UNIVERSITÉ DE POITIERS **2009-2010**
Master de Mathématiques et Applications
MATH 1Modan

Précision et erreurs de calculs

Dans un ordinateur, un nombre réel x est représenté en général par son approximation en virgule flottante :

$$x \approx \pm m \cdot b^p,$$

où b est la base de numérotation, m la mantisse et p l'exposant. Les calculs internes sont généralement effectués en base $b = 2$, même si les résultats affichés sont finalement traduits en base 10. La représentation $\pm m \cdot b^p$ n'est pas unique : par exemple, $3,1416$ en base $b = 10$ peut s'écrire $0,31416 \cdot 10^1$ ou $314,16 \cdot 10^{-2}$; pour garantir l'unicité, il faut définir une convention sur la mantisse (par exemple, virgule après le premier chiffre non nul, ou avant celui-ci,...).

Le codage d'un nombre en virgule flottante est fixé par la norme IEEE 754, norme adoptée par la majorité des ordinateurs (cf. article "Virgule flottante" de Wikipédia"), soit sur 32 bits ("simple précision"), soit sur 64 bits ("double précision"). La répartition des bits est la suivante :

	Encodage	**S**igne	**E**xposant	**M**antisse	Valeur d'un nombre
Simple p.	32 bits	1 bit	8 bits	23bits	$(-1)^S \times \mathrm{M} \times 2^{E-127}$
Double p.	64 bits	1 bit	11bits	52bits	$(-1)^S \times \mathrm{M} \times 2^{E-1023}$

Le tableau ci-dessus indique les bits représentés. Pour un nombre *normalisé* (virgule après le premier chiffre non nul), le premier bit de la mantisse est toujours 1 (qu'on ne stocke pas), et $1 \leq M < 2$, de sorte que

$$m = M = 1, a_1 a_2 a_3 \ldots a_N = 1 + \sum_{k=1}^{N} a_k 2^{-k},$$

où les chiffres $a_1, \ldots, a_N$ prennent la valeur 0 ou 1, et où N est le nombre de bits dans la mantisse. Cette écriture implique que m représente tous les réels de $[1, 2)$ valant m au dernier chiffre près, i.e. à $\pm 0.5 \cdot 2^{-N}$ près. Autrement dit, m représente un réel avec la précision (absolue) $\Delta m = 2^{-N}$ (NB : pour cette raison, on appelle parfois "précision" le nombre de N de bits utilisé pour la mantisse).

En SCILAB, les calculs sont faits en double précision (64bits=8octets ou "bytes" en anglais); la précision est de $N = 52$ bits, et elle est stockée dans la constante prédéfinie %eps=2^{-52}. La valeur de l'exposant p varie entre $E_{max} - 1023 = (2^{11} - 1) - 1023 = 1024$ et $E_{min} - 1023 = -1023$.

Pour raisonner sur la précision des calculs, un concept important est celui de *précision relative*. Ainsi, la mantisse d'un nombre normalisé est connu avec la précision relative

$$\frac{\Delta m}{m} \leq \frac{2^{-N}}{1} = 2^{-N}.$$

Si x est un réel représentable en virgule flottante par une matrice qui est un nombre normalisé, la précision relative sur x correspond à celle de sa mantisse : $\Delta x/x = \Delta m/m$.

Erreur d'arrondi sur une somme Ces erreurs d'arrondis provoquent de fait des erreurs de calcul dans les opérations élémentaires. Ainsi, supposons que les réels sont calculés avec 3 chiffres significatifs et arrondis à la décimale la plus proche. Si on additionne $x = 3,36$ et $y = 0,00245$ on trouve $x + y = 3.36245 \approx 3.36 = x$: on perd dans le calcul tous les chiffres significatifs de y. Additionnons maintenant x, y et $z = 0,00471$. Si on effectue les opérations dans le sens $(x + y) + z$, on trouve

$$x + y \approx 3.36 \text{ et } (x + y) + z \approx 3,3647 \approx 3.36;$$

si on effectue dans le sens opposé, $(z + y) + x$, on trouve

$$z + y = 0,00716 \text{ et } (z + y) + x \approx 3,36716 \approx 3.37.$$

On remarque que l'addition n'est plus associative. D'autre part, le résultat exact est $x + y + z = 3.36716$, donc la deuxième manière est meilleure parce que les petits termes s'additionnent. C'est un principe général : quand on veut sommer une suite de termes positifs, il vaut mieux le faire du plus petit au plus grand.

Phénomène de compensation Une source importante d'erreur réside dans les phénomènes de compensation, quand on additionne des termes de signes opposés. Ainsi, si les réels sont calculés avec 6 chiffres significatifs, si on a $x = 1,00004$ et $y = 1,00001$, tous deux connus avec 6 chiffres significatifs, la différence $x - y = 0,00003$ est connue avec seulement 1 chiffre significatif. La précision relative était de 10^{-5} sur x et y et elle est de l'ordre de 1 sur la différence !

Condition d'une matrice

Pour étudier la précision des calculs liés à la résolution d'un système linéaire "$Ax = b$", on introduit la notion de condition de matrice.

On rappelle que pour une matrice A à m lignes et n colones, et si on dispose de normes sur $\mathbb{R}^n$ et sur $\mathbb{R}^m$, la norme de A subordonnée est définie par

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

Sur la définition même, on voit que

$$\|Ax\| \leq \|A\| \, \|x\| \quad \forall x \in \mathbb{R}^n.$$

La norme subordonnée d'une matrice possède toutes les propriétés d'une norme. De plus, elle vérifie $\|I\| = 1$ pour la matrice identité et $\|A \cdot B\| \leq \|A\| \cdot \|B\|$.

