Analyse Numérique - Fondements

1 Introduction générale

Les méthodes utilisées en mathématiques classiques sont incapables de résoudre tous les problèmes. On ne sait pas, par exemple, donner une formule pour calculer exactement les racines des équations de degré 5 ou plus ; on ne sait pas non plus trouver la solution analytique de certaines équations différentielles ni calculer certaines intégrales définies. On remplace alors la résolution mathématique exacte du problème par sa résolution numérique qui est, en général, approchée.

L'analyse numérique est la branche des mathématiques qui s'intéresse à la mise en pratique des méthodes constructives de résolution numérique des problèmes. Par méthode constructive, on entend un ensemble de règles (on dit : algorithme) qui permet d'obtenir la solution numérique d'un problème avec une précision désirée après un nombre fini d'opérations arithmétiques.

Exemple : Pour trouver la valeur de la racine carrée de 2 à une certaine précision on peut procéder par la méthode suivante :

$$x_1 = 1; \qquad x_{n+1} = \frac{1}{2}(x_n + \frac{2}{x_n}).$$

On trouve

$$x_2 = 3/2 = 1.5 \qquad x_2 = 17/12 = 1,4167 \qquad x_3 = \frac{1}{2}(17/12 + 24/17) = 1,4142...$$

L'analyse numérique est une branche assez ancienne des mathématiques. Autrefois, en effet, les mathématiciens développaient les outils dont ils avaient besoin pour résoudre les problèmes posés par les sciences de la nature. C'est ainsi que Newton était avant tout un physicien, Gauss un astronome... Ils s'aperçurent rapidement que les problèmes pratiques qui se posaient étaient trop compliqués pour leurs outils et c'est ainsi que, peu à peu, s'élaborèrent les techniques de l'analyse numérique. Ces méthodes ne connurent cependant leur essor actuel qu'avec l'avènement des ordinateurs à partir des années 1945-1947.

L'objectif de ce cours est de donner à l'étudiant une introduction élémentaire de ce champs passionnant et actif des mathématiques appliquées.

2 Les nombres

2.1 Système de numérotation à base b

C'est un moyen de représenter les nombres avec b symboles. Selon sa place, le symbole indique une valeur particulière :

$$\begin{aligned}(a_n a_{n-1} \dots a_1 a_0, a_{-1}, \dots, a_{-k})_b = a_n \times b^n + a_{n-1} \times b^{n-1} + \cdots + a_1 \times b^1 \\ + a_0 \times b^0 + a_{-1} \times b^{-1} + \cdots + a_{-k} \times b^{-k}\end{aligned}$$

Système décimal

Il s'agit du système habituel. Ce système correspond à une représentation à base 10 ($b = 10$) :

$$3724 = (3724)_{10} = 3 \times 10^3 + 7 \times 10^2 + 2 \times 10^1 + 4 \times 10^0.$$
$$10,25 = (10,25)_{10} = 1 \times 10^1 + 0 \times 10^0 + 2 \times 10^{-1} + 5 \times 10^{-2}.$$

Système binaire

Il s'agit du système utilisé en générale sur les ordinateurs. Ce système correspond à une représentation à base de 2 ($b = 2$) :

$$\begin{aligned}(101)_2 &= 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = (5)_{10}, \\ (1111)_2 &= 1 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = (15)_{10}, \\ (10,1)_2 &= 1 \times 2^1 + 0 \times 2^0 + 1 \times 2^{-1} = (2,5)_{10}, \\ (1,01)_2 &= 1 \times 2^0 + 0 \times 2^{-1} + 0 \times 2^{-2} = (1,25)_{10}.\end{aligned}$$

Conversion du décimal au binaire

Cas des entiers :

Pour transformer un entier positif N dans sa représentation binaire habituelle, il faut déterminer les a_i tels que :

$$(N)_{10} = (a_n, a_{n-1}, \dots, a_1, a_0)_2$$

ou encore

$$N = a_n \times 2^n + a_{n-1} \times 2^{n-1} + \cdots + a_1 \times 2^1 + a_0 \times 2^0.$$

Exemple : Si $(N)_{10} = 30$

$$\begin{aligned} 30/2 = 15 \quad \text{reste} = 0 \quad &\Longrightarrow \quad a_0 = 0 \\ 15/2 = 7 \quad \text{reste} = 1 \quad &\Longrightarrow \quad a_1 = 1 \\ 7/2 = 3 \quad \text{reste} = 1 \quad &\Longrightarrow \quad a_2 = 1 \\ 3/2 = 1 \quad \text{reste} = 1 \quad &\Longrightarrow \quad a_3 = 1 \\ 1/2 = 0 \quad \text{reste} = 1 \quad &\Longrightarrow \quad a_4 = 1 \end{aligned}$$

Ainsi l'entier décimal 30 s'écrit 11110 en binaire.

