3.3

Travail demandé :

Il vous est demandé d'étudier puis de présenter le(s) texte(s) joint(s) à travers un exposé de synthèse d'une durée comprise entre 15 et 20 minutes.

Si l'étude de la totalité du dossier et la préparation d'un exposé cohérent dans la durée impartie ne vous paraît pas possible, vous pouvez décider de vous limiter à une partie du dossier.

Remarques générales :

1. Les textes proposés, quelle que soit leur origine, peuvent présenter des défauts (coquilles typographiques, négligences ou sous-entendus de l'auteur, voire erreurs...) qui, sauf exception, n'ont pas été corrigés.

2. Les textes proposés peuvent contenir des exercices qu'il n'est pas demandé de résoudre. Néanmoins, vous pouvez vous aider des énoncés de ces exercices pour enrichir votre exposé.

3. Vous pouvez annoter les documents qui vous sont fournis. Vos annotations ne seront pas regardées par l'examinateur.

Remarques particulières :

1. Dans l'étude du texte, on pourra se limiter à ne considérer que des matrices à coefficients réels. Dans ce cas, une matrice hermitienne est symétrique.

2. On note ici $\mathbb{K}$ le corps des réels ou le corps des complexes. Soit $\|\cdot\|$ une norme sur $\mathbb{K}^n$; on rappelle que l'application qui associe à une matrice $M \in \mathcal{M}_n(\mathbb{K})$ la quantité $\max\{\|M(x)\|; \|x\| = 1$ définit une norme sur $\mathcal{M}_n(\mathbb{K})$ qu'on appelle *norme subordonnée à* $\|\cdot\|$; une norme N sur $\mathcal{M}_n(\mathbb{K})$ est appelée *subordonnée* s'il existe une norme sur $\mathbb{K}^n$ à laquelle N est subordonnée.

3. Soit $M \in \mathcal{M}_n(\mathbb{K})$. On appelle spectre de M l'ensemble des zéros complexes de son polynôme caractéristique et on appelle *rayon spectral* de M, noté $\rho(M)$ le maximum des modules des valeurs propres de M.

4. Le Théorème de Householder, cité dans le document (p. 100), affirme que pour toute matrice $M \in \mathcal{M}_n(\mathbb{K})$ et tout réel $\epsilon > 0$, il existe une norme subordonnée N sur $\mathcal{M}_n(\mathbb{C})$ telle que $N(M) \leq \rho(M) + \epsilon$.

5. Une matrice M est dite à diagonale fortement (*resp.* strictement) dominante si on a

$$\forall i \ , \ |M_{ii}| \geq \sum_{j \neq i} |M_{ij}|,$$

où l'une au moins des inégalités est stricte (*resp.* toutes les inégalités sont strictes).
Une matrice $M \in \mathcal{M}_n(\mathbb{K})$ est dite irréductible s'il n'existe pas de partition non triviale $\{1, 2, \ldots, n\} = I \cup J$ où $M_{ij} = 0$ pour tout couple $(i, j) \in I \times J$.

Méthodes itératives de résolution de problèmes linéaires

Dans ce chapitre, le corps de base est $K = \mathbb{R}$ ou $\mathbb{C}$.

Le calcul direct de la solution d'un système linéaire de Cramer $Ax = b$, par exemple par la méthode de Gauss[1], est assez coûteux, de l'ordre de n^3 opérations. Il présente aussi plusieurs inconvénients. D'une part, il ne tire pas complètement parti de la présence éventuelle d'un grand nombre de zéros dans la matrice A ; il est pourtant fréquent que la matrice, de taille $n \times n$, n'ait que $\mathcal{O}(n)$ termes non nuls, au lieu de $\mathcal{O}(n^2)$. D'autre part, le calcul d'une décomposition LU est assez instable, dans le sens où les erreurs d'arrondis produites par le calculateur sont amplifiées au cours des étapes du calcul.

Pour ces raisons, on utilise plutôt en pratique une méthode itérative pour calculer une solution approchée x^m, au lieu d'une solution exacte. Ces méthodes profitent pleinement de la structure creuse de A. Le nombre d'opérations à réaliser est $\mathcal{O}(am)$, où a est le nombre de termes non nuls dans A. Le choix de m dépend de la précision demandée, mais il n'est jamais très grand car l'erreur $\|x^m - \bar{x}\|$ avec la solution exacte $\bar{x}$ est de l'ordre de cste $\cdot\, k^m$, où $k < 1$ dès lors que la méthode est convergente. Typiquement, une dizaine d'itérations donnent un très bon résultat et alors $\mathcal{O}(10a) << \mathcal{O}(n^3)$. Un autre avantage des méthodes itératives est de réduire les erreurs d'arrondis, au lieu de les amplifier.

1. Ou par toute autre méthode directe.

Principe général. On choisit une décomposition de A sous la forme $M - N$ et l'on réécrit le système, supposant que M est inversible :

$$x = M^{-1}(Nx + b).$$

Choisissant alors une première solution approchée $x^0 \in K^n$, qui peut éventuellement être très médiocre, on construit une suite $(x^m)_{m\in\mathbb{N}}$ par récurrence :

$$x^{m+1} = M^{-1}(Nx^m + b). \tag{9.1}$$

Dans la pratique, on ne calcule pas explicitement M^{-1}, mais on résout des systèmes linéaires $Mx^{m+1} = \cdots$. Il est donc important que cette résolution soit peu coûteuse. Ce sera le cas lorsque M est triangulaire ; dans ce cas, l'inversibilité de M se lit sur sa diagonale, les coefficients diagonaux doivent être non nuls.

9.1 CRITÈRE GÉNÉRAL DE CONVERGENCE

Définition: Supposons que A et M soient inversibles, A = M - N. On dit qu'une méthode itérative est convergente si pour tout couple (x^0 , b) la suite (x^m) converge vers $A^{-1}b$ lorsque m tend vers l'infini.

Proposition 9.1.1. *Une méthode itérative est convergente si et seulement si* $\rho(M^{-1}N) < 1$.

Démonstration. Si la méthode est convergente, alors, pour $b = 0$,

$$\lim_{m\to+\infty} (M^{-1}N)^m x^0 = 0,$$

pour tout $x^0 \in K^n$. Autrement dit,

$$\lim_{m\to+\infty} (M^{-1}N)^m = 0.$$

D'après le corollaire 4.4.1., cela implique $\rho(M^{-1}N) < 1$.