On appelle **condition de la matrice** A (relativement à la norme matricielle considérée) le nombre

$$\kappa(A) = \|A\|\|A^{-1}\|.$$

Si on considère un système linéaire $Ax = b$, ce nombre mesure la sensibilité de la solution x du système linéaire vis-à-vis des variations sur les données A et b. On a plus précisément le résultat fondamental suivant, où $\| \cdot \|$ est une norme quelconque sur $\mathbb{R}^n$.

Théorème 1 *Soit $A \in GL_n(\mathbb{R})$, soient $b, \hat{b} \in \mathbb{R}^n$ avec $b \neq 0$, et soient x, $\hat{x}$ les solutions des systèmes linéaires $Ax = b$ et $A\hat{x} = \hat{b}$. Alors*

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \kappa(A) \frac{\|\hat{b} - b\|}{\|b\|}.$$

De plus, il existe des vecteurs b et $\hat{b} \neq b$ pour lesquels cette inégalité est une égalité

Ce résultat dit que l'erreur relative commise sur x est au plus de l'ordre de l'erreur relative sur b multiplié par $\kappa(A)$.
Preuve. Comme $\hat{x} - x = A^{-1}(\hat{b} - b)$ et $b = Ax$,

$$\|\hat{x} - x\| \leq \|A^{-1}\| \|\hat{b} - b\| \text{ et } \|b\| \leq \|A\| \|x\|,$$

d'où le résultat. Pour le cas d'égalité, il suffit de choisir $x \neq 0$ tel que $\|Ax\| = \|A\|\|x\|$ et $\hat{b} \neq b$ tel que $\|A^{-1}(\hat{b} - b)\| = \|A^{-1}\| \|\hat{b} - b\|$, ce qui est possible d'après la définition d'une norme matricielle.

Il est facile de vérifier que pour tout matrice A inversible, $\kappa(A) \geq \kappa(I) = 1$, $\kappa(\alpha A) = \kappa(A)$ pour $\alpha \neq 0$ et $\kappa(A^{-1}) = \kappa(A)$. Si A est une matrice symétrique définie positive, et que l'on choisit la norme $\| \cdot \|_2$,

$$\kappa_2(A) = \frac{\lambda_n(A)}{\lambda_1(A)},$$

où $\lambda_n(A)$ et $\lambda_1(A)$ désignent la plus grande et la plus petite valeur propre de A respectivement.

Si $U \in O(n)$, alors $\kappa_2(U) = 1$. Un exemple célèbre de matrice mal conditionnée est la matrice de Van der Monde.

Si on prend la même norme dans les deux espaces, alors,

Exercices mathématiques

---

1. Montrer que $A \mapsto tr({}^t\overline{A} \cdot A)$ est une norme sur l'espace des matrices complexes de taille $m * n$, puis montrer qu'il ne s'agit pas d'une norme subordonnée.
2. Soit $A \in M_{m,n}(\mathbb{C})$. Etablir que si on prend la même norme sur $\mathbb{C}^m$ et $\mathbb{C}^n$, on a pour $\|x\|_1 = \sum_{i=1}^n |x_i|$, on a

$$\|A\|_1 = \max_{j=1,\ldots,n} \left(\sum_{i=1}^m |a_{ij}| \right); \tag{1}$$

pour $\|x\|_2 = (\sum_{i=1}^n |x_i|^2)^{1/2}$, on a

$$\|A\|_2 = \sqrt{\text{plus grande valeur propre de } {}^t A \cdot A}; \tag{2}$$

pour $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$, on a

$$\|A\|_\infty = \max_{i=1,\ldots,m} \left(\sum_{j=1}^n |a_{ij}| \right). \tag{3}$$

3. Soient $\|\cdot\|$ une norme sur $\mathbb{C}^n$ et $A \in M_n(\mathbb{R})$. Montrer que

$$\sup_{x\in\mathbb{C}^n,\ x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbb{R}^n,\ x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

Exercices de programmation

---

1. Vérifier que $log_2(\%eps) = -52$. Puis, retrouvez cette précision à l'aide d'une soustraction (dans la ligne de commande Scilab).
2. A l'aide d'une opération ou d'un programme élémentaire en SCILAB, déterminez le plus petit réel strictement positif, et le plus grand réel représentables. Trouve-t-on le résultat attendu ? En particulier, comment déduire de ces nombres le nombre de chiffres en écriture binaire de la mantisse M et de l'exposant E?.
3. Calculez le conditionnement en norme $\|\cdot\|_2$, $\|\cdot\|_1$ et $\|\cdot\|_\infty$ de la matrice $A = \begin{pmatrix} 1 & 10 \\ 0 & 1 \end{pmatrix}$.
4. Ecrire un programme qui trace sur un graphique la condition de la matrice A_N suivante en fonction de N, pour les normes $\|\cdot\|_2$ et $\|\cdot\|_1$:

$$A_N = \frac{1}{(N+1)^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}.$$

Vérifier numériquement que $\kappa_2(A_N) \sim 4N^2/\pi^2$ quand $N \to +\infty$.
5. Ecrire un programme qui tracer le conditionnement en fonction de N pour la matrice de Van der Monde

$$V[x_0, x_1, \dots, x_n] = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}$$

pour des points équirépartis sur $[0, 1]$, $x_i = i/n$, $i = 0, 1, \dots, n$. On tracera sur un même graphique les conditions en norme $\|\cdot\|_2$, $\|\cdot\|_1$ et $\|\cdot\|_\infty$. Même question avec les points de Tchebychev $x_i = \cos\frac{(2i+1)}{2n+2}\pi$, $i = 0, 1, \dots, n$.

Références :
J.-P. Demailly, *Analyse numérique et équations différentielles* (pour la partie erreurs de calcul)
Ciarlet, *Introduction à l'analyse numérique et à l'optimisation* (pour la partie condition d'une matrice)