ANALYSE NUMÉRIQUE MATRICIELLE

Exercice 13.1.1 Montrer que

1. $\|A\|_2 = \|A^*\|_2 =$ maximum des valeurs singulières de A,
2. $\|A\|_1 = \max_{1\leq j\leq n} \left(\sum_{i=1}^n |a_{ij}|\right)$,
3. $\|A\|_\infty = \max_{1\leq i\leq n} \left(\sum_{j=1}^n |a_{ij}|\right)$.

Correction.

1. Tout d'abord, on rappelle que les valeurs singulières de A sont les racines carré des valeurs propres de la matrice symétrique A^*A. Par définition, on a
$$\|A\|_2 = \left(\max_{x\in\mathbb{C}^n, x\neq 0} \frac{(Ax)^*\cdot Ax}{x^*\cdot x}\right)^{1/2}.$$
Ainsi,
$$\|A\|_2 = \left(\max_{x\in\mathbb{C}^n, x\neq 0} \frac{(A^*Ax)^*\cdot x}{x^*\cdot x}\right)^{1/2}$$
est bien le maximum des valeurs singulières de A (la matrice A^*A est symétrique, positive et diagonalisable).
On a pour tout $x\in\mathbb{C}^n$,
$$\|x\|_2 = \sup_{y\in\mathbb{C}^n, \|y\|_2\leq 1} |x\cdot y|.$$
Ainsi,
$$\|Ax\|_2 = \sup_{y\in\mathbb{C}^n, \|y\|_2\leq 1} |Ax\cdot y| = \sup_{y\in\mathbb{C}^n, \|y\|_2\leq 1} |x\cdot A^*y| \leq \|x\|_2\|A^*\|_2.$$
On en déduit que $\|A\|_2 \leq \|A^*\|_2$ et finalement $\|A\|_2 = \|A^*\|_2$.
2. $$\|A\|_1 = \max_{x\in\mathbb{C}, x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_{x\in\mathbb{C}, x\neq 0} \frac{\sum_i \left|\sum_k a_{ik}x_k\right|}{\sum_k |x_k|}.$$
Pour tout indice j, en choisissant $x_k = \delta_j^k$, on obtient
$$\|A\|_1 \geq \sum_i |a_{ij}|.$$
De plus,
$$\begin{aligned}\|A\|_1 &= \max_{x\in\mathbb{C}, x\neq 0} \frac{\sum_i \left|\sum_j a_{ij}x_j\right|}{\sum_j |x_j|} \leq \max_{x\in\mathbb{C}, x\neq 0} \frac{\sum_{i,j} |a_{ij}||x_j|}{\sum_j |x_j|} \\ &= \max_{x\in\mathbb{C}, x\neq 0} \frac{\sum_j \left(\sum_i |a_{ij}|\right)|x_j|}{\sum_j |x_j|} \leq \max_j \sum_i |a_{ij}|.\end{aligned}$$

3. On a
$$\|A\|_\infty = \max_{x\in\mathbb{C},x\neq 0}\left(\frac{\max_k\left|\sum_j a_{k,j}x_j\right|}{\max_k|x_k|}\right).$$
Soit $i\in\{1,\cdots n\}$ et $x\in\mathbb{C}^n$ telle que pour tout indice j, x_j soit égal au signe de $a_{i,j}$. On déduit de l'expression précédente que
$$\|A\|_\infty \geq \max_i\left(\sum_j |a_{i,j}|\right).$$
Réciproquement,
$$\|A\|_\infty \leq \max_{x\in\mathbb{C},x\neq 0}\left(\frac{\max_i\sum_j|a_{i,j}||x_j|}{\max_i|x_i|}\right)\leq \max_i\sum_j|a_{i,j}|.$$

