Chapitre 8

Systèmes linéaires

Les algorithmes de résolution de systèmes linéaires ($Ax = b$, $A \in \mathcal{M}_n(\mathbb{R})$), l'algèbre des matrices carrées) par des méthodes numériques directes (solution, x, trouvée en un nombre fini d'étapes, exacte : Gauss, Cholesky, $\cdots$) permettent de connaître :

— la non-dégénérescence du système linéaire (solution unique, $\det A \neq 0$)

— la solution exacte du système linéaire

avec une arithmétique exacte (précision infinie).
Sur un calculateur l'arithmétique flottante est à précision finie et par conséquent les résultats informatiques donnent des solutions approchées ou même aberrantes (méthode en pratique inutilisable pour un tel système linéaire) dans des conditions théoriques d'application mathématique (non-dégénérescence mathématique du système).
Il faut donc contrôler sur machine les résultats informatiques :

1. Le système n'est pas informatiquement dégénéré
2. Le contrôle de la validité de la solution informatique
3. L'évaluation de la précision de cette solution

1 Méthode (de substitution-élimination) de Gauss :

Soit le système d'équations linéaires :

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &=& b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &=& b_2 \\ \vdots \quad \vdots & & \vdots \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &=& b_n \end{array}$$

Si a_{11} est non nul, on peut soustraire à la i^{eme} équation $\dfrac{a_{i1}}{a_{11}}$ fois la première équation, pour $i = 2, ..., n$. On obtient un système équivalent où la variable x_1 est éliminée des $n-1$ dernières équations : on note $a_{ij}^{(1)} = a_{ij} - \dfrac{a_{i1}}{a_{11}}a_{ij}$ et on obtient :

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &=& b_1 \\ a_{22}^{(1)}x_2 + \cdots + a_{2n}^{(1)}x_n &=& b_2^{(1)} \\ \vdots \quad & & \vdots \quad \vdots \\ a_{n2}^{(1)}x_2 + \cdots + a_{nn}^{(1)}x_n &=& b_n^{(1)} \end{array}$$

Si $a_{22}^{(1)}$ est non nul, on peut soustraire à la i^{eme} équation $\dfrac{a_{i2}^{(1)}}{a_{22}^{(1)}}$ fois la deuxième équation, pour $i = 3, ..., n$. On obtient un système équivalent où la variable x_2 est éliminée des $n-2$ dernières équations.
De manière consécutive si $a_{kk}^{(k-1)}$ est non nul, on obtient un système équivalent

où la variable x_{k-1} est éliminée des $n-k$ dernières équations par les formules récurrentes :

$$(1)\ a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}$$

$$b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)}$$

$$k = 1, ..., n-1$$
$$i = k+1, ..., n$$
$$j = k+1, ..., n$$

Si $a_{kk}^{(k-1)}$, appelé pivot, est nul, il existe au moins un élément non nul dans sa colonne sous la diagonale, $a_{lk}^{(k-1)}$, avec $l > k$ (sinon $\det A = 0$ et on a plus l'unicité de la solution mathématique), on permute les lignes k et l et on applique les formules précédentes : $a_{lk}^{(k-1)}$ est le pivot.
A la $(n-1)^{eme}$ étape on obtient un système équivalent triangulaire $A^{(n-1)}x = b^{(n-1)}$. On le résout par remontée. Le nombre total (descente et remontée) d'opérations arithmétiques élémentaires (+, -, *, /) est de l'ordre de $\frac{2n^3}{3}$.
Si la matrice est obtenue à la k^{eme} étape notée $A^{(k)}$ sans échange de lignes, alors $\det A = \det A^{(k)}$ et on a pour $k = n-1$, $\det A = a_{11}a_{22}^{(1)} \cdots a_{nn}^{(n-1)}$ (le déterminant de A est le produit des pivots successifs en considérant $a_{nn}^{(n-1)}$ comme un pivot).
S'il y a échange de deux lignes, le déterminant de la nouvelle matrice change de signe. Les pivots sont non nuls si et seulement si les mineurs principaux de A sont non nuls.

