

# Généralités sur l'analyse numérique et le calcul scientifique

## I. Introduction :

L'analyse numérique (*numerical analysis*) est une branche des mathématiques appliquées s'intéressant au développement d'outils et de méthodes numériques pour le calcul d'approximations de solutions de problèmes de mathématiques qu'il serait difficile, voire impossible, d'obtenir par des moyens analytiques.

**Son objectif** est notamment *d'introduire des procédures calculatoires détaillées susceptibles d'être mises en œuvre par des calculateurs* (électroniques, mécaniques ou humains) et *d'analyser leurs caractéristiques et leurs performances*.

Elle possède des liens étroits avec deux disciplines à la croisée des mathématiques et de l'informatique :

- La première est l'analyse des algorithmes (*Analysis of Algorithms*), elle-même une branche de la théorie de la complexité (*Computational Complexity Theory*), qui fournit une mesure de l'efficacité d'une méthode en quantifiant le nombre d'opérations élémentaires, ou parfois la quantité de ressources informatiques (comme le temps de calcul, le besoin en mémoire...), qu'elle requiert pour la résolution d'un problème donné.
- La seconde est le calcul scientifique (*Scientific Computing*), qui consiste en l'étude de l'implémentation de méthodes numériques dans des architectures d'ordinateurs et leur application à la résolution effective de problèmes issus de la physique, de la biologie, des sciences de l'ingénieur ou encore de l'économie et de la finance.

Si l'introduction et l'utilisation de méthodes numériques précèdent de plusieurs siècles l'avènement des ordinateurs, c'est néanmoins avec l'apparition de ces outils modernes, vers la fin des années 1940 et le début des années 1950, que le calcul scientifique connut un essor sans précédent et que l'analyse numérique devint un domaine à part entière des mathématiques. La possibilité d'effectuer un grand nombre d'opérations arithmétiques très rapidement et simplement ouvrit en effet la voie au développement de nouvelles classes de méthodes nécessitant d'être rigoureusement analysées pour s'assurer de l'exactitude et de la pertinence des résultats qu'elles fournissent.

À ce titre, les travaux pionniers de Turing, avec notamment l'article [*Rounding-off errors in matrix processes 1948*] sur l'analyse des effets des erreurs d'arrondi sur la factorisation LU, et de Wilkinson, dont on peut citer l'ouvrage [*Rounding errors in algebraic processes 1994*] initialement publié en 1963, constituent deux des premiers éléments d'une longue succession de contributions sur le sujet.

### a. Différentes sources d'erreur dans une méthode numérique :

Les solutions de problèmes calculées par une méthode numérique sont affectées par des erreurs que l'on peut principalement classer en trois catégories :

- **Les erreurs d'arrondi** : Ce sont les erreurs dues au fait que la machine ne peut représenter les nombres réels qu'avec un nombre fini de chiffres (tout calculateur travaille en précision finie), c'est-à-dire dans un sous-ensemble discret du corps des réels  $\mathbb{R}$ , l'arithmétique naturelle étant alors approchée par une arithmétique de nombres à virgule flottante; A chaque opération mathématique élémentaire, il pourra y avoir une perte de chiffres significatifs. Le calculateur doit donc être vigilant quand le nombre d'opérations est très important.
- **Les erreurs sur les données** : Ce sont les erreurs imputables à une connaissance imparfaite des données du problème que l'on cherche à résoudre, comme lorsqu'elles sont issues de mesures physiques soumises à des contraintes expérimentales ou au fait que les données proviennent elle-même d'un calcul approché. Elles sont imposées, en quelque sorte, de l'extérieur et nous ne pouvons agir sur elles. Néanmoins, la manière dont elles se propagent au cours des calculs est davantage du ressort du calculateur. L'analyse de cette propagation sera étudiée en liaison avec notions de conditionnement et de stabilité.

- **Les erreurs de troncature, d'approximation ou de discrétisation** : introduites par les schémas de résolution numérique utilisés, comme le fait de tronquer le développement en série infini d'une solution analytique pour permettre son évaluation, d'arrêter d'un processus itératif dès qu'un itéré satisfait un critère donné avec une tolérance prescrite, ou encore d'approcher la solution d'une équation aux dérivées partielles en un nombre fini de points. (*calculer une intégrale à l'aide d'une somme finie, une dérivée à l'aide de différences finies ou bien la somme d'une série infinie à l'aide d'un nombre fini de ses termes ; approcher une fonction, solution d'une certaine équation fonctionnelle ou aux dérivées partielles, par une combinaison linéaire finie de fonctions élémentaires. Ce type d'erreurs est bien sûr fortement lié à la méthode employée.*)

On peut également envisager d'ajouter à cette liste les erreurs qualifiées d'« humaines », telles les erreurs de programmation, ou causées par des dysfonctionnements des machines réalisant les calculs.