Exercice 13.1.2 Soit une matrice $A\in\mathcal{M}_n(\mathbb{C})$. Vérifier que

1. $\mathrm{cond}(A)=\mathrm{cond}(A^{-1})\geq 1$, $\mathrm{cond}(\alpha A)=\mathrm{cond}(A)\ \forall\alpha\neq 0$,
2. pour une matrice quelconque, $\mathrm{cond}_2(A)=\frac{\mu_n(A)}{\mu_1(A)}$, où $\mu_1(A),\mu_n(A)$ sont respectivement la plus petite et la plus grande valeur singulière de A,
3. pour une matrice normale, $\mathrm{cond}_2(A)=\frac{|\lambda_n(A)|}{|\lambda_1(A)|}$, où $|\lambda_1(A)|,|\lambda_n(A)|$ sont respectivement la plus petite et la plus grande valeur propre en module de A,
4. pour toute matrice unitaire U, $\mathrm{cond}_2(U)=1$,
5. pour toute matrice unitaire U, $\mathrm{cond}_2(AU)=\mathrm{cond}_2(UA)=\mathrm{cond}_2(A)$.

Correction.

1.
$$\mathrm{cond}(A)=\|A\|\|A^{-1}\|=\|A^{-1}\|\|A\|=\mathrm{cond}(A^{-1}).$$
De plus d'après les propriétés élémentaires des normes subordonnées,
$$\mathrm{cond}(A)=\|A\|\|A^{-1}\|\geq\|AA^{-1}\|=\|\,\mathrm{Id}\,\|=1.$$
Enfin, $\mathrm{cond}(\alpha A)=\|\alpha A\|\|(\alpha A)^{-1}\|=|\alpha||\alpha|^{-1}\|A\|\|A^{-1}\|=\mathrm{cond}(A)$.
2. D'après l'Exercice 13.1.1, $\|A\|_2$ est la plus grande valeur singulière de A. Comme les valeurs singulières de A^{-1} sont les inverses des valeurs singulières de A, on en déduit que $\mathrm{cond}_2(A)=\frac{\mu_n(A)}{\mu_1(A)}$.
3. Pour une matrice normale (donc diagonalisable), les valeurs singulières sont les modules des valeurs propres. Ainsi, d'après le point précédent, pour toute matrice normale on a encore $\mathrm{cond}_2(A)=\frac{|\lambda_n(A)|}{|\lambda_1(A)|}$.
4. Pour une matrice unitaire, $\|U\|_2=\|U^{-1}\|_2=1$. Ainsi, $\mathrm{cond}_2(U)=1$.
5. Si U est une matrice unitaire, on a
$$(AU)(AU)^*=AUU^*A^*=AA^*\text{ et }(UA)^*(UA)=A^*A.$$

Ainsi, la plus grande valeur singulière de A est égale à la plus grande valeur singulière de UA tandis que la plus grande valeur singulière de A^* est égale à la plus grande valeur singulière de $(AU)^*$. On a donc

$$\|AU\|_2 = \|(AU)^*\|_2 = \|A^*\|_2 = \|A\|_2 = \|UA\|_2.$$

De plus, comme $(AU)^{-1}$ et $(UA)^{-1}$ sont le produit (à gauche et à droite) de A^{-1} avec la matrice unitaire U^*, on a également

$$\|(AU)^{-1}\|_2 = \|A^{-1}\|_2 = \|A\|_2 = \|(UA)^{-1}\|_2.$$

On en déduit que $\mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(A)$.

Exercice 13.1.3 Montrer que le conditionnement de la matrice de rigidité $\mathcal{K}_h$, donnée par (**6.12**) pour la méthode des éléments finis P_1 appliquée au Laplacien, est

$$\mathrm{cond}_2(\mathcal{K}_h) \approx \frac{4}{\pi^2 h^2}. \tag{13.1}$$

On montrera que les valeurs propres de $\mathcal{K}_h$ sont

$$\lambda_k = 4h^{-1}\sin^2\left(\frac{k\pi}{2(n+1)}\right) \quad 1 \le k \le n,$$

pour des vecteurs propres u^k donnés par leurs composantes

$$u_j^k = \sin\left(\frac{jk\pi}{n+1}\right) \quad 1 \le j,k \le n.$$

Correction. Dans un premier temps, on vérifie que les vecteurs u^k sont les vecteurs propres de $\mathcal{K}_h$. On a