Réciproquement, si $\rho(M^{-1}N) < 1$, alors

$$\lim_{m\to+\infty} (M^{-1}N)^m = 0,$$

A^{-1}

de sorte que

$$x^m - A^{-1}b = (M^{-1}N)^m(x^0 - A^{-1}b) \to 0. \qquad \square$$

De manière plus précise, si $\|\cdot\|$ est une norme sur K^n,

$$\|x^m - A^{-1}b\| \leqslant \|(M^{-1}N)^m\|\,\|x^0 - A^{-1}b\|.$$

D'après le théorème de Householder, il existe, pour tout $\varepsilon > 0$, une constante $C(\varepsilon) < \infty$ telle que

$$\|x^m - A^{-1}b\| \leqslant C(\varepsilon)\|x^0 - \bar{x}\|(\rho(M^{-1}N) + \varepsilon)^m.$$

Dans la plupart des cas (en fait lorsqu'il existe une norme induite vérifiant $\|M^{-1}N\| = \rho(M^{-1}N)$), on peut choisir $\varepsilon = 0$ dans cette inégalité, de sorte que

$$\|x^m - A^{-1}b\| = \mathcal{O}(\rho(M^{-1}N)^m).$$

Le choix d'un vecteur x^0 de sorte que $x^0 - A^{-1}b$ soit un vecteur propre associé à une valeur propre de module maximal, montre que cette inégalité ne peut pas être améliorée en général. Pour cette raison, on appelle *taux de convergence* de la méthode le nombre (positif, sinon la méthode ne converge pas)

$$\tau := -\log \rho(M^{-1}N).$$

De deux méthodes convergentes, on dira que la première converge plus vite que la seconde si $\tau_1 > \tau_2$, par exemple, elle converge deux fois plus vite si $\tau_1 = 2\tau_2$. En effet, avec une erreur de l'ordre de $\rho(M^{-1}N)^m = \exp(-m\tau)$, on voit qu'il faut deux fois moins d'itérations pour parvenir à la même précision garantie.

9.2 QUELQUES MÉTHODES COURANTES

Il y a trois méthodes principales, dont la première n'a qu'un intérêt historique ou théorique. Chacune utilise la décomposition de A en sa partie diagonale D et ses parties triangulaire inférieure $-E$ et supérieure $-F$:

$$A = D - E - F = \begin{pmatrix} d_1 & & \\ & \ddots & -F \\ & -E & \ddots \\ & & d_n \end{pmatrix}.$$

Dans tous les cas, on suppose que D est inversible : les coefficients diagonaux de A sont non nuls.

9.2.1 Méthode de Jacobi

On choisit $M = D$ et donc $N = E + F$. La matrice de l'itération est $J := D^{-1}(E + F)$. Connaissant le vecteur x^m, on calcule les composantes du vecteur x^{m+1} par les formules

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

9.2.2 Méthode de Gauss-Seidel

On choisit $M = D - E$ et donc $N = F$. La matrice de l'itération est $G := (D - E)^{-1}F$ (qu'on ne calcule jamais explicitement). Le calcul des solutions approchées se fait

par une récurrence *double*, sur m d'une part, sur $i \in \{1, \dots, n\}$ d'autre part :

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

Le changement, par rapport à la méthode de Jacobi, est qu'on utilise toujours les dernières valeurs calculées de chaque coordonnée.

9.2.3 Méthode de relaxation

On peut vouloir améliorer la méthode de Gauss-Seidel en cherchant une « meilleure » valeur approchée des x_j (avec $j < i$) lors du calcul de x_i^{m+1}. Au lieu d'être simplement x_j^m, comme dans la méthode de Jacobi, ou x_j^{m+1}, comme dans celle de Gauss-Seidel, cette meilleure valeur devrait être une interpolation de ces deux-ci (nous verrons qu'il s'agit plutôt d'une extrapolation). C'est ce qui justifie le choix de

$$M = \frac{1}{\omega} D - E, \quad N = \left(\frac{1}{\omega} - 1 \right) D + F,$$

où $\omega \in \mathbb{C}$ est un paramètre. Celui-ci reste en général constant au cours du calcul. La matrice de l'itération est

$$\mathcal{L}_\omega := (D - \omega E)^{-1}((1 - \omega)D + \omega F).$$

La méthode de Gauss-Seidel est un cas particulier, avec $\omega = 1$: $\mathcal{L}_1 = G$. Une attention toute particulière doit être accordée au choix de ω, afin de réaliser le minimum de $\rho(\mathcal{L}_\omega)$. Le calcul des solutions approchées résulte à nouveau d'une récurrence double :

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

En l'absence d'hypothèse complémentaire, relative à la matrice A, le seul résultat concernant la convergence est le suivant :

Proposition 9.2.1. *Si la méthode de relaxation converge pour une matrice $A \in M_n(\mathbb{C})$ et un paramètre $\omega \in \mathbb{C}$, alors*

$$|\omega - 1| < 1.$$

Autrement dit, il est nécessaire que ω soit dans le disque dont $]0, 2[$ est un diamètre.

Démonstration. Si la méthode est convergente, on a $\rho(\mathcal{L}_\omega) < 1$. Cependant,

$$\det \mathcal{L}_\omega = \frac{\det((1 - \omega)D + \omega F)}{\det(D - \omega E)} = \frac{\det((1 - \omega)D)}{\det D} = (1 - \omega)^n.$$

Ainsi,

$$\rho(\mathcal{L}_\omega) \geqslant |\det \mathcal{L}_\omega|^{1/n} = |1 - \omega|.$$

□

9.3 DEUX CAS DE CONVERGENCE

Dans cette section et la suivante, nous montrons que des hypothèses simples et naturelles sur A entraînent la convergence des méthodes classiques. Nous comparons aussi leur rapidité.

9.3.1 Cas à diagonale dominante

On suppose ici que l'une des deux propriétés ci-dessous est satisfaite :

1. A est à diagonale strictement dominante ;
2. A est irréductible, à diagonale fortement dominante.

Proposition 9.3.1. *Sous l'une ou l'autre des hypothèses (1) ou (2), la méthode de Jacobi converge, ainsi que la méthode de relaxation avec* $\omega \in]0, 1]$.