## b. Rappel :

### Conversion des entiers :

Décimal vers binaire	Binaire vers décimal
<p><math>35_{10} = 100011_2</math></p>	$10110111_2 = 1x2^0 + 1x2^1 + 1x2^2 + 0x2^3 + 1x2^4 + 1x2^5 + 0x2^6 + 1x2^7$ $= 1 + 2 + 4 + 0 + 16 + 32 + 0 + 128$ $= 183_{10}$
<b>Exemples :</b>	
<p>523, 875, 42, 64, 42587</p> <p><math>523_{10} = 1000001011_2</math> ; <math>875_{10} = 1101101011_2</math></p> <p><math>42_{10} = 101010_2</math> ; <math>64_{10} = 1000000_2</math> ;</p> <p><math>42587_{10} = 1010011001011011_2</math></p>	<p>111 ; 1101 ; 1011010011101011 ; 101110</p> <p><math>111_2 = 7_{10}</math> ; <math>1101_2 = 13_{10}</math> ; <math>101110_2 = 46_{10}</math> ;</p> <p><math>1011010011101011_2 = 46315_{10}</math>.</p>
Décimal vers hexadécimal/octal	Hexadécimal/octal vers décimal
<p>Le même processus avec le binaire :</p> <p>(Divisions successives sur 16/8)</p> <p><math>10_{10} = A_{16}</math> ; <math>11_{10} = B_{16}</math> ; <math>12_{10} = C_{16}</math> ; <math>13_{10} = D_{16}</math> ; <math>14_{10} = E_{16}</math> ; <math>15_{10} = F_{16}</math>.</p> <p><math>1295_{10} = 50_F</math> ; <math>19678_{10} = 4CDE_{16}</math></p> <p><math>500_{10} = 764_8</math> ; <math>125_{10} = 175_8</math></p>	<p>Le même processus avec le binaire :</p> <p>(sommés de puissances de 16/8)</p> <p><math>7CF_{16} = 7x16^3 + 12x16^2 + 15x16^1 + 1x16^0</math></p> <p><math>= 28672 + 3072 + 240 + 1 = 31985_{10}</math>.</p> <p><math>10762_8 = 1x8^4 + 7x8^3 + 6x8^2 + 2x8^1 = 4096 + 448 + 48 + 2 = 4594_{10}</math>.</p>
Binaire vers hexadécimal/octal	Hexadécimal/octal vers binaire
<p>Regroupement de chaque 4/3 bits (droite vers la gauche) ensemble puis conversion :</p> <p><math>11010110_2 = 1101\ 0110 = D6_{16}</math></p> <p><math>10111011110001_2 = 10\ 1110\ 1111\ 0001 = 2EF1_{16}</math>.</p> <p><math>11010110_2 = 11\ 010\ 110 = 326_8</math> ;</p> <p><math>10111011110001_2 = 10\ 111\ 011\ 110\ 001 = 27361_8</math>.</p>	<p>Chaque chiffre est converti en 4/3 bits :</p> <p><math>AB7_{16} = 1010\ 1011\ 0111 = 101010110111_2</math> ;</p> <p><math>123_{16} = 0001\ 0010\ 0011 = 100100011_2</math>.</p> <p><math>5270_8 = 101\ 010\ 111\ 000 = 101010111000_2</math> ;</p> <p><math>146_8 = 001\ 100\ 110 = 1100110_2</math>.</p>

## Conversion des réels :

On applique le même principe des entiers sur la partie entière d'un nombre réel. Cependant pour la partie fractionnaire (après la virgule) on calcule comme suit :

### Binaire vers décimal :

La somme des puissances de la base (puissances négatives).

Exemple :

$1011,011_2$  :

- La partie entière  $1011_2 = 2^3 + 2^1 + 2^0 = 11_{10}$  ;
- La partie fractionnaire  $0,011_2 = 0 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} = \frac{1}{4} + \frac{1}{8} = \frac{5}{8} = 0,65_{10}$ .
- Donc  $1011,011_2 = 11,65_{10}$ .

### Décimal vers binaire :

Multiplications successives de la partie fractionnaire par 2 et on stocke la partie entière qui résulte à chaque fois (0 ou 1). Les multiplications sont appliquées uniquement sur la partie fractionnaire.

Exemples :

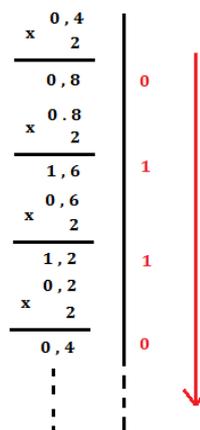
$123,04_{10} = ?_2$ .

La partie entière :  $123_{10} = 1111011_2$ .

La partie fractionnaire :

$0,4_{10} = 0,01100110\dots$

Donc :  $123,04_{10} = 1111011,011001100\dots_2$



### Exemples :

$11101,01_2 = 29,25_{10}$ .

$1,001_2 = 1,0625_{10}$ .

