

# ANALYSE NUMÉRIQUE

Prof. HAMRI Nasr-eddine  
Département de Mathématiques  
Université Abdelhafid Boussouf - Mila





# TABLE DES MATIÈRES

<b>1</b>	<b>NOTIONS D'ERREURS</b>	<b>7</b>
1	PRÉLIMINAIRES . . . . .	8
1.1	Exemples . . . . .	9
2	Erreurs absolues et Erreurs relatives . . . . .	9
2.1	Exemple . . . . .	10
2.2	Exemples . . . . .	10
3	PRINCIPALES SOURCES D'ERREURS . . . . .	11
4	PRECISION, CHIFFRES SIGNIFICATIFS . . . . .	11
4.1	Chiffres significatifs . . . . .	11
5	Cumulation des erreurs d'arrondi . . . . .	12
5.1	Erreurs d'arrondi sur une somme . . . . .	12
5.2	Erreurs d'arrondi sur un produit . . . . .	13
6	Représentation approchée des nombres réels . . . . .	13
6.1	Nombres en virgule flottante . . . . .	14
6.2	Non-associativité des opérations arithmétiques. . . . .	14
6.3	Phénomènes de compensation. . . . .	15
7	SERIE D'EXERCICES . . . . .	15
7	EXERCICES . . . . .	16
<b>2</b>	<b>APPROXIMATION</b>	<b>19</b>
1	GÉNÉRALITÉS . . . . .	20
2	APPROXIMATION . . . . .	20
2.1	Meilleure approximation . . . . .	20
3	APPROXIMATION AU SENS DES MOINDRES CARRÉS . . . . .	21
4	CARACTÉRISATION . . . . .	22
4.1	Norme . . . . .	23
<b>3</b>	<b>INTERPOLATION POLYNOMIALE</b>	<b>25</b>
1	GÉNÉRALITÉS . . . . .	26
2	POLYNOME DE LAGRANGE . . . . .	27
2.1	Cas où les points sont equidistants . . . . .	29
3	Estimation de l'erreur dans l'interpolation de Lagrange . . . . .	30
4	POLYNOME DE NEWTON . . . . .	32
4.1	Différences finies . . . . .	32
4.2	Différences divisées . . . . .	33
5	Polynôme d'interpolation de Newton : . . . . .	35
5.1	Erreur d'interpolation . . . . .	36
5.2	Autre écriture du polynôme d'interpolation de Newton . . . . .	36
6	INTERPOLATION CUBIQUE DE HERMITE . . . . .	37

<b>4</b>	<b>INTEGRATION ET DÉRIVATION NUMÉRIQUE</b>	<b>39</b>
1	INTÉGRATION NUMÉRIQUE . . . . .	40
1.1	Méthode Générale . . . . .	40
1.2	Approximation d'une intégrale . . . . .	40
1.3	Utilisation de l'interpolation polynomiale . . . . .	41
1.4	Etude de l'erreur d'intégration . . . . .	42
1.5	Convergence des méthodes d'intégration . . . . .	42
1.6	Formules de Newton Cotes . . . . .	44
1.7	Formule de type fermé : des trapèzes et de Simpson . . . . .	44
1.8	Formule de type ouvert : . . . . .	45
1.9	Intégration par la méthode de Gauss . . . . .	45
1.10	Calcul de $\int_a^b f(x)dx$ . . . . .	47
1.11	Erreur de l'intégration par la méthode de Gauss . . . . .	47
2	DÉRIVATION NUMÉRIQUE . . . . .	48
2.1	Généralités : . . . . .	48
2.2	Utilisation de l'interpolation polynomiale . . . . .	50
2.3	Erreur de dérivation . . . . .	50
2.4	Algorithmes de dérivation . . . . .	54
2.5	Formules centrales de dérivation . . . . .	56
2.6	Formules non centrales de dérivation . . . . .	56
<b>5</b>	<b>RESOLUTION D'UN SYSTEME LINEAIRE</b>	<b>57</b>
1	METHODES DIRECTES . . . . .	57
1.1	Rappel . . . . .	57
1.2	Systèmes linéaires . . . . .	57
1.3	Résolution d'un système triangulaire supérieur . . . . .	58
2	Méthode de Gauss . . . . .	59
2.1	Interprétation matricielle de la méthode de Gauss . . . . .	60
3	Méthodes LU . . . . .	61
3.1	Décomposition LU . . . . .	61
4	Méthode de Cholesky . . . . .	62
4.1	Factorisation de Cholesky . . . . .	63
4.2	Algorithme de décomposition de Cholesky . . . . .	64
5	METHODES INDIRECTES . . . . .	66
5.1	Les méthodes itératives . . . . .	66
5.2	Différentes décomposition de A . . . . .	67
5.3	Méthode de Jacobi . . . . .	67
5.4	Méthode de Gauss-Seidel . . . . .	67
5.5	Méthode de relaxation . . . . .	68
6	Convergence des méthodes itératives . . . . .	68
6.1	Cas général . . . . .	68
<b>6</b>	<b>CALCUL DES VALEURS PROPRES ET VECTEURS PROPRES</b>	<b>71</b>
1	Introduction . . . . .	72
2	RAPPELS . . . . .	72
3	Calcul direct de $\det(A - \lambda I)$ . . . . .	72
4	Méthode de Krylov . . . . .	72
5	MÉTHODE DE LEVERRIER . . . . .	74
6	Valeurs et Vecteurs Propres . . . . .	75