Démonstration. **Admisse.**

9.3.2 Cas d'une matrice hermitienne définie positive

Commençons par un résultat intermédiaire.

Lemme 9.3.1 *Si A est hermitienne définie positive, ainsi que $M^* + N$ (dans une décomposition $A = M - N$), alors $\rho(M^{-1}N) < 1$.*

Démonstration. Remarquons tout d'abord que $M^* + N = M^* + M - A$ est nécessairement hermitienne lorsque A l'est.
Il suffit donc de montrer que $\|M^{-1}Nx\|_A < \|x\|_A$ pour tout $x \in \mathbb{C}^n$, non nul, où on a noté $\|\cdot\|_A$ la norme associée à A :

$$\|x\|_A = \sqrt{x^*Ax}.$$

Nous avons $M^{-1}Nx = x - y$ avec $y = M^{-1}Ax$. Ainsi,

$$\begin{aligned} \|M^{-1}Nx\|_A^2 &= \|x\|_A^2 - y^*Ax - x^*Ay + y^*Ay \\ &= \|x\|_A^2 - y^*(M^* + N)y. \end{aligned}$$

On conclut en remarquant que y n'est pas nul, donc $y^*(M^* + N)y > 0$. ❑

Cette preuve donne un résultat un peu plus précis que celui qui était annoncé : en prenant le supremum de $\|M^{-1}Nx\|_A$ sur la boule unité, qui est compacte, on obtient $\|M^{-1}N\| < 1$, pour la norme matricielle induite par $\|\cdot\|_A$.

L'application essentielle de ce lemme est

Théorème 9.3.1. *Si A est hermitienne définie positive, la méthode de relaxation converge si et seulement si $|\omega - 1| < 1$.*

Démonstration. On a vu à la proposition 9.2.1. que la convergence entraîne $|\omega - 1| < 1$. Voyons la réciproque. On a $E^* = F$ et $D^* = D$, de sorte que

$$M^* + N = \left(\frac{1}{\omega} + \frac{1}{\overline{\omega}} - 1\right) D = \frac{1 - |\omega - 1|^2}{|\omega|^2} D.$$

Comme D est définie positive, $M^* + N$ l'est si et seulement si $|\omega - 1| < 1$. ❑

En revanche, le lemme 9.3.1 ne s'applique pas à la méthode de Jacobi, puisque l'hypothèse (A définie positive) n'entraîne pas que $M^* + N = D + E + F$ le soit. Nous verrons en exercice que cette méthode diverge pour certaines matrices $A \in HDP_n$, bien qu'elle converge lorsque $A \in HDP_n$ est tridiagonale.

9.4 CAS TRIDIAGONAL

Nous considérons ici le cas des matrices A tridiagonales qu'on rencontre fréquemment dans le traitement des équations aux dérivées partielles par des méthodes de différences finies ou d'éléments finis. La structure générale de A est la suivante :

$$A = \begin{pmatrix} x & x' & 0 & \cdots & 0 \\ x'' & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & y' \\ 0 & \cdots & 0 & y'' & y \end{pmatrix}.$$

Autrement dit, les coefficients a_{ij} sont nuls dès que $|j - i| \geqslant 2$.

Éventuellement, ces matrices sont tridiagonales par blocs, c'est-à-dire que les a_{ij} sont des matrices, les blocs diagonaux étant des matrices carrées. Dans ce cas, les méthodes itératives doivent aussi s'écrire par blocs, la décomposition $A = D - E - F$ étant faite par blocs. Les méthodes itératives correspondantes nécessitent l'inversion de matrices de petites tailles, à savoir les a_{ii}, qu'on réalise par une méthode directe. Nous ne détaillerons pas ici cette extension des méthodes classiques.

La structure de la matrice permet d'écrire une relation algébrique fort utile :

Lemme 9.4.1 *Soit μ un nombre complexe non nul et $C = C_0 + C_- + C_+$ une matrice tridiagonale. Alors*

$$\det C = \det\left(C_0 + \frac{1}{\mu}C_- + \mu C_+\right).$$

Démonstration. Il suffit de noter que la matrice C est conjuguée à

$$C_0 + \frac{1}{\mu}C_- + \mu C_+,$$

par la matrice de changement de base

$$Q_\mu = \begin{pmatrix} \mu & & & \\ & \mu^2 & & 0 \\ & & \ddots & \\ & 0 & & \ddots \\ & & & & \mu^n \end{pmatrix}.$$

❑

Appliquons le lemme au calcul du polynôme caractéristique P_ω de $\mathcal{L}_\omega$. On a

$$\begin{aligned}(\det D)P_\omega(\lambda) &= \det((D-\omega E)(\lambda I_n - \mathcal{L}_\omega)) \\ &= \det((\omega+\lambda-1)D - \omega F - \lambda\omega E) \\ &= \det\left((\omega+\lambda-1)D - \mu\omega F - \frac{\lambda\omega}{\mu}E\right),\end{aligned}$$

pour tout μ non nul. Choisissons pour μ n'importe quelle racine carrée de λ. Il vient

$$\begin{aligned}(\det D)P_\omega(\mu^2) &= \det((\omega+\mu^2-1)D - \mu\omega(E+F)) \\ &= (\det D)\det((\omega+\mu^2-1)I_n - \mu\omega J).\end{aligned}$$

Finalement,

Lemme 9.4.2 *Si A est tridiagonale et D inversible, alors*

$$P_\omega(\mu^2) = (\mu\omega)^n P_J\left(\frac{\mu^2+\omega-1}{\mu\omega}\right).$$

Commençons par analyser un cas simple, celui de la méthode de Gauss-Seidel, pour laquelle $G = \mathcal{L}_1$. Nous obtenons :

Proposition 9.4.1. *Si A est tridiagonale et D inversible, alors*

1) $P_G(X^2) = X^n P_J(X)$;
2) $\rho(G) = \rho(J)^2$;
3) *la méthode de Gauss-Seidel converge si et seulement si celle de Jacobi converge ; de plus en cas de convergence, la méthode de Gauss-Seidel converge deux fois plus vite que celle de Jacobi ;*
4) *le spectre de J est pair : $SpJ = -SpJ$.*