$11001,0\overline{1001}_2 = 25,3_{10} = 25 + \left(\frac{1}{4} + \frac{1}{32}\right) + \left(\frac{1}{64} + \frac{1}{512}\right) + \left(\frac{1}{1024} + \frac{1}{8192}\right) + \dots = 25 + \frac{9}{32} + \frac{9}{512} + \frac{9}{8192} + \frac{9}{131072} + \dots$

$= 25 + \frac{9}{2^5} + \frac{9}{2^9} + \frac{9}{2^{13}} + \frac{9}{2^{17}} + \dots = 25 + \frac{9}{2} \left[ \frac{1}{2^4} + \frac{1}{2^8} + \frac{1}{2^{12}} + \frac{1}{2^{16}} + \dots \right] = 25 + \frac{9}{2} \left[ \left(\frac{1}{2^4}\right)^1 + \left(\frac{1}{2^4}\right)^2 + \left(\frac{1}{2^4}\right)^3 + \left(\frac{1}{2^4}\right)^4 + \dots \right]$

$= 25 + \frac{9}{2} \left[ \frac{1 - \left(\frac{1}{16}\right)^n}{1 - \frac{1}{16}} - 1 \right] \quad (n \rightarrow +\infty)$

$= 25 + \frac{9}{2} \left[ \frac{16}{15} - 1 \right] = 25 + \frac{9}{2} \times \frac{1}{15} = 25 + \frac{3}{10} = 25,3$

$63,153_{10} = 111111,0010011100010001011010\dots_2$ .

$471,25_{10} = 111010111,01_2$ .

$1/9_{10} = 0,11111111\dots_{10} = 0,0001110001110\dots_2 = 0,\overline{000111}_2$

## II. Arithmétique en virgule flottante et erreur d'arrondi :

### a. Représentation des nombre dans la machine :

En informatique, les ordinateurs représentent les nombres réels sur un nombre fini de bits (représentation à virgule fixe ou à virgule flottante), ce qui ne permet la représentation exacte que d'un petit sous-ensemble des réels. Ainsi, la plupart des calculs conduisent à des résultats approchés qui résultent de la finitude de la représentation.

La virgule flottante est une méthode d'écriture de nombres réels fréquemment utilisée dans les ordinateurs. Elle consiste à représenter un nombre réel par :

- un signe (égal à  $-1$  ou  $1$ ),
- une mantisse (aussi appelée significande),
- et un exposant (entier relatif, généralement borné).

Un tel triplet représente le nombre réel :  $\text{signe} \times \text{mantisse} \times b^{\text{exposant}}$

Où  $b$  est la base de représentation (généralement 2 sur ordinateur). En faisant varier l'exposant, on fait « flotter » la virgule. La mantisse est représentée par une suite de chiffres en base  $b$ , généralement de taille fixée, dans laquelle on choisit de placer une virgule à une position fixe : juste avant ou juste après le premier chiffre, ou juste après le dernier chiffre ; dans ce dernier cas, la mantisse est un entier naturel<sup>1</sup>. Pour un nombre donné, la valeur de l'exposant dépend de ce choix.

La représentation en virgule flottante offre une étendue de nombres plus grande que celle de la représentation en virgule fixe, mais malheureusement cet avantage est au détriment de la précision.

### b. La norme IEEE754 :

Dans la norme IEEE 754, un nombre flottant est toujours représenté par un triplet ( $s, e, m$ ) :

- La première composante  $s$  détermine le signe du nombre représenté, ce signe valant 0 pour un nombre positif, et 1 pour un nombre négatif ;
- La deuxième  $e$  désigne l'exposant ;
- La troisième  $m$  désigne la mantisse.

La norme établit deux bases possibles pour la représentation :  $b = 2$  ou  $b = 10$ . Dans les deux cas, le nombre représenté par le triplet ( $s, e, m$ ) est :  $(-1)^s \times b^e \times m$ .

Ainsi, pour  $b = 10$ , les deux nombres  $31,41592 = 0,3141592 \times 10^2$  et  $-0,01732 = -0,1732 \times 10^{-1}$  sont représentés par les triplets  $(0, 2, 3141592)$  et  $(1, -1, 1732)$ .

Il reste à représenter les trois composantes ( $s, e, m$ ) dans un format binaire. Sur ce point la norme prévoit cinq formats possibles :

- Trois formats pour la représentation en base  $b = 2$  : sur 32, 64 ou 128 bits ;
- Deux formats pour la base  $b = 10$  : sur 64 et 128 bits.

#### Représentation dans la base 2 :

Un nombre réel  $x$  est représenté sur  $n$  bits dont : un bit (le plus significatif) est réservé au signe,  $w$  bits sont réservés pour coder l'exposant  $e$  et  $t$  bits sont réservés pour la mantisse  $m$ . Les valeurs des nombres  $w$  et  $t$  dépendent du format choisi, mais dans tous les cas  $n = 1 + w + t$ .

Le tableau suivant donne les valeurs des paramètres  $w$  et  $t$  selon les formats :

Paramètres	Format simple précision (32 bits)	Format double précision (64 bits)	Format quadruple précision (128 bits)
$w$	8	11	15
$t$	23	52	112

#### Le signe :

Représenté par le 1<sup>er</sup> bit le plus significatif (le plus à gauche) :

- La valeur 1 pour le signe -

- La valeur 0 pour le signe +.

**L'exposant :**

L'exposant, représenté sur  $w$  bits, peut prendre toute valeur entière comprise entre deux entiers  $e_{min}$  et  $e_{max}$ . Ces deux bornes doivent vérifier la relation  $e_{min} = 1 - e_{max}$ . De plus, si on considère l'entier naturel  $E$  codé par le mot des  $w$  bits de la représentation de  $e$ , on doit avoir  $1 \leq E \leq 2^w - 2$ .