1.1 Choix du pivot

Si l'élément $a_{kk}^{(k-1)}$ de la matrice $A^{(k-1)}$ est non nul, mais très petit, le facteur $\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$ peut amplifier les erreurs d'arrondi dues au calcul de $a_{kj}^{(k-1)}$ précédent.
D'où l'idée de prendre un pivot partiel qui revient à chercher avant la k^{eme} substitution, le plus grand élément en valeur absolue dans la k^{eme} colonne, $a_{lk}^{(k-1)}$ et à permuter les lignes k et l, avec $k \leq l \leq n$.
On peut aussi trouver ce pivot, dit total, dans la k^{eme} ligne et colonne comme le plus grand élément en valeur absolue, $a_{lm}^{(k-1)}$, $k \leq l \leq n$, $m \leq l \leq n$

Remarque :
Si le déterminant de A n'est pas petit par rapport à la taille des éléments de A, le choix du pivot partiel est satisfaisant, sinon les dernières étapes donnent de petits pivots ($\det A$ est le produit des pivots) et le choix du pivot partiel ne permet que d'éviter la propagation des erreurs d'arrondi dans les premières étapes, tandis que dans les dernières les erreurs d'arrondi s'amplifient.

1.2 Décomposition LU

Théorème : Soit $A \in \mathcal{M}_n(\mathbb{R})$ de mineurs principaux non nuls. Il existe une décomposition unique de $A = LU$ telle que L soit une matrice triangulaire

inférieure à diagonale unité avec sous la diagonale $l_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$ et U une matrice triangulaire supérieure (obtenue par la méthode de Gauss).
Remarque :
Si on veut obtenir informatiquement les matrices L et U on utilise l'espace mémoire de la matrice A en lui mettant la partie supérieure de U et inférieure strictement de L.
Applications :
Calcul du $\det A$, de A^{-1} par la méthode de Gauss sur n systèmes linéaires de second membre les n vecteurs de la base canonique.

2 Etude des erreurs sur les données :

On considère le système linéaire de R.S. Wilson

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$$

de solution théorique $x = (1, 1, 1, 1)^t$

Si on perturbe le second membre en $(32.1, 22.9, 33.1, 30.9)^t$ alors la solution du nouveau système est $(9.2, -12.6, 4.5, -1.1)^t$.
Une petite variation sur les éléments de b de l'ordre de 10^{-1} entraîne une variation sur les éléments de la solution x de l'ordre de 10 !
Si on perturbe la matrice du premier membre $\begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix}$
alors la solution est $(-81, 137, -34, 22)^t$.
Une petite variation sur les éléments de A de l'ordre de 10^{-1} entraîne une variation sur les éléments de la solution de l'ordre de 10^2 !
D'où le problème posé de savoir si une solution informatique peut être considérée comme une bonne approximation de la solution mathématique exacte sachant que l'on utilise une arithmétique machine à précision finie qui génère des erreurs d'arrondi sur les nombres utilisés.

3 Normes vectorielles et matricielles :

Soit E un espace vectoriel sur $\mathbb{R}$. Une norme sur E est une application de E dans $\mathbb{R}^+$: $x \to ||x||$ appelée norme de x qui vérifie :
— $\forall x, y \in E \quad ||x + y|| \leq ||x|| + ||y||$
— $\forall x \in E, \forall \lambda \in \mathbb{R} \quad ||\lambda x|| = |\lambda| ||x||$
— $||x|| = 0 \Leftrightarrow x = 0$

Exemples dans $\mathbb{R}^n$:
— $||x||_1 = \sum_{i=1}^{n} |x_i|$

— $||x||_2 = (\sum_{i=1}^{n} |x_i|^2)^{1/2} = (x^t x)^{1/2}$ est la norme euclidienne
— $||x||_\infty = \max_{i=1,\dots,n} |x_i|$ (norme du maximum)

Une norme matricielle est une application :

$$\begin{array}{c} \mathcal{M}_n(\mathbb{R}) \to \mathbb{R}^+ \\ A \to ||A|| \end{array}$$

telle que :

— $\forall A, B \in \mathcal{M}_n(\mathbb{R}) \quad ||A+B|| \leq ||A|| + ||B||$
— $\forall A, B \in \mathcal{M}_n(\mathbb{R}) \quad ||AB|| \leq ||A||||B||$
— $\forall A \in \mathcal{M}_n(\mathbb{R}) \quad ||A|| = 0 \Leftrightarrow A = 0$
— $\forall A \in \mathcal{M}_n(\mathbb{R}), \forall \lambda \in \mathbb{R} \quad ||\lambda A|| = |\lambda| \; ||A||$