7	La condition du calcul des valeurs propres . . . . .	75
7.1	Condition du calcul des vecteurs propres . . . . .	77
8	La méthode de la puissance . . . . .	78
9	Méthode de la puissance inverse de Wielandt . . . . .	79
10	VALEURS PROPRES ET VECTEURS PROPRES . . . . .	80
11	LA CONDITION DU CALCUL DES VALEURS PROPRES . . . . .	81
11.1	Condition du calcul des vecteurs propres . . . . .	83
12	LA METHODE DE LA PUISSANCE . . . . .	84
13	METHODE DE LA PUISSANCE INVERSE DE WIELANDT . . . . .	85
14	Transformation sous forme tridiagonale (ou de Hessenberg) . . . . .	87
14.1	a) A l'aide des transformations élémentaires . . . . .	87
14.2	b) A l'aide des transformations orthogonales . . . . .	88
14.3	Méthode de bisection pour des matrices tridiagonales . . . . .	88
14.4	Méthode de bisection. . . . .	90
15	L'itération orthogonale . . . . .	90
15.1	Généralisation de la méthode de la puissance (pour calculer les deux va- leurs propres dominantes). . . . .	91
15.2	Méthode de la puissance (pour le calcul de toutes les valeurs propres) . . . .	92
15.3	L' algorithme QR . . . . .	93
15.4	Accélération de la convergence . . . . .	94
15.5	Critère pour arrêter l'itération. . . . .	94
15.6	Le "double shift" de Francis . . . . .	95
15.7	Etude de la convergence . . . . .	96
16	Exercices . . . . .	96
17	TRANSFORMATION SOUS FORME TRIDIAGONALE (ou de HESSENBERG) . . . .	99
17.1	a) A l'aide des transformations élémentaires . . . . .	100
17.2	b) A l'aide des transformations orthogonales . . . . .	100
17.3	Méthode de bisection pour des matrices tridiagonales . . . . .	101
17.4	Méthode de bisection. . . . .	102
18	L'ITERATION ORTHOGONALE . . . . .	103
18.1	Généralisation de la méthode de la puissance (pour calculer les deux va- leurs propres dominantes). . . . .	103
18.2	Méthode de la puissance (pour le calcul de toutes les valeurs propres) . . . .	105
18.3	L' algorithme QR . . . . .	105
18.4	Accélération de la convergence . . . . .	106
18.5	Critère pour arrêter l'itération. . . . .	107
18.6	Le "double shift" de Francis . . . . .	108
18.7	Etude de la convergence . . . . .	109
19	Exercices . . . . .	109



# NOTIONS D'ERREURS



## 1 PRÉLIMINAIRES

L'outil fondamental en analyse numérique (qui nous fournit des méthodes de calcul pour l'étude et la solution approchée de problèmes mathématiques dont la résolution est généralement impossible ou impraticable), demeure la formule de Taylor. Ces solutions approchées sont le plus souvent calculées sur ordinateur au moyen d'algorithmes convenables. Dans ce qui suit nous rappelons quelques théorèmes dont la connaissance est impérative pour une meilleure compréhension de la suite.