Ces contraintes donnent les expressions suivantes pour les bornes de l'exposant  $e$  :

- $e_{max} = 2^{w-1} - 1$
- $e_{min} = 2 - 2^{w-1}$

La relation existant entre l'entier naturel  $E$  et l'exposant  $e$  est  $e = E - 2^{w-1} + 1$ .

Les valeurs possibles de l'exposant  $e$  selon le nombre de bits  $w$  de sa représentation, pour un nombre flottant ordinaire, sont données par le tableau suivant :

Formats	$e_{min}$	$e_{max}$
32 bits (w=8)	-126	+127
64 bits (w=11)	-1022	+1023
128 bits (w=15)	-16382	+16383

La représentation de l'exposant sur  $w$  bits nuls est réservée à la représentation de  $\pm 0$  et de nombres spéciaux qualifiés de « sous-normaux ». Et la représentation sur  $w$  bits à 1 est réservée aux flottants spéciaux  $\pm \infty$  et les NaN (Not a number).

**La mantisse :**

La mantisse doit être l'unique nombre réel  $m$  vérifiant :

- $1 \leq m < 2$
- $x = (-1)^s \times 2^e \times m$ :

Elle est représentée sur  $t$  bits. Dans la représentation binaire de  $m$ , le bit à gauche de la virgule est nécessairement un 1. Il est donc inutile de le représenter. Le nombre entier naturel  $M$  représenté par ces  $t$  bits possède la relation qui suit avec la mantisse  $m$  :

- $m = 1 + 2^{-t} \times M$ .

**Exemples :**

Soient  $x_1 = 29,75$  et  $x_2 = -123,025$  deux nombres réels en décimal. Donner leur représentation en binaire et en hexadécimal en virgule flottante dans les 3 formats cités précédemment.

**Format simple précision (1, w=8, t=23):**

$x_1 = 29,75 = (-1)^0 \times 2^4 \times 1,859375$

- Le signe : représenté par 0.
- L'exposant :  $e = 4$ , d'après la relation  $e = E - 2^{w-1} + 1 \rightarrow E = e + 2^{w-1} - 1 = 4 + 2^{8-1} - 1 = 4 + 2^7 - 1 = 4 + 127 = 131 = 10000011_2$ .
- La mantisse :  $m = 1,859375$ , d'après la relation  $m = 1 + 2^{-t} \times M$  on a  $M = (m-1) \times 2^t = (1,859375 - 1) \times 2^{23} = 0,859375 \times 8388608 = 7208960 = 1101110000000000000000_2$ .

Donc  $x_1 = 29,75$  est représenté au format simple précision par :

01000001111011100000000000000000 en binaire = 41EE0000 en hexadécimal.

$x_2 = -123,025 = (-1)^1 \times 2^6 \times 1,922265625$ .

- Le signe : représenté par 1.
- L'exposant :  $e = 6$ , d'après la relation  $e = E - 2^{w-1} + 1 \rightarrow E = 6 + 127 = 133 = 10000101_2$ .
- La mantisse :  $m = 1,922265625$ , d'après la relation  $m = 1 + 2^{-t} \times M$  on a  $M = (1,922265625 - 1) \times 2^{23} = 0,922265625 \times 8388608 = 7736524,8 = 11101100000110011001100, \bar{1100}_2 = 11101100000110011001100_2$  (avec une erreur de  $0, \bar{1100}$ ).

Donc  $x_1 = 29,75$  est représenté au format simple précision par :

110000101111011100000110011001100 en binaire = C2F60CCC en hexadécimal.



### c. Erreurs d'arrondis :

Une erreur d'arrondi est la différence entre la valeur approchée calculée d'un nombre et sa valeur mathématique exacte. Des erreurs d'arrondi naissent généralement lorsque des nombres exacts sont représentés dans un système incapable de les exprimer exactement. Les erreurs d'arrondi se propagent au cours des calculs avec des valeurs approchées ce qui peut augmenter l'erreur du résultat final. Dans le système décimal des erreurs d'arrondi sont engendrées, lorsqu'avec une troncature, un grand nombre (peut-être une infinité) de décimales ne sont pas prises en considération. Ce processus d'arrondi apporte des gains de temps de calcul au mépris de la précision.

Il y a au moins deux manières d'obtenir une représentation décimale limitée d'un nombre et de l'arrêter à une position donnée :

- En coupant ou en tronquant, c'est-à-dire en supprimant simplement tous les chiffres à partir d'une position donnée. Par exemple, considérons les nombres réels :

Nombres	Troncature à 4 décimales
5,2009002	5,2009
32,009891288	32,0098
-6,47009757	-6,4700

- En arrondissant, c'est-à-dire en additionnant 5 au chiffre suivant une décimale donnée, puis en coupant à partir de la décimale. Le résultat peut être arrondi par défaut ou par excès :
  - Arrondi par excès à la 5<sup>e</sup> décimale de  $(\frac{1}{7})$  : on ajoute 5 à la sixième décimale :  $0,142857+0,000005=0,142862$  et l'on tronque à la cinquième  $0,14286$  ;
  - Arrondi par défaut à la 2<sup>e</sup> décimale : on ajoute 5 à la troisième décimale :  $0,142+0,005=0,147$  et l'on tronque à la deuxième  $0,14$ .