C'est une norme sur l'algèbre des matrices $\mathcal{M}_n(\mathbb{R})$
Une norme matricielle induite est définie par une norme vectorielle :

$$||A|| = \sup_{x \neq 0} \frac{||Ax||}{||x||} = \sup_{||x||=1} ||Ax||$$

elle vérifie $||Ax|| \leq ||A||||x||$.
Les normes $||\;||_1$, $||\;||_2$, $||\;||_\infty$ vectorielles induisent les normes matricielles

— $||A||_1 = \max_{j=1,\dots,n} (\sum_{i=1}^{n} |a_{ij}|)$
— $||A||_\infty = \max_{i=1,\dots,n} (\sum_{j=1}^{n} |a_{ij}|)$
— $||A||_2 = (\rho(A^t A))^{1/2} = ||A^t||_2$ où $\rho(B)$ est le rayon spectral de B : le maximum des modules des valeurs propres de B.

Remarque :

$\rho(B) \leq ||B|| \; \forall B \in \mathcal{M}_n(\mathbb{R})$
Si A est symétrique $||A||_2 = \rho(A)$

L'application de $\mathcal{M}_n(\mathbb{R})$ dans $\mathbb{R}^+ : A \to ||A||_E = (Trace(A^t A))^{1/2} = (\sum_{i,j=1}^{n} |a_{ij}|^2)^{1/2}$
est une norme matricielle non induite (norme euclidienne sur l'espace vectoriel $\mathcal{M}_n(\mathbb{R})$)
Toute norme matricielle induite vérifie $||I|| = 1$ ($||I||_E = \sqrt{n}$) en notant I la matrice identité.
On a $||A||_2 \leq ||A||_E$.
Si $\rho(A) < 1$ alors la matrice $I - A$ est inversible, d'inverse la série de Neumann $\sum_{k=0}^{\infty} A^k$.

$$\rho(A) < 1 \text{ si et seulement si } \lim_{n \to \infty} A^n = 0$$

$$\rho(A) = \lim_{n \to \infty} ||A^n||^{1/n}$$

4 Conditionnement d'un système linéaire :

Il s'agit d'étudier la stabilité de la solution d'un système linéaire liée à la perturbation de la matrice ou du second membre (indépendemment de la méthode de résolution). C'est un conditionnement théorique. On prend des normes matricielles induites et compare les solutions x de $Ax = b$ et $x+\delta x$ de $A(x+\delta x) = b+\delta b$

On a $||\delta x|| \leq ||A^{-1}||\ ||\delta b||$ et $||b|| \leq ||A||\ ||x||$

d'où $\dfrac{||\delta x||}{||x||} \leq ||A^{-1}||\dfrac{||\delta b||}{||x||} \leq ||A^{-1}||||A||\dfrac{||\delta b||}{||b||}$

On mesure ainsi la sensibilité sur le second membre.
On compare les solutions x de $Ax = b$ et $x + \delta x$ de $(A + \delta A)(x + \delta x) = b$
On a $A\delta x + \delta Ax + \delta A\delta x = 0$, d'où $\delta x = -A^{-1}\delta A(x + \delta x)$

d'où $||\delta x|| \leq ||A^{-1}||\ ||\delta A||\ ||x + \delta x||$

Soit $\dfrac{||\delta x||}{||x + \delta x||} \leq ||A^{-1}||||A||\dfrac{||\delta A||}{||A||}$

On mesure ici la sensibilité sur la matrice du premier membre.
Dans les deux cas on remarque que l'erreur relative sur le résultat est majorée par l'erreur relative sur les données multipliée par le même facteur appelé conditionnement de A et noté $K(A)$.
C'est le plus petit nombre vérifiant ces inégalités.