*Théorème 1* (de la valeur intermédiaire). Soit  $f$  une fonction définie sur un intervalle  $[a, b]$ , on définit  $m = \inf_{x \in [a, b]} f(x)$  et  $M = \sup_{x \in [a, b]} f(x)$ . Alors pour tout  $y$  dans  $[m, M]$ , il existe au moins un point  $x$  dans  $[a, b]$  pour lequel

$$f(x) = y.$$

*Théorème 2* (des accroissements finis). Soit  $f$  une fonction définie et continue sur un intervalle  $[a, b]$ , différentiable sur  $]a, b[$ . Alors il existe au moins un point  $c$  dans  $]a, b[$  pour lequel

$$f(b) - f(a) = f'(c)(b - a).$$

*Théorème 3* (de la moyenne). Soit  $w(x)$  une fonction non négative définie et intégrable sur un intervalle  $[a, b]$ , et soit  $f(x)$  une fonction continue sur  $]a, b[$ . Alors

$$\int_a^b w(x)f(x)dx = f(\xi) \int_a^b w(x)dx$$

pour  $\xi \in [a, b]$ .

*Théorème 4* (de Taylor). Soit  $f$  une fonction définie et continue sur un intervalle  $[a, b]$ ,  $(n + 1)$  fois dérivable sur  $]a, b[$  pour  $n \geq 0$ , et soit  $x, x_0 \in [a, b]$ . Alors

$$f(x) = p_n(x) + R_{n+1}(x). \quad (1.1)$$

Où

$$\begin{aligned} p_n(x) &= f(x_0) + \frac{(x - x_0)}{1!} f'(x_0) \\ &+ \dots + \frac{(x - x_0)^n}{n!} f^n(x_0) \end{aligned}$$

Et

$$\begin{aligned} R_{n+1}(x) &= \frac{1}{n!} \int_{x_0}^x (x - t)^n f^{(n+1)}(t) dt \\ &= \frac{(x - x_0)^{n+1}}{(n + 1)!} f^{(n+1)}(\xi) \end{aligned} \quad (1.2)$$

pour  $\xi \in ]x_0, x[$

En utilisant la formule de Taylor on obtient par exemple les formules suivantes :

$$e^x = 1 + x + \frac{x^2}{2} + \dots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n + 1)!} e^{\xi_x} \quad (1.3)$$

$$\begin{aligned} \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \dots + (-1)^n \frac{x^{2n}}{2n!} + \\ &+ (-1)^{n+1} \frac{x^{2n+2}}{(2n + 2)!} \cos(\xi_x) \end{aligned} \quad (1.4)$$

$$(1 - x)^{-1} = 1 + x + x^2 + \dots + x^n + \frac{x^{n+1}}{1 - x}, \quad x \neq 0 \quad (1.5)$$



De cette dernière formule nous pouvons déduire :

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \quad |x| < 1 \quad (1.6)$$

On peut calculer les séries de Taylor de n'importe quelle fonction suffisamment dérivable avec autant de termes que l'on veut. Cependant à cause de la complexité de la différentiation de plusieurs fonctions, il est souvent préférable d'obtenir indirectement leur polynôme d'approximation de Taylor  $p_n(x)$  ou leur séries de Taylor, en utilisant l'un des développements limités connus. Les trois exemples qui suivent montrent que les erreurs sont plus simples que lorsque l'on utilise la formule de l'erreur (1.2).

### 1.1 Exemples

1.  $f(x) = e^{-x^2}$ , En remplaçant  $x$  par  $-x^2$  dans (1.3), on obtient :

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2} - \dots + (-1)^n \frac{x^{2n}}{n!} + (-1)^{n+1} \frac{x^{2n+2}}{(n+1)!} e^{\xi_x}$$

avec  $\xi_x \in [-x^2, 0]$ .

2.  $f(x) = \frac{1}{\tan(x)}$ , En posant  $x = -u^2$  dans le développement de  $\frac{1}{1-x}$  on a :

$$\frac{1}{1+u^2} = 1 - u^2 + u^4 - \dots + (-1)^n u^{2n} + (-1)^{n+1} \frac{u^{2n+2}}{1+u^2}$$

en intégrant sur  $[0, x]$  on aboutit à :

$$\frac{1}{\tan(x)} = x - \frac{x^3}{3} + \frac{x^5}{5} - \dots + (-1)^n \frac{x^{2n+1}}{2n+1} + (-1)^n \int_0^x \frac{u^{2n+2}}{1+u^2} du$$

En appliquant le théorème de la moyenne, on obtient :

$$\int_0^x \frac{u^{2n+2}}{1+u^2} du = \frac{x^{2n+3}}{2n+3} \cdot \frac{1}{1+\xi_x^2}$$

avec  $\xi_x \in [0, x]$ .

