{ de } C^n}} := \operatorname{Sup}_{\|x\|=1} \|Ax\| \leq \rho(A) + \varepsilon$$

Si $\rho(M^{-1}N) < 1$, il existe donc une norme matricielle induite telle que $\|M^{-1}N\| \leq \rho(M^{-1}N) + \varepsilon < 1$. Alors :

$$\|F(x) - F(y)\| = \|M^{-1}N(x-y)\| \leq \underbrace{\|M^{-1}N\|}_{<1} \|x - y\|$$

Donc $F : \mathbb{R}^n \to \mathbb{R}^n$ est une contraction.

Donc $\forall x_0 \in R^n$, la suite définie par (4) converge vers une limite $x \in \mathbb{R}^n$ unique, solution de $x = M^{-1}Nx + M^{-1}b$, c'est-à-dire $Ax = b$.

Vitesse de convergence : Plus $\rho(M^{-1}N)$ est petit, plus $\|M^{-1}N\|$ peut être choisie petite et plus la convergence est rapide. En effet, d'après (3) :

$$\|x - x_k\| \leq \frac{\|M^{-1}N\|^k}{1 - \|M^{-1}N\|}\|x_1 - x_0\|$$

Exemple de splitting : (peu utilisé)

$$M = \frac{1}{\alpha}I, \qquad N = \frac{1}{\alpha}I - A \qquad \Longrightarrow \qquad x_{k+1} = (I - \alpha A)x_k + \alpha b$$

(méthode de Richardson stationnaire, ou du gradient à pas fixe)

Elle converge si et seulement si $\forall \lambda \in Sp(A), |1 - \alpha\lambda| < 1$, c'est-à-dire toutes les valeurs propres de A se trouvent dans le disque (ouvert) de centre ($\frac{1}{\alpha}$ et rayon $\frac{1}{\alpha}$).

2 Méthode de Jacobi

On pose dans schéma (2) :

$$M = D \text{ avec } D \text{ diagonale et } d_{ii} = a_{ii}, \qquad N = D - A$$

Remarque 2 Cela suppose $a_{ii} \neq 0\ \forall i$ (si cette condition n'est pas vérifiée on peut permuter des lignes de A).

Theoreme 4 *Si A est à diagonale strictement dominante ($\rightarrow a_{ii} > 0$ et D inversible) alors la méthode de Jacobi converge.*

La démonstration sera vue en TD. On montre que le rayon spectral de la matrice $J = D^{-1}(D - A) = I - DA$ est < 1.

Nous avons rencontré ce type de matrices pour la discrétisation de problèmes aux limites dans le 1^{er} chapitre du cours.

3 Méthodes de Gauss-Seidel et SOR

On pose $A = D + L + U$ avec :

$$D = \begin{pmatrix} a_{00} & \cdots & 0 \\ \vdots & a_{ii} & \vdots \\ 0 & \cdots & a_{nn} \end{pmatrix}, L = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ a_{ij}(i > j) & \cdots & 0 \end{pmatrix}, U = \begin{pmatrix} 0 & \cdots & a_{ij}(j > i) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

Dans la méthode de Gauss-Seidel, on fixe :

$$M = D + L,\ N = -U$$

La méthode s'écrit donc :

$$Dx_{k+1} = -Lx_{k+1} - Ux_k + b$$

En notant $x_k = (x_1^{(k)}, \dots, x_n^{(k)})$ on obtient pour $i = 1..n$:

$$a_{ii}x_i^{(k+1)} = -\sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{(j>i)} a_{ij}x_j^{(k)} + b_i$$

Les méthodes de Jacobi et Gauss-Seidel ne sont guère utilisées. On leur préfère la méthode de relaxation.

La méthode SOR (“successive over-relaxation”, ou “méthode de relaxation”, environ 1950) généralise Gauss-Seidel en introduisant un paramètre de relaxation $\omega \neq 0$, que l'on ajuste afin d'accélérer la convergence de la méthode (avec un gain généralement très important si ω est bien choisi).