Démonstration. La formule 1) découle du lemme 9.4.2. Le spectre de G est donc formé de $\lambda = 0$ (qui est de multiplicité $[(n+1)/2]$ au moins) et des carrés des valeurs propres de J, ce qui prouve 2). Le point 3) en découle immédiatement. Enfin, si $\mu \in \mathrm{Sp}J$, alors $P_J(\mu) = 0$ et aussi $P_G(\mu^2) = 0$, de sorte que $(-\mu)^n P_J(-\mu) = 0$. Finalement, $P_J(-\mu) = 0$, ou bien $\mu = 0 = -\mu$, auquel cas $P_J(-\mu)$ est quand même nul. ❑

Nous retournons au cas général, avec une hypothèse supplémentaire : le spectre de J est réel et la méthode de Jacobi converge. Cette propriété est, par exemple, satisfaite lorsque A est hermitienne définie positive, puisque le théorème 9.3.1 et la proposition 9.4.1. assurent la convergence de la méthode de Jacobi, et puisque J est semblable à la matrice hermitienne $D^{-1/2}(E+F)D^{-1/2}$.

Nous choisissons aussi ω réel, c'est-à-dire $\omega \in]0,2[$, compte tenu de la proposition 9.2.1. Le spectre de J est donc constitué des valeurs propres

$$-\lambda_r < \cdots < -\lambda_1 \leqslant \lambda_1 < \cdots < \lambda_r = \rho(J) < 1,$$

d'après la proposition 9.4.1. Outre la valeur propre nulle, qui n'entre pas en ligne de compte pour le calcul du rayon spectral, les valeurs propres de $\mathcal{L}_\omega$ sont les nombres $\mu_{a\pm}^2$, où μ_{a-}, μ_{a+} sont les racines de

$$\mu^2 + \omega - 1 = \mu\omega\lambda_a, \quad 1 \leqslant a \leqslant r.$$

En effet, prendre $-\lambda_a$ au lieu de λ_a fournit les mêmes valeurs $\mu_{a\pm}^2$.

Notons $Q_{\lambda,\omega}(X)$ le polynôme $X^2 + \omega - 1 - X\omega\lambda$ et définissons $M(\omega,\lambda)$ comme le maximum des modules des deux racines de $Q_{\lambda,\omega}$. Si $\Delta := \omega^2\lambda^2 + 4(1-\omega)$ est négatif, les racines de $Q_{\lambda,\omega}$ sont complexes conjugées, de module $|\omega - 1|^{1/2}$. On a donc $M(\omega,\lambda) = |\omega - 1|^{1/2} < 1$. Dans le cas contraire, les racines sont réelles distinctes. Supposant λ positif (voir pourquoi plus haut), leur somme $\omega\lambda$ est positive, de sorte que l'une d'elles, la plus grande en module, est positive. De plus, l'une des racines au moins est dans $]-1,1[$ puisque leur produit $\omega - 1$ est de module strictement inférieur à 1. Enfin, $Q_{\lambda,\omega}(1)Q_{\lambda,\omega}(-1) = \omega^2(1-\lambda^2) > 0$ montre que $]-1,1[$ contient un nombre pair de racines.

En résumé, si $\Delta > 0$ et $\lambda \in [0,1[$, les deux racines de $Q_{\lambda,\omega}$ sont dans $]-1,1[$, la plus grande étant $M(\omega,\lambda)$. De plus, $M(\omega,\lambda) \geqslant |\omega - 1|^{1/2}$.

Montrons maintenant que $\lambda \mapsto M(\omega,\lambda)$ est une application croissante sur $[0,1[$. Soit $0 \leqslant \lambda < \eta < 1$. Si $\omega^2\lambda^2 + 4(1-\omega) \leqslant 0$, on sait que $M(\omega,\lambda) = |\omega - 1|^{1/2} \leqslant M(\omega,\eta)$. Si au contraire $\omega^2\lambda^2 + 4(1-\omega) > 0$, notons $\mu_- < \mu_+ = M(\omega,\lambda)$ les racines de $Q_{\lambda,\omega}$. Nous avons $Q_{\eta,\omega}(1) = \omega(1-\eta) > 0$, tandis que $Q_{\eta,\omega}(\mu_+) = \mu_+\omega(\lambda - \eta) < 0$, de sorte que $Q_{\eta,\omega}$ a une racine dans $]M(\omega,\lambda), 1[$, c'est-à-dire $M(\eta,\omega) > M(\omega,\lambda)$.

On en déduit que le rayon spectral de $\mathcal{L}_\omega$ vaut $M(\omega,\rho(J))^2$. C'est ce nombre qu'il faut rendre le plus petit possible en ajustant convenablement $\omega \in]0,2[$. Pour cela, on remarque que, tant que $\Delta_J := \omega^2\rho(J)^2 + 4(1-\omega)$ est positif, $\mu := M(\omega,\rho(J))$ satisfait :

$$(2\mu - \omega\rho(J))\frac{d\mu}{d\omega} + 1 - \mu\rho(J) = 0.$$

Comme 2μ est supérieur à la somme $\omega\rho(J)$ des racines et puisque $\mu, \rho(J) \in [0,1[$, on en déduit que $\omega \mapsto M(\omega,\rho(J))$ est décroissant sur l'intervalle de $]0,2[$ défini par $\Delta_J \geqslant 0$, qu'on note $]0,\omega_J[$. On a

$$\omega_J = 2\frac{1-\sqrt{1-\rho(J)^2}}{\rho(J)^2} = \frac{2}{1+\sqrt{1-\rho(J)^2}} \in [1,2[.$$

En effet, le trinôme Δ_J a deux racines réelles positives situées de part et d'autre de $\omega = 2$.

Sur $]\omega_J, 2[$, on a $M(\omega, \rho(J))^2 = |\omega - 1| = \omega - 1$, qui est croissant. Ainsi, $\omega \mapsto M(\omega, \rho(J))^2$, qui est continu, atteint son minimum en ω_J, ce minimum valant

$$\omega_J - 1 = \frac{1 - \sqrt{1 - \rho(J)^2}}{1 + \sqrt{1 - \rho(J)^2}}.$$

Notons aussi que $M(\omega, \rho(J))$ est toujours inférieur à un. Finalement :

Théorème 9.4.1. *On suppose que A est tridiagonale, D est inversible et J n'a que des valeurs propres réelles, toutes dans* $]-1, 1[$. *On suppose en outre que* $\omega \in \mathbb{R}$.