3.  $f(x) = \int_0^1 \sin(xt) dt$ , Utilisant le développement de  $\sin x$ , et intégrant, on écrit :

$$f(x) = \sum_{j=1}^n (-1)^{j-1} \frac{x^{2j-1}}{2j!} + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \int_0^1 t^{2n+1} \cos(\xi_{xt}) dt$$

avec  $\xi_{xt} \in [0, xt]$ . L'intégrale dont le reste est bornée par  $\frac{1}{2n+2}$ , mais on peut aussi la mettre sous une forme simplifiée, et en appliquant le théorème de la moyenne on a :

$$\int_0^1 \sin(xt) dt = \sum_{j=1}^n (-1)^{j-1} \frac{x^{2j-1}}{2j!} + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \cos(\xi_x)$$

avec  $\xi_x \in [0, x]$ .

## 2 Erreurs absolues et Erreurs relatives

Un nombre approché  $x$  est légèrement différent du nombre exact  $X$ , et qui dans les calculs remplace  $X$ .

- Si  $x < X$ ,  $x$  est dit valeur par *défaut*.
- Si  $x > X$ ,  $x$  est dit valeur par *excès*.

On note généralement  $x \approx X$ .

## 2.1 Exemple

$$1,41 < \sqrt{2} < 1,42$$

*Définition 5.* On appelle erreur  $\Delta x$  d'un nombre approché, la valeur :

$$\Delta x = X - x,$$

C'est-à-dire

$$X = x + \Delta x$$

*Définition 6.* On appelle erreur absolue  $\Delta$  d'un nombre  $x$  la valeur

$$\Delta = |X - x| \quad (2.1)$$

*Remarque 7.* 1. Si  $X$  est connu, l'erreur absolue est déterminée par (2.1).

2. Si  $X$  est inconnu, l'erreur absolue  $\Delta$  est impossible à déterminer. Dans ce cas on introduit la limite supérieure de l'erreur absolue.

*Définition 8.* On appelle borne supérieure de l'erreur absolue tout nombre supérieur ou égal à l'erreur absolue de ce nombre. C'est-à-dire :

$$\Delta = |X - x| \leq \Delta_x$$

si  $\Delta_x$  désigne la borne supérieure, donc

$$x - \Delta_x \leq X \leq x + \Delta_x$$

On note

$$X = x \pm \Delta_x.$$

## 2.2 Exemples

Trouver la borne supérieure d'erreur absolue de  $\pi = 3,14$ .

$$3,14 \leq \pi \leq 3,15$$

Donc  $|x - \pi| \leq 0,01$ , on peut poser  $\Delta_x = 0,01$ , comme  $3,140 \leq \pi \leq 3,142$ , une meilleure estimation de la borne d'erreur absolue est  $\Delta = 0,002$ .

*Définition 9.* On appelle erreur relative notée  $\delta$  d'un nombre  $x$ , le rapport suivant :

$$\delta = \frac{\Delta}{|X|}, \quad (X \neq 0)$$

C'est-à-dire

$$\Delta = |X| \delta.$$

*Définition 10.* La borne supérieure de l'erreur relative  $\delta_x$  est un nombre supérieur ou égal à l'erreur relative de ce nombre. C'est-à-dire :

$$\delta \leq \delta_x$$

c'est-à-dire

$$\frac{\Delta}{|X|} \leq \delta_x,$$

donc

$$\Delta \leq |X| \delta_x$$

On peut aussi utiliser

$$\Delta_x = |x| \delta_x$$

car  $X \approx x$ .

Remarque 11. Si l'on connaît une borne d'erreur relative  $\delta_x$  on a :

$$X = x(1 \pm \delta_x)$$

C'est-à-dire

$$\delta_x = \frac{\Delta_x}{x - \Delta_x}.$$

De même on obtient :

$$\Delta_x = \frac{x\delta_x}{1 - \delta_x}.$$

### Exemple

En cherchant la constante de gaz, on a obtenu  $R = 29,25$ , l'erreur relative étant  $1^0/_{00}$  trouver un encadrement de  $R$ . On a  $\delta_x = 0,001$ , donc  $\Delta_x = R\delta_x = 0,03$ , c'est-à-dire :

$$29,22 \leq R \leq 29,28.$$

## 3 PRINCIPALES SOURCES D'ERREURS

Les erreurs commises dans les problèmes peuvent être des :

**Erreurs inhérentes au problème :** Erreurs dues à la position même du problème. Le modèle théorique est très rarement fidèle au modèle réel. Lors de l'étude d'un phénomène de la nature on est souvent contraint d'admettre certaines conditions.