Remarque :

- $K(\lambda A) = K(A) \quad \forall \lambda \neq 0$
- $K(A) \geq 1$ ($1 = ||I|| = ||AA^{-1}|| \leq ||A||\ ||A^{-1}|| = K(A)$), donc dans l'inégalité associée au conditionnement, $K(A)$ est un facteur d'amplification de l'erreur relative de donnée
- $K(A) = K(A^{-1})$
- Si A est symétrique et en prenant la norme $||\ ||_2$, on a $K_2(A) = \dfrac{\max\limits_{1\leq i\leq n} |\lambda_i(A)|}{\min\limits_{1\leq i\leq n} |\lambda_i(A)|}$ où $\lambda_i(A)$ sont les valeurs propres (réelles non nulles) de A.
- $K_2(A) = \sqrt{\dfrac{\max\limits_{1\leq i\leq n} |\lambda_i(A^tA)|}{\min\limits_{1\leq i\leq n} |\lambda_i(A^tA)|}}$ où $\lambda_i(A^tA)$ sont les valeurs propres (strictement positives) de A^tA.
- Si A est orthogonale ($A^tA = I$), $K_2(A) = 1$

Une matrice est mathématiquement bien conditionnée si son conditionnement est proche de 1 ($K(A)$ représente le facteur d'amplification théorique sur l'erreur relative des données). Si le conditionnement est grand alors l'erreur relative sur le résultat peut être grande à erreur relative sur les données fixées (petites).

Exemples :
1) La matrice de R.S Wilson, notée A, est symétrique et définie positive, on a $||A||_2 = \rho(A)$

$$\max|\lambda_i(A)| \approx 30.2878$$
$$\min|\lambda_i(A)| \approx 0.01015$$

d'où $K_2(A) \approx 2984$
Ainsi pour les systèmes linéaires associés précédents on obtient :
$\frac{||\delta x||}{||x||} \approx 8.2$ et $K_2(A)\frac{||\delta b||}{||b||} \approx 9.95$.
La matrice ou le système sont mal conditionnés.
Remarque :
A n'a pas de grands éléments, son déterminant est égal à 1 et sa matrice inverse est la suivante :

$$\begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}$$

2) La matrice de Hilbert, symétrique et définie positive, $H_n = (\frac{1}{(i+j-1)})$ est mal conditionnée.

n	$K_2(H_n)$
4	1.5514e4
5	4.7661e5
6	1.4951e7
7	4.7537e8
8	1.5258e10
9	4.9315e11

Son conditionnement K_2 est de l'ordre de $e^{3.5n}$
On remarque expérimentalement que les erreurs de codage (affectation des données) ont plus de répercution que l'accumulation des erreurs d'arrondi (opération dans les calculs) commises dans la méthode d'élimination à partir de l'ordre 6 ou 7 suivant les calculateurs en arithmétique flottante.
Remarque :
Son déterminant est le suivant :

$\frac{(1!2!\cdots(n-1)!)^4}{1!2!\cdots(2n-1)!} \sim 2^{-2n^2}$

Sa matrice inverse est la suivante :

$(\frac{(-1)^{i+j}(n+i-1)!(n+j-1)!}{(i+j-1)[(i-1)!(j-1)!]^2(n-i)!(n-j)!})$

Calcul différentiel

Soit les systèmes linéaires $Ax = b$ et $(A+\epsilon E)x(\epsilon) = b$ avec la matrice réelle A carrée d'ordre n inversible, le vecteur $b \neq 0$ et le réel $\epsilon \geq 0$ assez petit pour que la matrice réelle $A(\epsilon) = A + \epsilon E$ soit inversible. On a $A(0) = A$, $A^{-1}(0) = A^{-1}$ et $x(0) = x$.

1) $A(\epsilon)$ et $A^{-1}(\epsilon)$ sont des fonctions différentiables de ϵ.
La dérivée de $A(\epsilon)$ par rapport à ϵ, notée $\frac{dA(\epsilon)}{d\epsilon}$ est égale à E (la matrice réelle E est indépendante de ϵ).
La dérivée de $A^{-1}(\epsilon)$ par rapport à ϵ est égale à $-A^{-1}(\epsilon)EA^{-1}(\epsilon)$:
En dérivant l'égalité $A(\epsilon)A^{-1}(\epsilon) = Id$ par rapport à ϵ on obtient :
$\frac{d(A(\epsilon)A^{-1}(\epsilon))}{d\epsilon} = A(\epsilon)\frac{dA^{-1}(\epsilon)}{d\epsilon} + \frac{dA(\epsilon)}{d\epsilon}A^{-1}(\epsilon) = 0$ d'où $\frac{dA^{-1}(\epsilon)}{d\epsilon} = -A^{-1}(\epsilon)\frac{dA(\epsilon)}{d\epsilon}A^{-1}(\epsilon)$
soit $\frac{dA^{-1}(\epsilon)}{d\epsilon} = -A^{-1}(\epsilon)EA^{-1}(\epsilon)$