Pour $i = 1, \dots, n$

$$\left\{ \begin{array}{lcl} a_{ii}\tilde{x}_i^{(k+1)} & = & -\sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)} + b_i \\ x_i^{(k+1)} & = & \omega\tilde{x}_i{}^{(k+1)} + (1-\omega)x_i^{(k)} \end{array} \right. \tag{5}$$

(Gauss-Seidel correspond à $\omega = 1$)

La méthode s'écrit (multiplier la seconde ligne par a_{ii}, et remplacer $a_{ii}\tilde{x}_i{}^{(k+1)}$ par son expression en fonction de x_{k+1} et x_k)

$$Dx_{k+1} = (1-\omega)Dx_k - \omega Lx_{k+1} - \omega Ux_k + \omega b$$

soit

$$(D + \omega L)x_{k+1} = [(1-\omega)D - \omega U]x_k + \omega b$$

On a donc :

$$M = \frac{1}{\omega}D + L, N = \frac{1-\omega}{\omega}D - U, M - N = D + L + U = A$$

On note :

$$\mathcal{L}_\omega := (\frac{1}{\omega}D + L)^{-1}(\frac{1-\omega}{\omega}D - U)$$

SOR converge si $\rho(\mathcal{L}_\omega < 1)$

Theoreme 5 (demo en TD) 1. *Soit $A \in M_n(\mathbb{R})$ inversible, avec $\forall i, a_{ii} \neq 0$ Une condition nécessaire pour que SOR converge est que $\omega \in]0, 2[$.*

2. *Si A est symétrique définie positive, alors $\forall \omega \in]0, 2[$, SOR converge*

Rappel 1 A est symétrique définie positive si A est symétrique, ${}^t xAx = 0 \Rightarrow x = 0$, et $\forall x \in \mathbb{R}^n, {}^t xAx \geq 0$

Méthode		Itérations	Temps CPU
Jacobi		34 900	902
Gauss-Seidel		20 450	1121
Relaxation	$\omega = 1,8$	3 270	180
Relaxation	$\omega = 1,93$	1 200	66
Relaxation	$\omega = 1,98$	530	24,7
Gradient conjugué		539	24,6
Gradient conjugué. Tridiagonale		443	44
Gradient conjugué SSOR	$\omega = 1,0$	152	15
Gradient conjugué SSOR	$\omega = 1,8$	57	5,8
Gradient conjugué SSOR	$\omega = 1,93$	40	4,1

FIGURE 1 – Exemples pratiques

Corollaire 1 Si A est symétrique définie positive, alors la méthode de Gauss-Seidel converge. Nous avons rencontré ce type de matrices pour la discrétisation des problèmes aux limites dns le 1^{er} chapitre du cours (paragraphe 2), cas où la fonction p est identiquement nulle.

Remarque 3

1. Il y a des exemples où A est symétrique définie positive et où la méthode de Jacobi n'est pas convergente.
2. Si A est tridiagonale ($a_{ij} = 0$ si $|i - j| > 1$) et D inversible, on peut montrer que $\rho(\mathcal{L}_1) = \rho(J)^2$. Donc la méthode de Gauss-Seidel converge si et seulement si celle de Jacobi converge (et Gauss-Seidel converge plus vite).
3. Pour quelques types de matrices, on connaît la valeur de ω qui minimise $\rho(\mathcal{L}_\omega)$. Pour A tridiagonale (avec D inversible), et si les valeurs propres de J sont réelles, alors le paramètre de relaxation optimal dans SOR (c'est-à-dire la valeur de ω qui minimise $\rho(\mathcal{L}_\omega)$) est > 1 (donc Gauss-Seidel ne donne pas la vitesse optimale de convergence).
4. Pour optimiser empiriquement le choix de ω dans SOR, on peut évaluer le facteur de contractivité $\frac{\|x_{k+1}-x_k\|}{\|x_k-x_{k-1}\|}$ à partir du moment où $\|x_{k+1} - x_k\|$ décroît vers 0.