L'analyse numérique essaie spécifiquement d'évaluer l'erreur lorsque sont utilisés des approximations de solutions d'équations ou des algorithmes numériques, plus particulièrement quand un nombre fini de chiffres est utilisé pour la représentation des nombres réels.

#### Exemple :

Nombre	Représentation	Valeur approchée	Erreur
$\frac{1}{7}$	0,142857142857... = $0, \overline{142857}$	0,142857	$\frac{1}{7} \times 10^{-6}$
Ln(2)	0,693147180559945...	0,693147	$0,180559945 \times 10^{-6}$
$\sqrt{2}$	1,414213562373095..	1,414213	$0,562373095 \times 10^{-6}$
e	2,718281828459045..	2,718281	$0,828459045 \times 10^{-6}$
$\pi$	3,141592653589793..	3,141592	$0,653589793 \times 10^{-6}$

### d. Chiffres significatifs :

Le nombre de chiffres significatifs indique la précision d'une mesure physique. Il s'agit des chiffres connus avec certitude plus le premier chiffre incertain. La précision (ou l'incertitude) avec laquelle on connaît la valeur d'une grandeur dépend du mesurage (ensemble d'opérations ayant pour but de déterminer une valeur d'une grandeur).

**Exemple :** La longueur **028,500 m** contient **5 chiffres significatifs (028,500 m s'écrit 028,500±0,001m)** :

- Le zéro le plus à gauche n'est pas significatif.

- Les chiffres 2, 8, et 5 sont significatifs.
- Les deux chiffres 0 se trouvant le plus à droite sont significatifs : ils indiquent que cette longueur est précise au millième de près.

### Comment déterminer les chiffres significatifs dans une mesure :

- Tous les chiffres (1,..,9) constituant un nombre sont considérés significatifs.
- Tous les 0 situés au milieu sont significatifs (avant ou après la virgule) : 102,03 (5 chiffres).
- Les 0 situés au plus à droite sont significatifs (même s'ils figurent après la virgule) : 1.50 (3 chiffres).
- Tous les 0 situés au plus à gauche ne sont pas significatifs : 01203 (4 chiffres) ; 0,55 (2 chiffres) ; 0,0350 (3 chiffres).
- Tous les chiffres constituant un nombre écrit selon la notation scientifique sont significatifs sauf la puissance de 10 qui n'est pas prise en compte : 2,5 $\times 10^4$  (2 chiffres).
- Les 0 à droite d'un entier puissance de 10 qui n'est pas exprimé en notation scientifique seront pris en considération.

#### Exemple :

<b>Valeurs</b>	0,10	7300	0,0073	3,5889	0,0009	0023,2	3000,05	0,0203 $\times 10^5$
<b>Nombre de chiffres significatifs</b>	2	4	2	5	1	3	6	3

### Comment déterminer les chiffres significatifs du résultat d'une opération :

**Multiplication et division :** Le résultat possède une précision équivalente à celle du terme dont la précision est la plus faible.

#### Exemples :

- $2,689$  et  $3,6 \times 10^5 = 9,6804 \times 10^5 \approx 9,7 \times 10^5$ .
- $500/100 = 5 = 5,00$ .

**Addition et soustraction :** La précision du résultat est équivalente à celle du nombre le moins précis (on se concentre sur le nombre de chiffres après la virgule, pas sur le nombre de chiffres significatifs).

#### Exemples :

- $256,3 + 1,89 = 258,19 \approx 258,2$ .

## e. Perte de chiffres significatifs :

**Exemple1 :** Supposons que dans un calcul apparaisse la quantité  $x = \pi - 3.1415$  (où  $\pi = 3.141592653589793 \dots$ ). Si on travaille avec 8 chiffres significatifs (comme beaucoup de calculettes), le nombre  $\pi$  sera représenté par :  $\pi^* = 0.31415927 \cdot 10^{+1}$  en virgule flottante normalisée. On aura donc :

$$x = 0.31415927 \cdot 10^{+1} - 0.31415 \cdot 10^{+1} = 0.0000927 = 0.927 \cdot 10^{-4}.$$

On constate que  $x$  ne contient en fait que 3 chiffres significatifs et non 8, soit une perte sèche de 5 chiffres significatifs. Ce genre de phénomènes peut surgir à tout moment d'un calcul.

**Exemple2 :** Soit à calculer le quotient

$$A = \frac{XN}{XD} = \frac{\pi - 3,1415}{10^4(\pi - 3,1415) - 0,927}$$

En travaillant avec 8 chiffres significatifs, on a :  $XD = 10^4(0,927 \cdot 10^{-4}) - 0,927 = 0,0 \rightarrow A = \text{ERREUR}$

Avec 9 chiffres on obtient :  $A = -0,1853$

Avec 10 chiffres on obtient :  $A = -0,197134$

**Exemple3 :** L'addition numérique n'est pas associative : en virgule flottante à  $n$  chiffres significatifs :

$(a + b) + c$  peut être différent de  $a + (b + c)$ . C'est le cas avec le choix suivant où les calculs sont faits avec 8 chiffres significatifs.