2) Le développement de Taylor de $A^{-1}(\epsilon)$ en $\epsilon = 0$ à l'ordre 1 est donné par :
$A^{-1}(\epsilon) = A^{-1} - \epsilon\frac{dA^{-1}(0)}{d\epsilon} + O(\epsilon^2) = A^{-1} - \epsilon A^{-1}EA^{-1} + O(\epsilon^2)$
On a $x(\epsilon) = A^{-1}(\epsilon)b = x - \epsilon A^{-1}Ex + O(\epsilon^2)$ d'où $x(\epsilon) - x = -\epsilon A^{-1}Ex + O(\epsilon^2)$
soit $||x(\epsilon) - x|| = ||\epsilon A^{-1}Ex|| + O(\epsilon^2) \leq \epsilon||A^{-1}||\,||E||\,||x|| + O(\epsilon^2)$
D'où $\frac{||x(\epsilon)-x||}{||x||} \leq \epsilon K(A)\frac{||E||}{||A||} + O(\epsilon^2)$.
Remarque :
A l'ordre 2 on a la majoration suivante :
$\frac{||x(\epsilon)-x||}{||x||} \leq \epsilon K(A)\frac{||E||}{||A||} + \epsilon^2(K(A))^2\frac{||E||^2}{||A||^2} + O(\epsilon^3)$.

Conditionnement et déterminant de matrice

Proposition :
On a l'égalité (1) $\frac{1}{K_2(A)} = \inf\limits_{B\in S_n} \frac{||A-B||_2}{||A||_2}$ avec S_n l'ensemble des matrices carrées d'ordre n complexes singulières (non inversibles).

démonstration :
L'égalité (1) est équivalente à l'égalité (2) $\frac{1}{||A^{-1}||_2} = \inf\limits_{B\in S_n} ||A - B||_2$.
Dans l'inf, $||A||_2$ ne dépend pas de la condition $B \in S_n$ donc on peut sortir cette norme de l'inf et par définition du conditionnement $K_2(A)$ on obtient le résultat.

2) il n'existe pas de matrice $B \in S_n$ telle que $||A - B||_2 < \frac{1}{||A^{-1}||_2}$.
Supposons qu'il existe une matrice $B \in S_n$ telle que $||A - B||_2 < \frac{1}{||A^{-1}||_2}$.
Par propriété du rayon spectral d'une matrice on a le rayon spectral, $\rho(A^{-1}(A - B)) \leq ||A^{-1}(A - B)||_2 \leq ||A^{-1}||_2||A - B||_2 < 1$ par hypothèse.
Par propriété du rayon spectral (si le rayon spectral d'une matrice, $\rho(M) < 1$ alors la matrice $I - M$ est inversible) on a en prenant $M = A^{-1}(A - B)$:

$I - A^{-1}(A - B) = A^{-1}B$ inversible donc puisque A est inversible, B l'est aussi. Or $B \in S_n$ d'où la contradiction.

Donc $||A-B||_2 \geq \frac{1}{||A^{-1}||_2}$ pour toute matrice $B \in S_n$ d'où $\inf\limits_{B\in S_n} ||A-B||_2 \geq \frac{1}{||A^{-1}||_2}$.

l'inf est un minimum.

3) Soit $u \in \mathbb{C}^n$ avec $||u||_2 = 1$ tel que $||A^{-1}||_2 = ||A^{-1}u||_2$. Un tel vecteur u existe par définition de la norme matricielle.
On prend $B_0 = A - \frac{u(A^{-1}u)^*}{||A^{-1}||_2{}^2}$.

3.1) $B_0 \in S_n$
Montrons que $A^{-1}u \neq 0$.

Par l'absurde supposons que $A^{-1}u = 0$ alors $u = 0$ or $||u||_2 = 1$ d'où la contradiction.

Montrons que $Ker(B_0) \neq \{0\}$.