Alors la méthode de relaxation converge si et seulement si $\omega \in]0, 2[$. *De plus, le taux de convergence est maximal pour le paramètre*

$$\omega_J := \frac{2}{1 + \sqrt{1 - \rho(J)^2}} \in [1, 2[,$$

le rayon spectral de $\mathcal{L}_{\omega_J}$ *étant*

$$(\omega_J - 1 =) \quad \frac{1 - \sqrt{1 - \rho(J)^2}}{1 + \sqrt{1 - \rho(J)^2}} = \left(\frac{1 - \sqrt{1 - \rho(J)^2}}{\rho(J)} \right)^2.$$

L'analyse qui précède est traduite dans la figure 9.1.

Remarque. La méthode de Gauss-Seidel n'est donc pas en général la plus performante ; on n'a $\omega_J = 1$ que lorsque $\rho(J) = 0$, alors que dans la réalité, $\rho(J)$ est proche de l'unité. C'est par exemple le cas lorsqu'on résout une EDP elliptique par la méthode des éléments finis.

Pour des valeurs de $\rho(J)$ qui ne sont pas trop proches de l'unité, la méthode de relaxation avec le paramètre optimal ω_J, bien qu'elle améliore la rapidité de convergence, n'a pas une supériorité écrasante. En effet,

$$\rho(G)/\rho\left(\mathcal{L}_{\omega_J}\right) = \left(1 + \sqrt{1 - \rho(J)^2}\right)^2$$

reste compris entre 1 (pour $\rho(J)$ tendant vers 1) et 4 (pour $\rho(J) = 0$), de sorte que le rapport $\log \rho(\mathcal{L}_{\omega_J}) / \log \rho(G)$ reste assez petit. En revanche, dans le cas réaliste où $\rho(J)$ est proche de l'unité, on a

$$\log \rho(G) / \log \rho(\mathcal{L}_{\omega_J}) \sim \frac{1}{\sqrt{2}} \sqrt{1 - \rho(J)},$$

qui est très petit. Le nombre d'itérations à mener pour garantir une certaine précision est multiplié par ce rapport quand on remplace la méthode de Gauss-Seidel par la méthode de relaxation avec paramètre optimal.

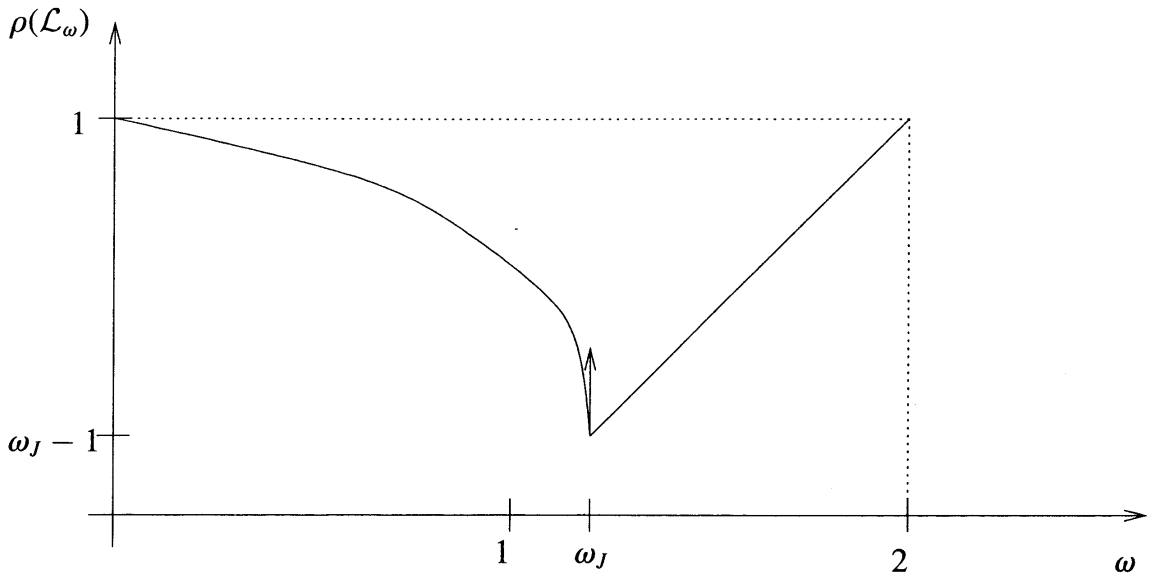

Figure 9.1 $\rho(\mathcal{L}_\omega)$ dans le cas tridiagonal.

9.5 MÉTHODE DU GRADIENT CONJUGUÉ

Nous présentons ici la méthode du *gradient conjugué* dans le cadre qui lui sied le mieux, à savoir celui des systèmes $Ax = b$ où A est réelle symétrique définie positive ($A \in SDP_n$). Comme nous le verrons, c'est une méthode *directe*, au sens où elle fournit la solution $\bar{x}$ en un nombre fini d'itérations (au plus n). Cependant, les erreurs d'arrondis polluent le résultat final et l'on préfère voir le gradient conjugué comme une méthode *itérative* dans laquelle un nombre N d'itérations, très inférieur à n, permet d'obtenir une bonne approximation de $\bar{x}$. Nous verrons que le choix de N est lié au *conditionnement* de la matrice A.

Nous notons $(\cdot,\cdot)$ le produit scalaire canonique sur $\mathbb{R}^n$. Lorsque $A \in SDP_n$ et $b \in \mathbb{R}^n$, la fonction

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

est strictement convexe et tend vers l'infini quand $\|x\| \to +\infty$. Elle atteint donc sa borne inférieure en un point unique $\bar{x}$, qui est l'unique vecteur où le gradient de J s'annule. Nous noterons r (pour *résidu*) le gradient de J : $r(x) = Ax - b$. Ainsi $\bar{x}$ est la solution du système linéaire $Ax = b$.