- $a := 0,23371258 \cdot 10^{-4}$
- $b := 0,33678429 \cdot 10^2$
- $c := -0,33677811 \cdot 10^2$

$$a + b = 0,00000023(371258) \cdot 10^2 + 0,33678429 \cdot 10^2 = 0,33678452 \cdot 10^2.$$

On remarque que, dans cette addition, les 6 derniers chiffres de a sont perdus. Ainsi :

$$(a + b) + c = 0,33678452 \cdot 10^2 - 0,33677811 \cdot 10^2 = 0,00000641 \cdot 10^2 = 0,641 \cdot 10^{-3}.$$

Par ailleurs :

$$b + c = 0,33678429 \cdot 10^2 - 0,33677811 \cdot 10^2 = 0,00000618 \cdot 10^2 = 0,618 \cdot 10^{-3}$$

$$a + (b + c) = 0,23371258 \cdot 10^{-3} + 0,61800000 \cdot 10^{-3} = 0,64137126 \cdot 10^{-3}.$$

Essayons d'analyser le phénomène ci-dessus. Si  $vf(a + b)$  désigne le résultat de l'addition  $a + b$ , on a :

$vf(a + b) = (a + b)(1 + \varepsilon_1)$ , où  $\varepsilon_1$  désigne l'erreur relative commise dans le calcul : celle-ci dépend de la précision de la machine. On sait qu'elle est majorée par  $\frac{1}{2}b^{1-n}$  soit ici  $5 \cdot 10^{-8}$ .

Posons  $\eta = vf(a + b)$ . On a alors

$$vf((a + b) + c) = vf(\eta + c) = (\eta + c)(1 + \varepsilon_2) \quad |\varepsilon_2| \leq 5 \cdot 10^{-8} = [(a + b)(1 + \varepsilon_1) + c](1 + \varepsilon_2)$$

$$= a + b + c + (a + b)\varepsilon_1(1 + \varepsilon_2) + (a + b + c)\varepsilon_2.$$

Ainsi :

$$\frac{vf((a + b) + c) - (a + b + c)}{a + b + c} = \frac{a + b}{a + b + c} \varepsilon_1(1 + \varepsilon_2) + \varepsilon_2.$$

De la même façon :

$$\frac{vf(a + (b + c)) - (a + b + c)}{a + b + c} = \frac{b + c}{a + b + c} \varepsilon_3(1 + \varepsilon_4) + \varepsilon_4.$$

On voit que les erreurs  $\varepsilon_1(1 + \varepsilon_2)$  et  $\varepsilon_3(1 + \varepsilon_4)$ , environ égales à  $5 \cdot 10^{-8}$  sont soumises à des coefficients amplificateurs :

$$\frac{a+b}{a+b+c} \simeq 5 \cdot 10^4 \text{ et } \frac{b+c}{a+b+c} \simeq 0,9$$

Ceci explique pourquoi le second calcul est plus précis que le premier.

**Remarque :** Dans les calculs où interviennent des nombres d'ordres de grandeur différents, il est en général préférable d'effectuer les opérations en groupant ceux d'ordres de grandeur similaires pour éviter les pertes de chiffres significatifs.

### III. Notions de conditionnement et de stabilité :

Ces deux notions, toujours présentes en analyse numérique, sont relatives à la propagation plus ou moins importante des erreurs d'arrondi dans un calcul donné. Nous les étudions ici pour le calcul d'une fonction.

$$x \in \mathbb{R} \rightarrow f(x) \in \mathbb{R}.$$

#### a. Conditionnement :

Le conditionnement décrit la sensibilité de la valeur d'une fonction à une petite variation de son argument, c'est-à-dire :

$$\frac{f(x)-f(x^*)}{f(x)} \text{ en fonction de } \frac{x-x^*}{x} \text{ lorsque } x-x^* \text{ est petit.}$$

Pour une fonction suffisamment régulière, on a évidemment :

$$\left| \frac{f(x)-f(x^*)}{f(x)} / \frac{x-x^*}{x} \right| \simeq \left| \frac{xf'(x)}{f(x)} \right|$$

**Définition :**

On appelle conditionnement d'une fonction  $f$  en un point  $x$  le nombre :

$$\text{cond}(f) = \left| \frac{xf'(x)}{f(x)} \right|$$

**Exemple 1 :**  $f(x) = \sqrt{x} \rightarrow \text{cond}(f) = \frac{1}{2}$ . Cela signifie que l'erreur relative sur  $f$  sera au plus la moitié de l'erreur sur relative sur  $x$  (c'est un bon conditionnement).

**Exemple 2 :**  $f(x) = a - x \rightarrow \text{cond}(f) = \left| \frac{x}{a-x} \right|$ . Un conditionnement très mauvais si  $x$  est voisin de  $a$ .

#### b. Stabilité :

La stabilité décrit la sensibilité d'un algorithme numérique pour le calcul d'une fonction  $f(x)$ .

Exemple :

$$f(x) = \sqrt{x+1} - \sqrt{x}$$

Le conditionnement :

$$\text{cond}(f) = \left| \frac{xf'(x)}{f(x)} \right| = \left| \frac{x(\sqrt{x}-\sqrt{x+1})}{2\sqrt{x(x+1)}} - \frac{1}{\sqrt{x+1}-\sqrt{x}} \right| = \frac{1}{2} \sqrt{\frac{x}{x+1}}$$

Cette dernière expression étant proche de  $\frac{1}{2}$  pour  $x$  grand. Donc, si  $x$  est grand, le conditionnement de  $f$  est bon. Cependant, dans un calcul à 6 chiffres significatifs, on a :

$$f(12345) = \sqrt{12346} - \sqrt{12345} = 111,113 - 111,108 = 0,50000 \times 10^{-2}.$$

Tandis qu'un calcul plus précis donne :  $f(12345) = 0,4500032 \dots \times 10^{-2}$ .