On a $B_0(A^{-1}u) = B_0A^{-1}u = u - \frac{u(A^{-1}u)^*A^{-1}u}{||A^{-1}||_2^2} = u - \frac{<A^{-1}u, A^{-1}u>u}{||A^{-1}||_2^2} = u - \frac{||A^{-1}u||_2^2 u}{||A^{-1}||_2^2} = 0$ par définition de u.
d'où $A^{-1}u \in Ker(B_0)$, or $A^{-1}u \neq 0$ donc $B_0 \in S_n$.
d'où $||A - B_0||_2 \geq \frac{1}{||A^{-1}||_2}$

3.2) On a $||A - B_0||_2 = \frac{1}{||A^{-1}||_2^2}||u(A^{-1}u)^*||_2$ égale à (par définition de la norme 2 matricielle) $\frac{1}{||A^{-1}||_2^2} \max\limits_{x\neq 0} \frac{||u((A^{-1}u)^*x||_2}{||x||_2}$. Mais dans l'expression $u(A^{-1}u)^*x$, $(A^{-1}u)^*x$ est un produit hermitien égal à $< x, A^{-1}u >$ donc $\max\limits_{x\neq 0} \frac{||u((A^{-1}u)^*x||_2}{||x||_2} = \max\limits_{x\neq 0} \frac{|<x,A^{-1}u>|}{||x||_2}$ (avec $||u||_2 = 1$).

Par l'inégalité de Cauchy-Schwarz on a $\frac{|<x,A^{-1}u>|}{||x||_2}\} \leq ||A^{-1}u||_2$

donc le max est réalisé pour $x = \lambda A^{-1}u$ avec $\lambda \in \mathbb{C}^*$ (cas de l'égalité de Cauchy-Schwarz) d' où $\max\limits_{x\neq 0} \frac{|<x,A^{-1}u>|}{||x||_2} = ||A^{-1}u||_2$
On a alors $||A - B_0||_2 = \frac{||A^{-1}u||_2}{||A^{-1}||_2^2} = \frac{||A^{-1}||_2}{||A^{-1}||_2^2}$ (par définition de u) soit $\frac{1}{||A^{-1}||_2}$.

Interprétation de l'égalité (1).

L'égalité (1) permet une caractérisation des matrices mal conditionnées en norme 2. Elles sont relativement proches de matrices singulières.

Algorithme de Hager du calcul du conditionnement en norme 1 d'une matrice

Soit la sphère unité de $\mathbb{R}^n$ munie de la norme 1, $S = \{x \in \mathbb{R}^n, ||x||_1 = 1\}$. On considère l'application f de $\mathbb{R}^n$ dans $\mathbb{R}$ telle que $f(x) = ||A^{-1}x||_1$ avec A une matrice carrée réelle inversible. Le conditionnement de la matrice A en norme 1, $K_1(A)$ est égal à $||A||_1 \max\limits_{x\in S} f(x)$ avec $||A||_1 = \max\limits_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|$.

1) f atteint son maximum sur S en un vecteur de la base canonique de $\mathbb{R}^n$:

Il existe j_0 tel que $||A^{-1}||_1 = \underset{1\leq j\leq n}{max} \sum_{i=1}^n |A^{-1}|_{ij} = \sum_{i=1}^n |A^{-1}|_{ij_0} = ||(A^{-1})_{j_0}||_1$
où le vecteur $(A^{-1})_{j_0}$ est la $j_0$ème colonne de A^{-1}.
$\underset{x\in S}{max}||A^{-1}x||_1 = ||A^{-1}||_1 = ||A^{-1}e_{j_0}||_1 = f(e_{j_0})$

2) Soit $x \in \mathbb{R}^n$, on note $\tilde{x}$ la solution de $A\tilde{x} = x$ et $\bar{x}$ la solution de $A^t\bar{x} = s$, avec s vecteur signe de $\tilde{x}$ défini par $s_i = -1$ si $\tilde{x_i} < 0$, $s_i = 0$ si $\tilde{x_i} = 0$ et $s_i = 1$ si $\tilde{x_i} > 0$. On a $f(x) = \tilde{x}^t s$.

$f(x) = \sum_{i=1}^n |(A^{-1}x)_i| = \sum_{i=1}^n |\tilde{x_i}| = \sum_{i=1}^n s_i\tilde{x_i} =< \tilde{x}, s >$

3) On a pour tout $a \in \mathbb{R}^n$, on a $f(x) + \bar{x}^t(a - x) \leq f(a)$:

$\bar{x}^t(a - x) = ((A^{-1})^t s)^t(a - x) = s^t A^{-1}(a - x) =< A^{-1}a, s > - < \tilde{x}, s >$.
$f(x)+\bar{x}^t(a-x) =< A^{-1}a, s >= \sum_{j=1}^n (A^{-1}a)_j s_j \leq (\sum_{j=1}^n |(A^{-1}a)_j| = ||A^{-1}a||_1 = f(a)$.