$$\begin{aligned}
&(\mathcal{K}_h u^k)_j = h^{-1}(-u_{j-1}^k + 2u_j^k - u_{j+1}^k)\\
&= h^{-1}\left(\sin\left(\frac{(j-1)k\pi}{n+1}\right) + 2\sin\left(\frac{jk\pi}{n+1}\right) - \sin\left(\frac{(j+1)k\pi}{n+1}\right)\right)\\
&= (2ih)^{-1}\left(-e^{\frac{i(j-1)k\pi}{n+1}} + 2e^{\frac{i(j)k\pi}{n+1}} - e^{\frac{i(j-1)k\pi}{n+1}} - e^{-\frac{i(j-1)k\pi}{n+1}} + 2e^{-\frac{i(j)k\pi}{n+1}} - e^{-\frac{i(j-1)k\pi}{n+1}}\right)\\
&= (2ih)^{-1}\left(e^{\frac{ijk\pi}{n+1}} - e^{-\frac{ijk\pi}{n+1}}\right)\left(-e^{\frac{ik\pi}{n+1}} + 2 - e^{\frac{-ik\pi}{n+1}}\right)\\
&= 4h^{-1}\sin\left(\frac{jk\pi}{n+1}\right)\sin^2\left(\frac{k\pi}{2(n+1)}\right)\\
&= 4h^{-1}\sin^2\left(\frac{k\pi}{2(n+1)}\right)u_j^k.
\end{aligned}$$

La matrice $\mathcal{K}_h$ étant normale,

$$\mathrm{cond}_2(\mathcal{K}_h) = |\lambda_n(\mathcal{K}_h)|/|\lambda_1(\mathcal{K}_h)|.$$

La plus grande valeur propre de $\mathcal{K}_h$ est $4h^{-1}\sin^2(n\pi/2(n+1)) \approx 4h^{-1}$ et la plus petite $4h^{-1}\sin^2(\pi/2(n+1)) \approx 4h^{-1}(\pi/2(n+1))^2 = h\pi^2$. La matrice $\mathcal{K}_h$ étant normale, le conditionnement de $\mathcal{K}_h$ est

$$\mathrm{cond}_2(\mathcal{K}_h) \approx \frac{4h^{-1}}{h\pi^2} = \frac{4}{\pi^2 h^2}.$$

Exercice 13.1.4 Montrer que les factorisations LU et de Cholesky conservent la structure bande des matrices.

Correction. Considérons le cas de la factorisation LU. Soit A une matrice bande de demi largeur de bande p. On raisonne par récurrence afin de prouver que les matrices L et U sont également des matrices bande de demi largeur de bande p. Les composantes des matrices L et U sont déterminées en fonction des composantes de A colonnes par colonnes. Supposons que les $j-1$ premières colonnes de L et U soit de demi largeur de bande p. Les composantes de la j ème colonne de U sont définies pour $1 \leq i \leq j$ par

$$u_{i,j} = a_{i,j} - \sum_{k=1}^{i-1} l_{i,k}u_{k,j}.$$

La matrice A étant une matrice bande de demi-largeur p, on a $a_{i,j} = 0$ pour tout i tels que $j > i+p$. Par une (nouvelle) récurrence (sur i cette fois), on en déduit que $u_{i,j} = 0$ pour tout i tel que $j > i+p$. Ainsi, la $j$ème colonne de U est celle d'une matrice bande creuse de demi largeur de bande p. La $j$ème colonne de L est déterminée pour $j+1 \leq i \leq n$ par

$$l_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} l_{i,k}u_{k,j}}{u_{j,j}}.$$

D'après l'hypothèse de récurrence sur la structure bande des j premières colonnes de L, on a $l_{i,k} = 0$ dès que $i-k > p$. Ainsi, le terme de somme apparaissant dans l'expression de $l_{i,j}$ est nul dès que $i-(j-1) > p$ et en particulier dès que $i-j > p$. Ainsi, $l_{i,j} = 0$ dès que $i-j > p$ et les j premières colonnes de L on une structure de matrice bande de demi largeur p. Ceci achève la récurrence et prouve que la structure bande est conservée par la factorisation LU. La matrice B issue de la factorisation de Cholesky n'est autre que le produit de L par une matrice diagonale, si L est une matrice bande, B l'est également. La factorisation de Cholesky conserve donc également la structure bande.