On a donc une erreur de 10% ce qui est important et peu en accord avec le bon conditionnement de  $f$ . Ceci est dû à l'algorithme utilisé dans ce calcul que l'on peut expliciter comme suit :

- $x_0 = 12345$
- $x_1 = x_0 + 1$
- $x_2 = \sqrt{x_1}$
- $x_3 = \sqrt{x_0}$
- $x_4 = x_2 - x_3$

Il y a quatre fonctions à intervenir et, a priori, même si le conditionnement de  $f$  est bon, il se peut que le conditionnement d'une ou plusieurs fonctions utilisées dans l'algorithme soit supérieur à celui de  $f$ .

En conclusion, le choix d'un bon algorithme numérique est essentiel. Par exemple, ci-dessus, un meilleur algorithme est obtenu en utilisant :

$$f(x) = \sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}$$

Dans ce cas, toujours avec 6 chiffres significatifs :  $f(12345) = 0,450002 \times 10^{-2}$  ce qui donne une erreur relative de 0,0003%.

Exemple d'instabilité :

Le calcul de  $e^{-12}$  à l'aide de sa série de Taylor :  $e^x = \sum_{n=0}^N \frac{x^n}{n!}$

Voici les valeurs successives de cette somme en fonction de N pour  $x = -12$ . Le calcul est fait avec 10 chiffres significatifs :

N	S <sub>N</sub>	N	S <sub>N</sub>	N	S <sub>N</sub>
2	-11,00..	19	1629,87..	36	-0,001432..
3	61,0..	20	-996,45..	37	0,000472..
4	-227,0..	21	579,34..	38	-0,0001454..
5	637,0..	22	-321,11..	39	0,000049726..
6	-1436,6..	23	170,04..	40	-0,000010319..
7	2710,6..	24	-86,20..	41	0,000007694..
8	-4398,88..	25	41,91..	42	0,000002422..
9	6265,34..	26	-19,58..	43	0,000003928..
10	-7953,62	27	8,80..	44	0,000003508..
11	9109,137..	28	-3,8130..	45	0,000003623..
12	-9504,78..	29	1,5937..	46	0,000003592..
13	9109,13..	30	-0,6435	47	0,000003600..
14	-8072,94..	31	0,2513	48	0,000003598..
15	6654,55..	32	-0,0950	49	0,000003599..
16	-5127,44..	33	0,0348..	50	0,000003598..
17	3709,05..	34	-0,01238..		
18	-2528,47..	35	0,004283..		

La valeur de  $e^{-12}$  est en fait 0,0000061442... . La formule  $e^{-12} = \frac{1}{e^{12}}$  donne lieu à un calcul plus stable même si  $e^x$  est calculé comme ci-dessus.

## IV. Conclusion :

Afin de limiter la propagation des erreurs d'arrondi, il faut essayer d'anticiper en utilisant des algorithmes dont la stabilité est optimisée par un choix d'opérations intermédiaires à bon conditionnement.

Les phénomènes soulignés dans ce chapitre ne pouvant être, en tout état de cause, complètement éliminés, on peut essayer d'évaluer l'erreur totale à laquelle un algorithme est susceptible de donner lieu :

- En faisant un calcul en double précision et en confrontant le résultat au même calcul fait en simple précision. C'est ce que nous avons fait dans l'exemple 1.2. Cependant, cette technique est très coûteuse en temps machine puisqu'elle peut multiplier le temps de calcul par un facteur 8.
- En faisant une analyse mathématique de l'erreur : ce peut être une analyse rétrograde de l'erreur comme celle utilisée plus haut pour comparer  $(a + b) + c$  et  $a + (b + c)$ . Des méthodes statistiques peuvent être également utilisées. Nous renvoyons à la littérature spécialisée pour ces questions le plus souvent délicates.

## Références bibliographiques :

- Jean-Michel Muller " Elementary Functions – Algorithms and implementation" 3rd Edition Birkhauser 2016.
- Vincent Lefèvre & Paul Zimmermann " Arithmétique flottante" rapport de recherche INRIA 2004.
- Guillaume Legendre "Introduction à l'analyse numérique et au calcul scientifique" Cours de Méthodes numériques - Université Paris Dauphine 2010.
- Takeo Takahashi "Analyse numérique" Cours électif CE33 Ecole des Mines de Nancy 2014.

## Annexe :

Voici quelques exercices pour s'entraîner :

### Exercice 1 :

1. Convertir les nombres suivants du décimal vers le binaire, puis vers le hexadécimal :

a	b	c	d	e	f
823	412,5	325,45	$\frac{1}{7}$	0,04	$\frac{16}{3}$

2. Convertir les nombres suivants du binaire vers le décimal, puis vers le hexadécimal :

a	b	c	d	e	f
10101	11011,1001	0,010110	1110110,1101	0,001101	101010111100

3. Donner la représentation en virgule flottante en binaire en simple et en double précision des nombres donnés dans la 1<sup>ère</sup> question.