Plus précisemment, si $A\bar{x} = b$ et $x \in \mathbb{R}^n$, $x \neq \bar{x}$, on a

$$J(x) = J(\bar{x}) + \frac{1}{2}(A(x - \bar{x}), x - \bar{x}) > J(\bar{x}). \tag{9.3}$$

Le gradient conjugué est donc une méthode de descente. Mais elle est loin d'être naïve,

9.5.1 Analyse théorique

Soit $x_0 \in \mathbb{R}^n$. Nous notons $e_0 = x_0 - \bar{x}$, $r_0 = r(x_0) = Ae_0$. Nous pouvons supposer que $e_0 \neq 0$, sans quoi nous connaissons déjà la solution. Pour $k \geqslant 1$, définissons l'espace vectoriel

$$\mathcal{H}_k := \{P(A)r_0 \mid P \in \mathbb{R}[X], \deg P \leqslant k-1\}, \quad \mathcal{H}_0 = \{0\}.$$

Dans $\mathcal{H}_{k+1}$, $\mathcal{H}_k$ est de codimension 0 ou 1. Dans le premier cas, $\mathcal{H}_{k+1} = \mathcal{H}_k$; il s'ensuit alors que $\mathcal{H}_{k+2} = A\mathcal{H}_{k+1} + \mathcal{H}_{k+1} = A\mathcal{H}_k + \mathcal{H}_k = \mathcal{H}_{k+1} = \mathcal{H}_k$ et donc, par récurrence, $\mathcal{H}_k = \mathcal{H}_m$ pour tout $m > k$. Notons l le plus petit indice tel que $\mathcal{H}_l = \mathcal{H}_{l+1}$. Pour $k < l$, $\mathcal{H}_k$ est donc de codimension un dans $\mathcal{H}_{k+1}$, tandis que, si $k \geqslant l$, $\mathcal{H}_k = \mathcal{H}_{k+1}$. Il s'ensuit que $\dim \mathcal{H}_k = k$ si $k \leqslant l$. En particulier, $l \leqslant n$.

On peut toujours trouver, par le procédé d'orthogonalisation de Gramm-Schmidt, une base A-orthogonale[2] (c'est-à-dire telle que $(Ap_j, p_i) = 0$ si $i \neq j$) $\{p_0, \ldots, p_{l-1}\}$ de $\mathcal{H}_l$, telle que $\{p_0 \ldots, p_{k-1}\}$ soit une base de $\mathcal{H}_k$ quand $k \leqslant l$. Les vecteurs p_j, qui ne seront pas choisis unitaires *a priori*, sont définis, de manière unique à multiplication par un scalaire non nul près, par

$$p_k \in \mathcal{H}_{k+1}, \quad p_k \perp_A \mathcal{H}_k.$$

Le signe $\perp_A$ indique l'orthogonalité pour le produit scalaire défini par A. On dit que les vecteurs p_j sont deux à deux *conjugués* (sous-entendu, pour le produit scalaire induit par A). C'est l'origine du nom de la méthode.

La fonction quadratique J, strictement convexe, atteint sa borne inférieure sur l'espace affine $x_0 + \mathcal{H}_k$ en un vecteur unique, que nous notons x_k. Cette notation est cohérente pour $k = 0$. Si $x = y + \gamma p_k \in \mathcal{H}_{k+1}$ avec $y \in \mathcal{H}_k$, on a

$$\begin{aligned} J(x) &= J(\bar{x}) + E(x - \bar{x}) \\ &= J(\bar{x}) + E(y - \bar{x}) + \gamma^2 E(p_k) + \gamma(Ap_k, y - \bar{x}) \\ &= J(y) + \gamma^2 E(p_k) - \gamma(Ap_k, \bar{x}), \end{aligned}$$

puisque $(Ap_k, y) = 0$. Ainsi, la minimisation de J sur $\mathcal{H}_{k+1}$ revient à celle de J sur $\mathcal{H}_k$ et celle de $\gamma^2 E(p_k) - \gamma(Ap_k, \bar{x})$ sur $\mathbb{R}$. On a donc

$$x_{k+1} - x_k \in \mathbb{R}p_k. \tag{9.4}$$

Par définition de l, il existe un polynôme P non nul de degré l tel que $P(A)r_0 = 0$, c'est-à-dire $AP(A)e_0 = 0$. Supposons que $P(0)$ soit nul. Alors $P(X) = XQ(X)$ avec $\deg Q = l - 1$. On a donc $A^2Q(A)e_0 = 0$. Mais comme A est symétrique, donc diagonalisable, $A^2w = 0$ implique $Aw = 0$. On a donc $AQ(A)e_0 = 0$, c'est-à-dire $Q(A)r_0 = 0$. Mais cela contredit le fait que $\dim \mathcal{H}_{l-1} = l - 1$. Ainsi, $P(0) \neq 0$ et l'on peut supposer que $P(0) = 1$. Alors $P(X) = 1 - XR(X)$, où $\deg R = l - 1$. Soit $z := P(A)e_0$. On a $z = e_0 - R(A)r_0 \in e_0 + \mathcal{H}_l$. Mais $Az = P(A)r_0 = 0$ montre que $z = 0$.

2. On doit distinguer dans cette section entre les deux produits scalaires $(\cdot, \cdot)$ et $(A\cdot, \cdot)$.

Autrement dit, $e_0 \in \mathcal{H}_l$, ou encore $\bar{x} \in x_0 + \mathcal{H}_l$. Réciproquement, si $k \leqslant l$ et $\bar{x} \in x_0 + \mathcal{H}_k$, alors $e_0 \in \mathcal{H}_k$, c'est-à-dire $e_0 = Q(A)r_0$, où $\deg Q \leqslant k-1$. Alors $Q_1(A)e_0 = 0$, pour $Q_1(X) = 1 - XQ(X)$. On a donc $Q_1(A)r_0 = 0$, $Q_1(0) \neq 0$ et $\deg Q_1 \leqslant k$, donc $k \geqslant l$, c'est-à-dire $k = l$.

En résumé, $\bar{x} \in x_0 + \mathcal{H}_l$ mais $\bar{x} \notin x_0 + \mathcal{H}_{l-1}$. On a donc $x_l = \bar{x}$ et $x_k \neq \bar{x}$ si $k < l$.