Exercice 13.1.5 Montrer que, pour une matrice bande d'ordre n et de demie largeur de bande p, le compte d'opérations de la factorisation LU est $\mathcal{O}(np^2/3)$ et celui de la factorisation de Cholesky est $\mathcal{O}(np^2/6)$.

Correction. Voir les remarques du répertoire... N'y aurait-il pas une erreur d'énoncé ?

Exercice 13.1.6 Soit A une matrice hermitienne définie positive. Montrer que pour tout $\omega \in]0,2[$, la méthode de relaxation converge.

Correction. La matrice A étant hermitienne définie positive, sa diagonale D est constituée de réels strictement positif. La matrice $M = D/\omega - E$ est donc inversible et la méthode de relaxation correctement définie. De plus, d'après les Lemmes **13.1.26** et **13.1.27**, la méthode de relaxation est convergente dès que $M^* + N$ est définie positive. Or

$$M^* + N = \frac{2-\omega}{\omega} D,$$

qui est définie positive pour tout $\omega \in]0,2[$.

Exercice 13.1.7 Montrer que, pour la méthode de relaxation, on a toujours

$$\rho(M^{-1}N) \geq |1-\omega| \ , \ \forall \omega \neq 0,$$

et donc qu'elle ne peut converger que si $0 < \omega < 2$.

Correction. Le vecteur $e_1 = (1, 0, \cdots, 0)$ est un vecteur propre de $M^{-1}N$ de valeur propre $1-\omega$. Ainsi, $\rho(M^{-1}N) \geq |1-\omega|$ et la méthode de relaxation ne peut converger que pour $\omega \in]0,2[$.

Exercice 13.1.8 Soit A une matrice symétrique définie positive. Soit $(x_k)_{0 \leq k \leq n}$ la suite de solutions approchées obtenues par la méthode du gradient conjugué. On pose $r_k = b - Ax_k$ et $d_k = x_{k+1} - x_k$. Montrer que

(i) l'espace de Krylov K_k est aussi égal à

$$K_k = [r_0, ..., r_k] = [d_0, ..., d_k],$$

(ii) la suite $(r_k)_{0 \leq k \leq n-1}$ est orthogonale

$$r_k \cdot r_l = 0 \text{ pour tout } 0 \leq l < k \leq n-1,$$

(iii) la suite $(d_k)_{0 \leq k \leq n-1}$ est conjuguée par rapport à A

$$Ad_k \cdot d_l = 0 \text{ pour tout } 0 \leq l < k \leq n-1.$$

Correction.

(i) On rappelle que r_k est définie par $r_k = r_0 - Ay_k \perp K_{k-1}$, où $y_k \in K_{k-1}$. On a $Ay_k \in K_k$. Ainsi, $r_k \in K_k$ et $(r_0, \cdots, r_k)$ est une famille de K_k. Reste à montrer que cette famille est génératrice. On raisonne par récurrence. Supposons que $K_{k-1} = [r_0, \cdots, r_{k-1}]$. Si r_k n'appartient pas à K_{k-1}, on a

$$\dim([r_0, \cdots, r_k]) = \dim([r_0, \cdots, r_{k-1}]) + 1 = \dim(K_{k-1}) + 1 \geq \dim(K_k).$$

L'espace $[r_0, \cdots, r_k]$ étant inclus dans K_k et de même dimension, ils sont égaux. Reste à considérer le cas où r_k appartient à K_{k-1}. Comme r_k est orthogonal à K_{k-1}, on a dans ce cas $r_k = 0$ et $r_0 = Ay_k$. Or $y_k \in [r_0, \cdots, A^{k-1}r_0]$.