**Erreurs de la méthode :** Il arrive qu'il soit difficile ou même impossible de résoudre un problème énoncé en termes exacts. On le remplace par un problème approché.

**Erreurs de troncature :** Associées aux processus infinis. Les fonctions données dans les formules le sont sous forme de suites infinies ou de séries, on est donc obligé de mettre fin à un certain terme de la suite. Par exemple, l'approximation d'une somme infinie par une somme finie, l'approximation de la limite d'une suite par un terme de "grand indice" ou encore l'approximation d'une intégrale par une somme finie.

**Erreurs initiales :** Dûes à la présence dans les formules de paramètres dont les valeurs sont approchées.

**Erreurs d'arrondi :** Dûes au système de numérisation.

**Erreurs propagées :** Les erreurs des données de départ se repercutent sur le résultat des calculs.

## 4 PRECISION, CHIFFRES SIGNIFICATIFS

### 4.1 Chiffres significatifs

*Définition 12.* On appelle chiffre significatif d'un nombre tous les chiffres de son écriture à partir du premier chiffre différent de zéro à gauche.

**Exemple** Les chiffres significatifs des nombres  $x = 0,03045$  et  $x = 0,03045000$  sont ceux soulignés. Ils sont 4 chiffres dans le premier cas et 7 dans le deuxième.

*Définition 13.* Un chiffre significatif est dit *exact* si l'erreur absolue sur le nombre ne dépasse pas l'unité de l'ordre correspondant.

**Exemple**  $x = 0,03045$ ,  $\Delta(x) = 0,000003$ ;  $x = 0,03045000$ ,  $\Delta(x) = 0,0000007$ ; les chiffres soulignés sont exacts.

*Définition 14.* Si tous les chiffres significatifs sont *exact*s, on dit que le nombre est écrit avec tous les chiffres exacts.

**Exemple**  $x = 0,03045$ ,  $\Delta(x) = 0,000003$ ,  $x$ , est écrit avec 4 chiffres exacts. On parle souvent de *décimales exactes*. Dans le dernier exemple le nombre  $x$  est écrit avec 5 décimales exactes.

### 4.1.1 Règle pour arrondir les nombres

Soit  $x$  un nombre approché sous forme décimale. Pour l'arrondir jusqu'à  $n$  chiffres significatifs, c'est-à-dire le remplacer par un nombre  $x_1$ <sup>1</sup> avec  $n$  chiffres significatifs, on rejette tous les chiffres à droite du  $n^{\text{ième}}$  chiffre significatif ou s'il faut conserver les rangs, on les remplace par des zéros.

Dans ces cas :

- Si le premier des chiffres rejetés est inférieur à 5, les chiffres restent inchangés.
- Si le premier des chiffres rejetés est supérieur à 5, on ajoute une unité au dernier chiffre restant.
- Si le premier des chiffres rejetés est égal à 5, et si parmi les autres chiffres rejetés il y en a des non nuls, le dernier chiffre restant est augmenté de l'unité.
  - Mais si le premier des chiffres rejetés est égal à 5 alors que les autres chiffres rejetés sont nuls, le dernier chiffre conservé reste inchangé s'il est pair ou on lui ajoute une unité s'il est impair.

**Exemple** En arrondissant le nombre

$$x = 3,045166382535$$

jusqu'à 5; 4 et 3 chiffres significatifs, on obtient les nombres approchés 3,0452, 3,045 et 3,05

## 5 Cumulation des erreurs d'arrondi

### 5.1 Erreurs d'arrondi sur une somme

Soient  $X, Y$  des nombres réels supposés représentés sans erreur avec  $N$  chiffres significatifs :

$$\begin{aligned} X &= 0, a_1 a_2 \dots a_N \cdot b^p, & b^{-1+p} \leq X < b^p \\ Y &= 0, a'_1 a'_2 \dots a'_N \cdot b^q, & b^{-1+q} \leq Y < b^q \end{aligned}$$

Et  $\Delta(X+Y)$  l'erreur d'arrondi commise sur le calcul de  $X+Y$ . Supposons  $p \geq q$ .