### Exercice 2 :

1. Pour chaque valeur de x suivante donner sa valeur approchée en appliquant la troncature, l'arrondi, l'arrondi par excès, puis l'arrondi par défaut : au dixième, au centième ; au millième puis à la 4<sup>ème</sup> décimale.
2. Calculer l'erreur absolue  $\Delta x$  et l'erreur relative  $\rho$  pour les valeurs de x et ses valeurs approchées :

x	$2 \times 10^{-1}$	$\frac{10}{3}$	$\frac{1}{7}$	3,1415926535...	$\sqrt{2}$
$\hat{x}$	0,199	3,34	0,142 857	3,1416	1,4142

### Exercice 3 :

1. Quel est le nombre de chiffres significatifs pour les valeurs de x suivantes :

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
3,14	100,010	012,5	0,0014	$0,51 \times 10^{-5}$

2. Calculer les résultats des opérations suivantes en utilisant le nombre de chiffres significatifs qui convient (et en appliquant la troncature puis l'arrondi s'il le faut) :

$$y_1 = 3,141 \times 2,50$$

$$y_4 = y_1 \times y_2 - (y_2 - y_1) \times y_3$$

$$y_2 = 0,040 \times 125$$

$$y_5 = \frac{y_3 - y_1}{y_2 \times y_4} + 0,5$$

$$y_3 = 12,1 + 23,25 - 1,44$$

$$y_6 = y_1^2 - 0,78 \times 10^{-1}$$

3. Soit une surface rectangulaire de longueur  $L = 15,6\text{m} \pm 5\text{cm}$  et une largeur de  $7\text{m} \pm 0,01\text{m}$ . Calculer sa surface S et son périmètre P.

- Calculer la surface  $S$  et le volume  $V$  de la terre sachant que le rayon de la terre est  $R = 6,370 \times 10^3 \text{ km}$ , la valeur de  $\pi = 3,1415926535 \dots$ , la surface d'une sphère est  $S = 4\pi R^2$ , et son volume est  $V = \frac{4\pi R^3}{3}$ .
- Calculer l'erreur absolue  $\Delta$  et l'erreur relative  $\rho$  pour les calculs faits dans les deux dernières questions si la valeur approchée du rayon de la terre est  $R = 6,370 \times 10^3 \text{ km} \pm 10 \text{ km}$ .

#### Exercice4 :

- Estimer l'erreur faite dans le calcul de  $f(x) = \cos(x)e^{10x^2}$  autour de  $x = 2$  quand l'erreur sur  $x$  est d'ordre  $10^{-6}$ .
- Calculer en arrondissant à quatre chiffres significatifs chaque calcul intermédiaire, la somme des nombres suivants dans le sens croissant puis dans le sens décroissant :  
 $0,1580 \quad 0,2653 \quad 0,2581 \cdot 10^1 \quad 0,6266 \cdot 10^2 \quad 0,7889 \cdot 10^3 \quad 0,8999 \cdot 10^4$   
 Comparer les résultats.
- Soit la suite donnée ci-dessous. Calculer les valeurs successives avec 4 décimales puis avec 8 décimales et comparer les résultats (une différence importante apparaît dès la 5<sup>ème</sup> étape).

$$S = \begin{cases} X_0 = 1 \\ X_1 = \frac{1}{6} \\ X_{n+1} = \frac{37}{6}X_n - X_{n-1} \end{cases}$$

#### Exercice5 :

- Soit  $N$  une notation scientifique avec 2 chiffres significatifs en décimal (dans laquelle un nombre réel  $X$  est écrit sous la forme  $X = X_1 X_2 \cdot 10^n$  avec  $X_1 \geq 1$ ). Donner la représentation des nombres suivants dans cette notation en utilisant la troncation :  
 $0,0452 \quad 251 \quad 0,152 \cdot 10^2 \quad 0,6 \cdot 10^{-1}$
- Supposons que le nombre  $n$  qui détermine la puissance dans cette notation  $N$  est limité entre  $-2$  et  $3$  ( $n \in [-2, 3]$ ). Quels est le plus petit nombre et le plus grand nombre qu'on peut représenter ainsi?
- Selon la valeur de  $n$  déterminer le pas de cette représentation ?
- Quelle sera l'erreur absolue  $\Delta$  et l'erreur relative  $\rho$  lors de la représentation des nombres précédents (question 1).
- Déduire les valeurs min et max pour  $\Delta$  et  $\rho$  ( $\Delta_{\min}, \rho_{\min}, \Delta_{\max}, \rho_{\max}$ ).
- Représenter la variation des valeurs de  $\Delta$  et  $\rho$  par rapport à la valeur de  $X$  sur une feuille millimétrique.
- Que peut-on remarquer ?

Remarque : utiliser 2 graphes : dans le 1<sup>er</sup> représenter les valeurs de  $X \in [1,0 \times 10^{-2}, 9,9 \times 10^{-2}]$  avec une échelle de  $0,01 \times 10^{-2} \rightarrow 1 \text{ mm}$  et dans le 2<sup>ème</sup> les valeurs  $X \in [1,0 \times 10^3, 9,9 \times 10^3]$  avec une échelle de  $10 \rightarrow 1 \text{ mm}$ .