Ainsi, $r_0 \in [Ar_0, \cdots, A^k r_0]$. La famille $(r_0, \cdots A^k r_0)$ n'est pas libre et K_k est un espace de dimension strictement inférieure à k. Dans ce cas, on a

$$K_k = K_{k-1} = [r_0, \cdots, r_{k-1}] = [r_0, \cdots, r_k].$$

Comme y_k appartient à K_{k-1}, le vecteur $d_k = y_{k+1} - y_k$ appartient à K_k. Ainsi, $[d_0, \ldots, d_k]$ est un sous espace de K_k. Supposons que pour un k donné, on ait $K_{k-1} = [d_0, \ldots, d_{k-1}]$. Si y_{k+1} n'appartient pas à K_{k-1}, d_k n'appartient pas à K_{k-1} et $K_k = [d_0, \ldots, d_k]$. Dans le cas contraire (y_{k+1} appartient à K_{k-1}), on a $y_k = y_{k+1}$ et $r_{k+1} = r_k$. En particulier, r_{k+1} appartient à K_k et est orthogonal à K_k. On a donc $r_{k+1} = 0$. On en déduit que r_k est nul et que $K_k = K_{k-1}$. On a donc a nouveau $K_k = [d_0, \ldots, d_k]$.

(ii) Le vecteur r_k est orthogonal à $K_{k-1} = [r_0, \ldots, r_{k-1}]$.

(iii) On a $\langle A^{-1}r_0 - y_k, y\rangle_A = 0$ pour tout $y \in K_{k-1}$. Ainsi, $\langle y_{k+1} - y_k, y\rangle_A = 0$ pour tout $y \in K_{k-1}$. En d'autres termes,

$$\langle d_k, y\rangle_A = 0, \quad \forall y \in [d_0, \ldots, d_{k-1}].$$

Exercice 13.1.9 Si on considère la méthode du gradient conjugué comme une méthode directe, montrer que dans le cas le plus défavorable, $k_0 = n - 1$, le nombre d'opérations (multiplications seulement) pour résoudre un système linéaire est $N_{op} = n^3\,(1 + o(1))$.

Correction. A chaque itérations, on effectue de l'ordre de n^2 opérations, l'essentiel du temps étant consacré au calcul de Ap_k. Dans le cas le plus défavorable, l'algorithme converge au bout de n itérations. Dans ce cas, le nombre d'itérations est de l'ordre de n^3.

Exercice 13.1.10 On note avec un tilde $\tilde{\cdot}$ toutes les quantités associées à l'algorithme du gradient conjugué appliqué au système linéaire (**13.12**). Soit $x_k = B^{-*}\tilde{x}_k$, $r_k = B\tilde{r}_k = b - Ax_k$, et $p_k = B^{-*}\tilde{p}_k$. Montrer que l'algorithme du gradient conjugué pour (**13.12**) peut aussi s'écrire sous la forme

$$\texttt{initialisation} \quad \begin{cases} \texttt{choix initial } x_0 \\ r_0 = b - Ax_0 \\ p_0 = z_0 = C^{-1}r_0 \end{cases}$$

$$\texttt{itérations } k \geq 1 \quad \begin{cases} \alpha_{k-1} = \frac{z_{k-1}\cdot r_{k-1}}{Ap_{k-1}\cdot p_{k-1}} \\ x_k = x_{k-1} + \alpha_{k-1}p_{k-1} \\ r_k = r_{k-1} - \alpha_{k-1}Ap_{k-1} \\ z_k = C^{-1}r_k \\ \beta_{k-1} = \frac{z_k\cdot r_k}{z_{k-1}\cdot r_{k-1}} \\ p_k = z_k + \beta_{k-1}p_{k-1} \end{cases}$$

où $C = BB^*$.

Correction. L'algorithme du gradient conjugué associé à (**13.12**) consiste à calculer itérativement

$$\begin{aligned}
&\alpha_{k-1} = \frac{\|\tilde{r}_{k-1}\|^2}{\tilde{A}\tilde{p}_{k-1}\cdot\tilde{p}_{k-1}}\\
&\tilde{x}_k = \tilde{x}_{k-1} + \alpha_{k-1}\tilde{p}_{k-1}\\
&\tilde{r}_k = \tilde{r}_{k-1} - \alpha_{k-1}\tilde{A}\tilde{p}_{k-1}\\
&\beta_{k-1} = \frac{\|\tilde{r}_k\|^2}{\|\tilde{r}_{k-1}\|^2}\\
&\tilde{p}_k = \tilde{r}_k + \beta_{k-1}\tilde{p}_{k-1}.
\end{aligned}$$