4) S'il existe j tel que $\bar{x}_j > x^t\bar{x}$ alors $f(e_j) > f(x)$ où e_j est le jème vecteur de la base canonique :

D'après la partie 3) en prenant $a = e_j$ on a $f(e_j) - f(x) \geq \bar{x}^t(e_j - x) =< \bar{x}, ej > - < x, \bar{x} >= \bar{x}_j - < x, \bar{x} > > 0$ par hypothèse donc $f(e_j) > f(x)$.

5) On suppose que pour tout $j \quad \tilde{x_j} \neq 0$.
Pour y suffisamment proche de x, on a $f(y) = f(x) + s^t A^{-1}(y - x)$:

a) D'après la partie 2) on a $f(y) =< \tilde{y}, s' >$ avec s' le vecteur signe de $\tilde{y}$.
Pour y suffisamment proche de x et par hypothèse $\tilde{x_i} \neq 0$ pour tout $i = 1, \cdots, n$ on a $s' = sg(\tilde{y}) = sg(A^{-1}y) = sg(A^{-1}x) = sg(\tilde{x}) = s$.
On a $f(y) =< \tilde{y}, s > + < \tilde{x}, s > - < \tilde{x}, s >= f(x) + < \tilde{y} - \tilde{x}, s >= f(x) + < A^{-1}(y - x), s >= f(x) + s^t A^{-1}(y - x)$.

b) Si $||\bar{x}||_\infty \leq x^t\bar{x}$ alors x est un maximum local de f sur la sphère unité S :

$x \in S$ est un maximum local de f sur S si et seulement si pour tout y proche de x on a $s^t A^{-1}(y - x) \leq 0$ puisque d'après la partie a) $f(y) - s^t A^{-1}(y - x) = f(x)$.
Montrons que $s^t A^{-1}(y - x) \leq 0$.
On a $s^t A^{-1}(y - x) =< A^{-1}(y - x), s >=< y - x, (A^{-1})^t s >=< y - x, \bar{x} >=< y, \bar{x} > - < x, \bar{x} >\leq< y, \bar{x} > -||\bar{x}||_\infty$ par hypothèse.
On a $< y, \bar{x} >\leq | < y, \bar{x} > | \leq \sum_{i=1}^n |y_i\bar{x}_i| \leq (\sum_{i=1}^n |y_i|) \underset{1\leq i\leq n}{max} |\bar{x}_i|\}$.
D'où $< y, \bar{x} > -||\bar{x}||_\infty \leq (||y||_1 - 1)||\bar{x}||_\infty = 0$ car $y \in S$.

6) Algorithme de calcul de $K_1(A)$:

choix de $x \in S$ (par exemple $x_i = 1/n$)
calcul de $\tilde{x}, s, \bar{x}$

tant que $||\bar{x}||_\infty >< x, \bar{x} >$ faire
calcul de j tel que $|\bar{x}_j| = ||\bar{x}||_\infty$
choix de $x = e_j$
calcul de $\tilde{x}, s, \bar{x}$
si $||\bar{x}||_\infty \leq < x, \bar{x} >$ alors $K_1(A) \simeq ||A||_1 ||\tilde{x}||_1$
fin tant que

L'intérêt est de ne pas calculer $||A^{-1}||_1$.

Pour la matrice de R.S Wilson, notée A, le conditionnement théorique, $K_1(A)$, est égal à 4488 et on obtient les résultats informatiques suivants :

```
MATRICE :

  10.000000    7.000000    8.000000    7.000000
  7.000000    5.000000    6.000000    5.000000
  8.000000    6.000000    10.000000    9.000000
  7.000000    5.000000    9.000000    10.000000

CONDITIONNEMENT :
4488.000000
```

FIGURE 1 – Conditionnement $K_1(A)$ avec l'algorithme de Hager codé en C

$$A := \langle\langle 10, 7, 8, 7\rangle|\langle 7, 5, 6, 5\rangle|\langle 8, 6, 10, 9\rangle|\langle 7, 5, 9, 10\rangle\rangle$$

$$A := \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}$$

$$ConditionNumber(A, 1);$$

$$4488$$

FIGURE 2 – Conditionnement $K_1(A)$ avec un logiciel de calcul formel