- Si  $X+Y < b^p$ , le calcul de  $X+Y$  s'accompagne de la perte des  $p-q$  derniers chiffres de  $Y$  correspondants aux puissances  $b^{-k+q} < b^{-N+p}$ ; donc  $\Delta(X+Y) \leq b^{-N+p}$ , alors que  $X+Y \geq X \geq b^{-1+p}$ .
- Si  $X+Y \geq b^p$ , la décimale correspondant à la puissance  $b^{-N+p}$  est elle aussi perdue, d'où  $\Delta(X+Y) \leq b^{1-N+p}$ .

Dans les deux cas :

$$\Delta(X+Y) \leq \varepsilon(|X| + |Y|),$$

Où  $\varepsilon = b^{1-N}$  est la précision relative. Ceci est vrai quel que soit le signe de  $X$  et de  $Y$ . En général, les réels  $X, Y$  ne sont connus que par des valeurs approchées  $x, y$  avec des erreurs respectives  $\Delta_x = |X-x|, \Delta_y = |Y-y|$ . A ces erreurs s'ajoutent l'erreur d'arrondi :

$$\Delta(x+y) \leq \varepsilon(|x| + |y| + \Delta_x + \Delta_y).$$

Les erreurs  $\Delta_x, \Delta_y$  sont elles mêmes le plus souvent d'ordre  $\varepsilon$  par rapport à  $|x|$  et  $|y|$ , de sorte que l'on pourra "négliger" les termes  $\varepsilon\Delta_x$  et  $\varepsilon\Delta_y$ . On aura :

$$\Delta(x+y) \leq \Delta_x + \Delta_y + \varepsilon(|x| + |y|).$$

*Remarque 15.* Pour calculer une somme de réels positifs  $\sum_{k=1}^n u_k$ , on calcule les sommes partielles  $s_k = u_1 + u_2 + \dots + u_k$  de proche en proche par les formules de récurrence :

$$\begin{cases} s_0 &= 0 \\ s_k &= s_{k-1} + u_k, \quad k \geq 1 \end{cases}$$

1. le nombre  $x_1$  est choisi de façon à minimiser l'erreur d'arrondi  $|x_1 - x|$ .

Si les  $u_k$  sont connus exactement, on aura sur  $s_k$  des erreurs  $\Delta_{s_k}$  telles que  $\Delta_{s_1} = 0$  et

$$\Delta_{s_k} \leq \Delta_{s_{k-1}} + \varepsilon(s_{k-1} + u_k) = \Delta_{s_{k-1}} + \varepsilon s_k.$$

L'erreur globale sur  $s_n$  vérifie

$$\Delta_{s_n} \leq \varepsilon(s_2 + s_3 + \dots + s_n).$$

## 5.2 Erreurs d'arrondi sur un produit

Le produit de deux mantisses de  $N$  chiffres donne une mantisse de  $2N$  ou  $2N - 1$  chiffres dont les  $N$  ou  $N - 1$  derniers vont être perdus. Dans le calcul d'un produit  $XY$  il y aura donc une erreur d'arrondi

$$\Delta(xy) \leq \varepsilon |XY|, \quad \text{où } \varepsilon = b^{1-N}.$$

Si  $X$  et  $Y$  ne sont connus que par des valeurs approchées  $x, y$  et si  $\Delta_x = |X - x|$ ,  $\Delta_y = |Y - y|$ , on a une erreur initiale :

$$|XY - xy| = |X(y - Y) + (x - X)y| \leq |X| \Delta_y + \Delta_x |y|$$

A cette erreur s'ajoute une erreur d'arrondi :

$$\Delta(xy) \leq \varepsilon |xy| \leq \varepsilon(|x| + \Delta_x)(|y| + \Delta_y).$$

Ce qui donne la formule approximative :

$$\Delta(xy) \leq |x| \Delta_y + \Delta_x |y| + \varepsilon |xy|.$$

Cette dernière formule nous permet d'obtenir par récurrence :

$$\Delta(x_1 x_2 \dots x_k) \leq (k - 1)\varepsilon |x_1 x_2 \dots x_{k-1} \cdot x_k|.$$

*Remarque 16.* La majoration de l'erreur d'un produit ne dépend pas de l'ordre des facteurs.