En utilisant les expressions de x_k, r_k et p_k en fonction de $\tilde{x}_k$, $\tilde{r}_k$ et $\tilde{p}_k$, on obtient

$$\begin{aligned}
&\alpha_{k-1} = \frac{\|B^{-1}r_{k-1}\|^2}{Ap_{k-1}\cdot p_{k-1}} = \frac{C^{-1}r_{k-1}.r_{k-1}}{Ap_{k-1}\cdot p_{k-1}}\\
&x_k = B^{-*}\tilde{x}_k = x_{k-1} + \alpha_{k-1}p_{k-1}\\
&r_k = B\tilde{r}_k = r_{k-1} - \alpha_{k-1}Ap_{k-1}\\
&\beta_{k-1} = \frac{\|B^{-1}r_k\|^2}{\|B^{-1}r_{k-1}\|^2} = \frac{C^{-1}r_k\cdot r_k}{C^{-1}r_{k-1}\cdot r_{k-1}}\\
&p_k = B^{-*}\tilde{p}_k = C^{-1}r_k + \beta_{k-1}p_{k-1}.
\end{aligned}$$

L'algorithme du gradient préconditionné s'écrit donc bien sous la forme annoncée.

Exercice 13.1.11 Soit A la matrice d'ordre n issue de la discrétisation du Laplacien en dimension $N = 1$ avec un pas d'espace constant $h = 1/(n+1)$

$$A = h^{-1}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0\\ -1 & 2 & -1 & \ddots & \vdots\\ 0 & \ddots & \ddots & \ddots & 0\\ \vdots & \ddots & -1 & 2 & -1\\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}.$$

Montrer que pour la valeur optimale

$$\omega_{opt} = \frac{2}{1+2\sin\frac{\pi}{2n}} \simeq 2(1-\frac{\pi}{n+1})$$

le conditionnement de la matrice $C_\omega^{-1}A$ est majoré par

$$\mathrm{cond}_2(C_\omega^{-1}A) \leq \frac{1}{2} + \frac{1}{2\sin\frac{\pi}{2(n+1)}},$$

et donc que, pour n grand, on gagne un ordre en n dans la vitesse de convergence.

Correction. On note B_ω la matrice définie par

$$B_\omega = \sqrt{\frac{\omega}{2-\omega}}\left(\frac{D}{\omega} - E\right)D^{-1/2}.$$

On a $C_\omega = B_\omega B_\omega^T$. Ainsi,

$$C_\omega^{-1}A = B_\omega^{-T}B_\omega^{-1}A = B_\omega^{-T}(B_\omega^{-1}AB_\omega^{-T})B_\omega^T = B_\omega^{-T}\tilde{A}_\omega B_\omega^T,$$

où $\tilde{A}_\omega = B_\omega^{-1} A B_\omega^{-T}$. Les matrices $C_\omega^{-1}A$ et $\tilde{A}_\omega$ étant semblables, elles ont les mêmes valeurs propres et

$$\mathrm{cond}_2(C_\omega^{-1}A) = \mathrm{cond}_2(\tilde{A}_\omega) = \|\tilde{A}_\omega\|_2 \|\tilde{A}_\omega^{-1}\|_2.$$

Afin de déterminer une majoration du conditionnement, il suffit de majorer $\|\tilde{A}_\omega\|_2$ et $\|\tilde{A}_\omega^{-1}\|_2$. On a

$$\|\tilde{A}_\omega\|_2 = \max_{x\neq 0} \frac{\langle \tilde{A}_\omega x, x\rangle}{\langle x, x\rangle} = \max_{x\neq 0} \frac{\langle B_\omega^{-1} A B_\omega^{-T} x, x\rangle}{\langle x, x\rangle} = \max_{x\neq 0} \frac{\langle A B_\omega^{-T} x, B_\omega^{-T} x\rangle}{\langle x, x\rangle}.$$