Cas des fractions :

Soit f une fraction décimale comprise entre 0 et 1. Il faut trouver les a_i tels que :

$$(f)_{10} = (0, a_{-1}, a_{-2}, , a_{-3} \ldots)_2$$

ou encore

$$f = a_{-1} \times 2^{-1} + a_{-2} \times 2^{-2} + a_{-3} \times 2^{-3} + \ldots$$

Si on multiplie f par 2, on obtient a_{-1} plus une fraction. En appliquant le même raisonnement à $(2f - a_{-1})$, on obtient a_{-2} On poursuit ainsi jusqu'à ce que la partie fractionnaire soit nulle ou que l'on ait atteint le nombre maximal de chiffres de la mantisse.

Exemple : Si $f = 0,0625$ on a :

$$\begin{aligned} 0,0625 \times 2 = 0,1250 \quad &\Longrightarrow \quad a_{-1} = 0 \\ 0,1250 \times 2 = 0,2500 \quad &\Longrightarrow \quad a_{-2} = 0 \\ 0,2500 \times 2 = 0,500 \quad &\Longrightarrow \quad a_{-3} = 0 \\ 0,500 \times 2 = 1,000 \quad &\Longrightarrow \quad a_{-4} = 1 \end{aligned}$$

Ainsi $(0,0625)_{10} = (0,0001)_2$.

2.2 Représentation des réels

Notation scientifique

Cette notation consiste à représenter un nombre donné x sous la forme

$$x = \pm m \times 10^l$$

avec

$$m = d_1, d_2 d_3 \ldots \qquad 0 \leq d_i \leq 9 \text{ et } d_1 \neq 0.$$

$\pm$: le signe,
$l \in \mathbb{Z}$: l'exposant,
m : la mantisse.

Exemple : Les nombres suivants sont représentés selon la notation scientifique

$$1,25020 \times 10^{-6} \quad -1,0125 \times 10^{10} \quad 7 \times 10^{0} \quad -5,0105242 \times 10^{2}$$

Ces écritures ne correspond pas à la notation scientifiques.

$$0,125020 \times 10^{-5} \quad -10,125 \times 10^{9} \quad 7 \quad -501,05242$$

Notation en virgule flottante

La convention adopté pour le stockage des réels sur les calculateurs/ordinateurs consiste à représenter un réel donné x sous la forme

$$x = \pm m \times b^{l}.$$

avec

$$m = 0, d_1 d_2 d_3 \ldots \qquad 0 \leq d_i \leq b-1 \text{ et } d_1 \neq 0.$$

$\pm$: le signe,
$l \in \mathbb{Z}$: l'exposant,
m : la mantisse.
Exemple : Les nombres suivants sont représentés selon la norme de la représentation virgule flottante

$$0,125020 \times 10^{-5} \quad -0,10125 \times 10^{11} \quad 0,7 \times 10^{1} \quad -0,50105242 \times 10^{3}$$

$$(0,1010)_2 \times 2^{(10)_2} \qquad (0,100)_2 \times 2^{(10)_2}$$

Ces écritures ne correspond pas à la notation virgule flottante.

$$1,25020 \times 10^{-5} \quad -10,125 \times 10^{9} \quad 7 \quad -0,015242$$

2.3 Troncature et Arrondi

Supposons que l'on veuille employer n chiffres pour représenter un nombre.

Troncature

On abondonne les derniers chiffres au delà du n ème chiffre.
Exemple : $3,1415$ est la troncature de $\pi = 3,14159265\ldots$ à 5 chiffres
$1,41$ est la troncature de $\sqrt{(2)} = 1,4142135623\ldots$ à 3 chiffres.

Arrondi

On ajoute 5 au $(n+1)$ ème chiffre puis on fait la troncature.
Exemple : $3,1416$ est l'arrondi de $\pi = 3,14159265\ldots$ à 5 chiffres
$3,14$ est l'arrondi de $\pi = 3,14159265\ldots$ à 3 chiffres
$1,414214$ est l'arrondi de $\sqrt{(2)} = 1,4142135623\ldots$ à 7 chiffres.