## 6 Représentation approchée des nombres réels

L'objet de cette section est de mettre en évidence les principales difficultés liées à la pratique des calculs numériques sur ordinateur. La capacité mémoire d'un ordinateur est par construction finie. Si  $X$  est un nombre réel, il est donc nécessaire de représenter  $X$  sous forme approchée. La notation la plus utilisée est la représentation avec *virgule flottante* :

$$X \simeq \pm m \cdot b^p$$

Où  $b$  désigne la base de numération,  $m$  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 mantisse  $m$  est un nombre écrit avec virgule fixe et possédant un nombre maximum  $N$  de chiffres significatifs (imposé par la mémoire de l'ordinateur) : suivant les machines,  $m$  s'écrira

$$m = 0, a_1 a_2 \dots a_N = \sum_{k=1}^N a_k b^{-k}, \quad b^{-1} \leq m < 1.$$

Ceci entraîne que la précision dans l'approximation d'un nombre réel est toujours une précision relative :

$$\frac{\Delta_x}{X} = \frac{\Delta_m}{m} \leq \frac{b^{-N}}{b^{-1}} = b^{1-N}.$$

On note  $\varepsilon = b^{1-N}$  cette précision relative. **Exemple.** La même écriture peut représenter des nombres différents dans des bases différentes : 123,45 en base 10 représente le nombre  $x = 1.10^2 + 2.10 + 3 + 4.10^{-1} + 5.10^{-2}$ , en base 6 il représente le nombre  $y = 1.6^2 + 2.6 + 3 + 4.6^{-1} + 5.6^{-2}$ . D'autre part, le même nombre peut avoir un nombre fini de chiffres dans une base, et un nombre infini dans une autre base :  $x = 1/3$  donne  $x_3 = 0,1$  en base 3 et  $x_{10} = 0,3$  en base 10.

## 6.1 Nombres en virgule flottante

De nombreuses manières ont été proposées pour représenter les nombres par un ordinateur mais la plus utilisée aujourd'hui est donc la représentation dite en "virgule flottante", et en base 2 (On utilise aussi la base 8 et la base 16 (numérotation hexadécimale à seize chiffres, de 0 à 9, auxquels on rajoute les lettres A,B,C,D,E et F)). La manière courante d'écrire les nombres aujourd'hui est la notation de position en base dix. On se donne donc dix symboles  $0; 1; 2 = 1 + 1; 3 = 2 + 1; \dots; 9 = 8 + 1$ . La représentation d'un nombre entier est simplement une suite finie de tels symboles. Par exemple 27821 n'est que le symbole 2 suivit du symbole 7, . . . Pour mettre en évidence l'aspect suite de symboles, nous écrivons :

$$\boxed{2} \boxed{7} \boxed{8} \boxed{2} \boxed{1}$$

Pour savoir à quel nombre correspond cette suite de symboles, on les interprète selon leur position. Ici la base est  $b = 10 = 9 + 1$ . Cela signifie que chaque fois qu'on se déplace d'un chiffre vers la gauche, la puissance de dix augmente de 1 :

$$\boxed{2[10^4]} \quad \boxed{7[10^3]} \quad \boxed{8[10^2]} \quad \boxed{2[10^1]} \quad \boxed{1[10^0]}$$

Ainsi, le nombre représenté est  $2b^4 + 7b^3 + 8b^2 + 2b^1 + 1b^0 = 2 \cdot 10^4 + 7 \cdot 10^3 + 8 \cdot 10^2 + 2 \cdot 10^1 + 1 \cdot 10^0$ . Ceci c'était pour les nombres entiers. Qu'en est-il des écritures avec virgule? Par exemple, que

représente 0,2? Simplement, c'est  $2/10 = 2 \cdot 10^{-1}$ . Et 1,74 représente  $1 \cdot 10^0 + 7 \cdot 10^{-1} + 4 \cdot 10^{-2}$ . De manière générale,

$$\boxed{\pm \quad a_n \quad \dots \quad a_1 \quad a_0 \quad , a_{-1} \quad a_{-2} \quad \dots}$$

représente le nombre  $\pm(a_n 10^n + \dots + a_1 10^1 + a_0 10^0 + a_{-1} 10^{-1} + a_{-2} 10^{-2} + \dots)$ , c'est-à-dire