En posant $y = B_\omega^{-T} x$, on en déduit que

$$\|\tilde{A}_\omega\|_2 = \max_{y\neq 0} \frac{\langle Ay, y\rangle}{\langle B_\omega^T y, B_\omega^T y\rangle} = \max_{y\neq 0} \frac{\langle Ay, y\rangle}{\langle B_\omega^{-T} B_\omega^T y, y\rangle} = \max_{y\neq 0} \frac{\langle Ay, y\rangle}{\langle C_\omega y, y\rangle}.$$

De même, on a

$$\|\tilde{A}_\omega^{-1}\|_2 = \max_{y\neq 0} \frac{\langle C_\omega y, y\rangle}{\|\tilde{A}_\omega\|_2}.$$

Ainsi,

$$\mathrm{cond}_2(C_\omega^{-1}A) = \max_{x\neq 0} \frac{\langle Ax, x\rangle}{\langle C_\omega x, x\rangle} \left(\min_{x\neq 0} \frac{\langle Ax, x\rangle}{\langle C_\omega x, x\rangle} \right)^{-1}.$$

Il reste à déterminer un encadrement

$$0 < \alpha \leq \frac{\langle Ax, x\rangle}{\langle C_\omega x, x\rangle} \leq \beta.$$

Majoration . On décompose C_ω sous la forme

$$C_\omega = A + \frac{\omega}{2-\omega} F_\omega D^{-1} F_\omega^T,$$

avec $F_\omega = \frac{\omega-1}{\omega} D - E$. Pour tout $x \neq 0$, on a

$$\frac{2-\omega}{\omega} \langle (A_\omega - C)x, x\rangle = -\langle F_\omega D^{-1} F_\omega^T x, x\rangle = -\langle D^{-1} F_\omega^T x, F_\omega^T x\rangle \leq 0,$$

puisque la matrice D^{-1} est définie positive. Il en découle que $\beta = 1$.
Minoration . On écrit cette fois $(2-\omega)C_\omega = A + aD + \omega G$ avec

$$G = E D^{-1} E^T - \frac{D}{4} \quad \text{et} \qquad a = \frac{(2-\omega)^2}{4\omega}.$$

Pour $x \neq 0$, on calcule le rapport

$$(2-\omega)\frac{\langle C_\omega x, x\rangle}{\langle Ax, x\rangle} = 1 + a\frac{\langle Dx, x\rangle}{\langle Ax, x\rangle} + \omega\frac{\langle Gx, x\rangle}{\langle Ax, x\rangle}.$$

Puisque $\langle Gx, x\rangle = -\frac{|x_1|^2}{2h}$, on a

$$(2-\omega)\frac{\langle C_\omega x, x\rangle}{\langle Ax, x\rangle} \leq 1 + a\frac{\langle Dx, x\rangle}{\langle Ax, x\rangle} = 1 + \frac{2a}{h}\frac{\|x\|^2}{\langle Ax, x\rangle} \leq 1 + \frac{2a}{h\lambda_{\min}(A)},$$

où $\lambda_{\min}(A) = 4h^{-1}\sin^2\frac{\pi}{2(n+1)}$ est la plus petite valeur propre de A. On peut donc prendre

$$\begin{aligned}\alpha &= (2-\omega)\left(1 + \frac{a}{2\sin^2\frac{\pi}{2(n+1)}}\right)^{-1} \\ &= \left(\frac{1}{2-\omega} + \frac{2-\omega}{2\omega\sin^2\frac{\pi}{2(n+1)}}\right)^{-1}.\end{aligned}$$

et

$$\mathrm{cond}_2(C_\omega^{-1}A) \leq \frac{1}{2-\omega} + \frac{2-\omega}{2\omega\sin^2\frac{\pi}{2(n+1)}}.$$

La minimisation du terme de droite par rapport à ω conduit à la valeur optimale

$$\omega_{opt} = \frac{2}{1 + 2\sin\frac{\pi}{2n}} \simeq 2(1 - \frac{\pi}{n+1})$$

et à la majoration

$$\mathrm{cond}_2(C_\omega^{-1}A) \leq \frac{1}{2} + \frac{1}{2\sin\frac{\pi}{2(n+1)}},$$