Lemme 9.5.1 *Notons $\lambda_n \geqslant \cdots \geqslant \lambda_1 (> 0)$ les valeurs propres de A. Si $k \leqslant l$, on a*

$$E(x_k - \bar{x}) \leqslant E(e_0) \cdot \min_{\deg Q \leqslant k-1} \max_j |1 + \lambda_j Q(\lambda_j)|^2.$$

Démonstration.

$$\begin{aligned} E(x_k - \bar{x}) &= \min\{E(x - \bar{x}) \mid x \in x_0 + \mathcal{H}_k\} \\ &= \min\{E(e_0 + y) \mid y \in \mathcal{H}_k\} \\ &= \min\{E((I_n + AQ(A))e_0) \mid \deg Q \leqslant k-1\} \\ &= \frac{1}{2} \min\{\|(I_n + AQ(A))A^{1/2}e_0\|_2^2 \mid \deg Q \leqslant k-1\}, \end{aligned}$$

où on a utilisé l'égalité $(Aw, w) = \|A^{1/2}w\|_2^2$. Ainsi

$$\begin{aligned} E(x_k - \bar{x}) &\leqslant \frac{1}{2} \min\{\|I_n + AQ(A))\|_2^2 \|A^{1/2}e_0\|_2^2 \mid \deg Q \leqslant k-1\} \\ &= E(e_0) \min\{\rho(I_n + AQ(A))^2 \mid \deg Q \leqslant k-1\}, \end{aligned}$$

car $\rho(S) = \|S\|_2$ pour une matrice symétrique réelle. ❑

On peut déduire du lemme 9.5.1 une estimation de l'erreur $E(x_k - \bar{x})$, en majorant par

$$\min_{\deg Q \leqslant k-1} \max_{t \in [\lambda_1, \lambda_n]} |1 + tQ(t)|^2.$$

Classiquement, ce minimum est atteint pour

$$1 + XQ(X) = \omega_k T_k \left(\frac{2X - \lambda_1 - \lambda_n}{\lambda_n - \lambda_1} \right),$$

où $\omega_k = 1/T_k(0)$ et T_k est un polynôme de Tchebycheff :

$$T_k(t) = \begin{cases} \cos k \arccos t & \text{si} \quad |t| \leqslant 1, \\ \operatorname{ch} k \operatorname{argch} t & \text{si} \quad |t| \geqslant 1. \end{cases}$$

On a alors

$$\max_{[\lambda_1, \lambda_n]} |1 + tQ(t)| = |\omega_k|,$$

et

$$|\omega_k| = \frac{1}{T_k\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)} = \frac{1}{\operatorname{ch} k \operatorname{argch} \frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}}.$$

on a donc $E(x_k - \bar{x}) \leqslant |\omega_k|^2 E(e_0)$. Cependant, si

$$\theta := \operatorname{argch} \frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1},$$

on a $|\omega_k| = (\operatorname{ch} k\theta)^{-1} \leqslant 2\exp(-k\theta)$, où $\exp(-\theta)$ est la racine inférieure à un du trinôme

$$T^2 - 2\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1} T + 1.$$

Notant $K(A) := \|A\|_2 \|A^{-1}\|_2 = \lambda_n/\lambda_1$ le *conditionnement* de A, on obtient

$$e^{-\theta} = \frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1} - \sqrt{\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)^2 - 1} = \frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}} = \frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1}.$$

Le résultat final est donc

Théorème 9.5.1. *Si $k \leqslant l$, on a*

$$E(x_k - \bar{x}) \leqslant 4E(x_0 - \bar{x}) \left(\frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1}\right)^{2k}.$$

Nous notons maintenant $r_k = r(x_k) = A(x_k - \bar{x})$. Nous avons vu que $r_l = 0$ et que $r_k \neq 0$ si $k < l$. En fait r_k est le gradient de J en x_k. La minimalité de J en x_k sur $x_0 + \mathcal{H}_k$ entraîne donc que $r_k \perp \mathcal{H}_k$ (pour le produit scalaire usuel). Autrement dit, nous avons $(r_k, p_j) = 0$ si $j < k$. Cependant, $x_k - \bar{x} \in e_0 + \mathcal{H}_k$ s'écrit aussi $x_k - \bar{x} = Q(A)e_0$ avec $\deg Q \leqslant k$, donc implique $r_k = Q(A)r_0$, d'où $r_k \in \mathcal{H}_{k+1}$. Si $k < l$, on a donc $\mathcal{H}_{k+1} = \mathcal{H}_k \oplus \mathbb{R}r_k$.

Nous normalisons maintenant p_k (ce qui n'avait pas encore été fait) par

$$p_k - r_k \in \mathcal{H}_k.$$

En d'autre termes, p_k est la projection A-orthogonale de $r_k = r(x_k)$, parallèlement à $\mathcal{H}_k$. C'est bien un élément de $\mathcal{H}_{k+1}$ puisque $r_k \in \mathcal{H}_{k+1}$, non nul puisque $r_k \notin \mathcal{H}_k$. On remarquera que r_k est orthogonal à $\mathcal{H}_k$ pour le produit scalaire usuel, mais que p_k l'est pour A ; c'est pourquoi p_k et r_k sont généralement différents.

Si $j \leqslant k - 2$, nous calculons $(A(p_k - r_k), p_j) = -(Ar_k, p_j) = -(r_k, Ap_j) = 0$. Nous avons utilisé successivement la conjugaison des p_k, la symétrie de A, le fait que $Ap_j \in \mathcal{H}_{j+2}$ et l'orthogonalité de r_k et $\mathcal{H}_k$. Nous avons donc $p_k - r_k \perp_A \mathcal{H}_{k-1}$, de sorte que

$$p_k = r_k + \delta_k p_{k-1} \tag{9.5}$$

pour un nombre δ_k convenable.