3 Erreur

Erreur absolue

Soit x, un nombre, et x^* une approximation de ce nombre. L'erreur absolue est définie par

$$\Delta x = |x - x^*|$$

Exemple : Si $x = 2,224$ et $x^* = 2,223$ alors l'erreur absolue $\Delta x = |x - x^*| = 0,001 = 1 \times 10^{-3}$.

Erreur relative

Soit x, un nombre, et x^* une approximation de ce nombre. L'erreur relative est définie par

$$E_r = \frac{|x - x^*|}{|x|} = \frac{\Delta x}{|x|}.$$

Exemple : Si $x = 2,224$ et $x^* = 2,223$ alors l'erreur relative $E_r = \dfrac{0,001}{2,224} = 4,496 \times 10^{-6}$.

Remarque

1) En pratique, il est difficile d'évaluer les erreurs absolues et relatives puisque on dispose pas en générale de la valeur exacte de x et on n'a que x^*. Dans le cas de quantités mesurées on dispose souvent d'une borne supérieure pour l'erreur absolue qui dépend des instruments utilisés. Cette borne est quand même appelée erreur absolue alors qu'en fait on a que

$$|x - x^*| \leq \Delta x \Longleftrightarrow x^* - \Delta x \leq x \leq x^* + \Delta x$$

On écrit parfois

$$x = x^* \pm \Delta x.$$

2) L'erreur absolue donne une mesure quantitative de l'erreur commise et l'erreur relative en mesure l'importance. Par exemple, si on fait l'usage d'un chronomètre dont la précision est de l'ordre de $0,1\,s$. Est-ce une erreur importante ?
   a) Dans un marathon d'une durée de $2h\ 20min$ on a

$$E_r = \frac{0,1}{2 \times 60^2 + 20 \times 60} = 0,0000119$$

qui est une valeur très faible et ne devrai pas avoir de conséquence sur le classement.

b) Dans une course de $100m$ d'une durée d'environ $10s$, l'erreur relative est plus importante

$$E_r = \frac{0,1}{10} = 0,01 = 1\%$$

Avec telle erreur, on ne pourra vraisemblablement pas faire la différence entre le premier et le deuxième coureur.

Chiffres significatifs

Soit x un nombre et x^* une approximation de ce nombre écrite selon la notation scientifique :

$$x^* = d_1, d_2 d_3 \cdots \times 10^k \quad d_1 \neq 0,\ k \in \mathbb{Z}$$

Si l'erreur absolue $\Delta x = |x - x^*|$ vérifie

$$\Delta x \leq 5 \times 10^{k-m}$$

alors on dit que x^* est une approximation de x avec m chiffres significatifs.

Exemples

i) On obtient une approximation de π au moyen de la quantité

$$x^* = 22/7 = 3,142857 \cdots \times 10^0 \quad (k = 0)$$

On en conclut que :

$$\Delta x = |\pi - 22/7| = 0,00126 \cdots = 1,26 \cdots \times 10^{-3} \leq 5 \times 10^{-3} \quad (k - m = 0 - m = -3)$$

Par conséquent $m = 3$ et on a en tout 3 chiffres significatifs $(3, 14)$.

ii) 5.1 est une approximation de 5 avec un seul chiffre significatif (5). On a

$$5 = 5 \times 10^0 \Longrightarrow k = 0$$

$$\Delta x = |5,1 - 5| = 0,1 = 1 \times 10^{-1} \leq 5 \times 10^{-1} \Longrightarrow k - m = 0 - m = -1$$

Inversement, si un nombre est donné avec m chiffres significatifs, alors on peut estimer l'erreur absolue.

Exemple On a mesuré le poids d'une personne et trouvé 90,567 kg. On vous assure que l'appareil utilisé est suffisamment précis pour que tous les chiffres fournis soient significatifs. C'est à dire $m = 5$ On a

$$90,567 = 9,0567 \times 10^1 (k = 1)$$

D'après la définition cela signifie que :

$$\Delta x \leq 5 \times 10^{k-m} = 5 \times 10^{-4}$$

En pratique en conclut que

$$x = (90,567 \pm 5 \times 10^{-4})kg$$

Influence de l'erreur de l'arrondis sur les opérations

Supposons par exemple que les réels sont à calculer avec 3 chiffres significatifs et soit à calculer la somme $x + y + z$ avec $x = 8,22$; $y = 0,00317$ et $z = 0,00432$
On a

$$x + y = 8,22317 \approx 8,22 \quad (x + y) + z \approx 8,22432 \approx 8,22$$

Or

$$y + z = 0,00749 \quad x + (y + z) \approx 8,22749 \approx 8,23$$

L'addition est donc non associative par suite d'erreurs d'arrondis ! !