$$\pm \sum_{i=-\infty}^n a_i 10^i.$$

Notons que la suite des décimales  $a_{-1}, a_{-2}, \dots$  peut être finie ou infinie et que dans ce dernier cas la série converge. Représenter les nombres en base deux se fait exactement de la même manière

qu'en base dix excepté qu'on remplace dix par deux! Ainsi, on se donne deux symboles 0 et 1. A une suite de tels symboles

$$\boxed{\pm \quad a_n \quad \dots \quad a_1 \quad a_0 \quad , a_{-1} \quad a_{-2} \quad \dots} \quad a_i \in \{0, 1\},$$

on fait correspondre le nombre

$$\pm \sum_{i=-\infty}^n a_i 2^i = \pm(a_n 2^n + \dots + a_1 2 + a_0 + a_{-1} 2^{-1} + \dots)$$

Nous noterons ce nombre  $(\pm a_n \dots a_1 a_0, a_{-1} a_{-2} \dots)_2$ . On a donc  $(110)_2 = 2^2 + 2^1 = 5$  et  $(1, 11)_2 = 2^0 + 2^{-1} + 2^{-2} = 1,75$ .

## 6.2 Non-associativité des opérations arithmétiques.

Supposons par exemple que les réels soient calculés avec  $N = 3$  chiffres significatifs et arrondis à la décimale la plus proche. Soient

$$x = 8,22 = 0,822 \cdot 10, \quad y = 0,00317 = 0,317 \cdot 10^{-2}, \quad z = 0,00432 = 0,432 \cdot 10^{-2}.$$

On veut calculer la somme  $x + y + z$ .  $(x + y) + z$  donne :

$$x + y = 8,22317 \simeq 0,822.10$$

$$(x + y) + z \simeq 8,22432 \simeq 0,822.10$$

$x + (y + z)$  donne :

$$y + z = 0,00749 \simeq 0,749.10^{-2}$$

$$x + (y + z) = 8,22749 \simeq 0,823.10$$

L'addition est donc non associative par suite des erreurs d'arrondi.

*Remarque 17.* En générale, dans une sommation de réels, l'erreur a tendance à être minimisée lorsqu'on somme en premier les termes ayant la plus petite valeur absolue.

### 6.3 Phénomènes de compensation.

Lorsqu'on tente d'effectuer des soustractions de valeurs très voisines, on peut avoir des pertes importantes de précision. **Exemple.** On veut résoudre l'équation  $x^2 - 1634x + 2 = 0$  en effectuant les calculs avec  $N = 10$  chiffres significatifs. On obtient

$$\Delta' = 667487, \quad \sqrt{\Delta'} = 816,9987760,$$

$$x_1 = 817 + \sqrt{\Delta'} \simeq 1633,998776,$$

$$x_2 = 817 - \sqrt{\Delta'} \simeq 0,0012240.$$

On voit qu'on a une perte de 5 chiffres significatifs sur  $x_2$ . Ici le remède est simple : il suffit d'observer que  $x_1 \cdot x_2 = 2$ , d'où

$$x_2 = \frac{2}{x_1} = 1,223991125.10^{-3}.$$

C'est donc l'algorithme numérique utilisé qui doit être modifié.

## 7 SERIE D'EXERCICES

*Exercice 18.* Trouver une borne de l'erreur absolue du nombre  $x = 3.14$  qui remplace  $\pi$  ( $\pi = 3.1415926\dots$ ) dans les deux cas suivants :

1.  $3.14 < \pi < 3.142$ .
2.  $3.14 < \pi < 3.15$ .

*Exercice 19.* Supposons que  $x_1, x_2, x_3, \dots, x_n$  approchent respectivement  $X_1, X_2, X_3, \dots, X_n$  et que dans chaque cas la borne supérieure de l'erreur absolue est  $\varepsilon$ . Montrer que la borne supérieure de l'erreur de la somme des  $x_i$  ( $i = 1, 2, \dots, n$ ) est égale à  $n\varepsilon$ .

*Exercice 20.* Donner les bornes des erreurs relatives de  $a = 1.414$  et  $b = 1.41$  qui approchent  $\sqrt{2} = 1.414214\dots$ .

*Exercice 21.* En recherchant la constante des gaz de l'air on a obtenu  $R \simeq 29.25$ . La borne de l'erreur relative de cette valeur étant  $1^0/00$ , trouver les limites entre lesquelles est comprise  $R$ .