9.5.2 Mise en œuvre de la méthode

Le point essentiel est la simplicité du calcul des vecteurs x_k par récurrence. Tout d'abord, $p_0 = r_0 = Ax_0 - b$. Supposons maintenant que x_k et p_{k-1} sont connus. Alors $r_k = Ax_k - b$. Si $r_k = 0$, le calcul est terminé. Sinon, les formules (9.4 et 9.5) montrent qu'en fait x_{k+1} minimise J dans le plan $x_k + \mathbb{R}r_k \oplus \mathbb{R}p_{k-1}$. On a donc $x_{k+1} = x_k + \alpha_k r_k + \beta_k p_{k-1}$, où les coefficients α_k, β_k s'obtiennent en résolvant le système linéaire de deux équations

$$\alpha_k(Ar_k, r_k) + \beta_k(Ar_k, p_{k-1}) + \|r_k\|^2 = 0\ ; \quad \alpha_k(Ar_k, p_{k-1}) + \beta_k(Ap_{k-1}, p_{k-1}) = 0$$

(on a utilisé $(r_k, p_{k-1}) = 0$). On a alors $\delta_k = \beta_k/\alpha_k$. Noter que α_k n'est pas nul, car sinon β_k le serait et r_k aussi.

En résumé, l'algorithme s'écrit :

- choisir x_0, poser $p_0 = r_0 = r(x_0) := Ax_0 - b$;
- pour $k \geqslant 0$ avec incrément de un, faire
  - calculer $r_k = r(x_k) = Ax_k - b$. Si $r_k = 0$, alors $\bar{x} = x_k$;
  - sinon, minimiser $J(x_k + \alpha r_k + \beta p_{k-1})$, d'où α_k, β_k ;
  - définir

$$p_{k+1} = r_k + (\beta_k/\alpha_k)p_{k-1}, \quad x_{k+1} = x_k + \alpha_k p_k.$$

À priori, ce calcul fournit la solution exacte $\bar{x}$ en l itérations. Cependant, l vaut n en général et chaque itération coûte $O(n^2)$ si la matrice A est pleine. Le gradient conjugué, vu comme méthode directe, est donc assez lent. On ne l'utilise donc que pour des matrices creuses, dont le nombre maximal m d'éléments non nuls par lignes est petit devant n. La complexité d'une itération est alors $O(mn)$. Mais même dans ce cas, c'est une méthode directe assez coûteuse ($O(mn^2)$ au total), alors que les méthodes itératives profitent aussi de la forme creuse de la matrice.

C'est pourquoi on préfère voir le gradient conjugué comme une méthode *itérative*, dans laquelle on ne fait qu'un petit nombre d'itération $N << n$. D'après le théorème 9.5.1, le taux de convergence satisfait

$$\tau_{GC} \leqslant -2\log\frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1}. \tag{9.6}$$

Cette méthode n'est vraiment une méthode itérative que si $n\tau_{GC} << 1$, car on peut alors choisir $N << n$. Clairement, une condition suffisante est que $K(A) << n^2$.

Application. Voyons le cas de la résolution par éléments finis du problème de Dirichlet pour le laplacien dans un ouvert borné Ω de $\mathbb{R}^d$:

$$\Delta u = f \text{ dans } \Omega, \quad u = 0 \text{ sur } \partial\Omega.$$

La matrice A est bien symétrique, reflétant la symétrie de la formulation variationnelle

$$\int_\Omega (\nabla u \cdot \nabla v + fv)\, dx = 0, \quad \forall v \in H_0^1(\Omega).$$

Si le diamètre du maillage est h avec $0 < h << 1$, et si ce maillage est régulier, le nombre de degrés de liberté (la taille de la matrice) n est de l'ordre de C/h^d, où C est une constante. La matrice est creuse avec $m = O(1)$. Chaque itération coûte donc $O(n)$ opérations. Enfin, le conditionnement de A est de l'ordre de c/h^2. Ainsi, un nombre d'itérations $N >> 1/h$ suffit, ce qui est intéressant dès que $d \geqslant 2$. Curieusement, la méthode est d'autant plus utile que d est grand, et le seuil $1/h$ est indépendant de la dimension.

La comparaison des performances des méthodes du gradient conjugué d'une part, et de la relaxation avec paramètre optimal d'autre part est intéressante. On trouve deux taux de convergence équivalents, de l'ordre de $\text{cste}_d h$ (par exemple $2\pi/n$ en dimension $d = 1$). Les deux méthodes sont donc aussi performantes l'une que l'autre à première vue.

Cependant, la relaxation n'est optimale qu'au prix du calcul de $\rho(J)$ qui est assez coûteux. De plus, une sous-estimation de $\rho(J)$ nuit gravement à l'efficacité de la relaxation, à cause de la tangente verticale de la figure 9.1.

Enfin, le théorème 9.4.1 porte sur la méthode par blocs pour une matrice tridiagonale par blocs, sa mise en œuvre demande donc l'inversion explicite des blocs diagonaux de A, ainsi que le calcul de $\rho(J)$ pour la méthode de Jacobi par blocs.

Par sa simplicité, le gradient conjugué est donc nettement préférable à la relaxation lorsque la dimension d du problème de Dirichlet est supérieure ou égale à deux.

Préconditionnement. En pratique, on améliore la performance de la méthode en *préconditionnant* la matrice A. L'idée est de remplacer le système $Ax = b$ par ${}^tBABy = {}^tBb$, où B est facile à inverser, par exemple triangulaire ou bien diagonale par (petits) blocs. Si ${}^tBB$ est assez proche de A^{-1}, alors le conditionnement de la nouvelle matrice est plus petit et le nombre d'itérations en est réduit d'autant. Rien d'étonnant à cela. Lorsque le conditionnement atteint sa borne inférieure $K = 1$, on a $A = I_n$ et la solution $\overline{x} = b$ est obtenue sans calcul. Le conditionnement le plus simple consiste à prendre $B = D^{-1/2}$. Son efficacité est visible sur le cas (trivial) où A est diagonale, car la matrice du nouveau système est I_n, le conditionnement étant donc ramené à un. Notons que la technique du préconditionnement est également utilisée avec la méthode de la relaxation, car elle permet de diminuer la valeur de $\rho(J)$, donc aussi le taux de convergence.

On verra en exercice que, si $A \in SDP_n$ est tridiagonale et si $D = dI_n$ (ce qui correspond au préconditionnement évoqué ci-dessus), le gradient conjugué est au moins aussi rapide que la relaxation avec paramètre optimal, c'est-à-dire

$$\tau_{GC} \leqslant \tau_{RL}.$$

En fait, cette inégalité est la combinaison de deux inégalités qui sont strictes en général (dont (9.6). Le gradient conjugué est donc vraiment plus performant dans ce cas.