CHAPITRE 0 : Analyse d'erreurs

1- Introduction :Le calcul scientifique

Le calcul scientifique est une discipline qui regroupe un ensemble de champs mathématiques et informatiques permettant la simulation numérique (par ordinateur) d'un phénomène ou un processus décrit par un modèle mathématique.
Un modèle mathématique est une traduction de la réalité pour pouvoir lui appliquer les outils, les techniques et les théories mathématiques, puis généralement, en sens inverse, la traduction des résultats mathématiques obtenus en prédictions ou opérations dans le monde réel.
Lors de l'établissement d'un modèle, on est constamment tenu de trancher entre deux exigences contradictoires

1. Représenter la réalité avec précision et dans tous ses aspects
2. Ne pas compliquer exagérément le modèle

Après l'étude théorique d'un modèle mathématique, commence la phase de calcul scientifique qui est constitué par trois étapes :

1. Analyse numérique : L'objet de l'analyse numérique est la conception et l'étude de méthodes de résolution numérique des modèles mathématiques des sciences de l'ingénieur, des sciences expérimentales, de l'économie etc....
   L'efficacité d'une méthode dépend aussi bien de ses propriétés mathématiques (convergence, précision, etc...) que de sa facilité d'implémentation et de son comportement sur la machine.
2. Programmation : Implémentation de l'algorithme à l'aide d'un langage de programmation (C, C++, Matlab, ...)
3. Calculs : Donner des solutions approches des problèmes mathématiques
4. Vérification : Validation de la méthode numérique sur des cas tests pour vérifier le comportement de la méthode dans des situations bien connues

Exemple (Modélisation en dynamique des populations)

Les premiers modèles de croissance de populations datent de la fin du 18^e *siècle avec le modèle de Malthus.*
Ce modèle est donné par l'équation différentielle $\frac{dN(t)}{dt} = rN(t)$
Avec : - $N(t)$ *est la taille de la population*
- r est le taux de croissance de cette population.
La solution de cette équation est $N(t) = N_0 e^{rt}, \quad N_0 = N(0)$

- *Si* $r < 0$, *la taille de la population diminue,*
- *Si* $r = 0$, *la population reste constante ,*
- *Si* $r > 0$, *la population augmente de manière exponentielle.*

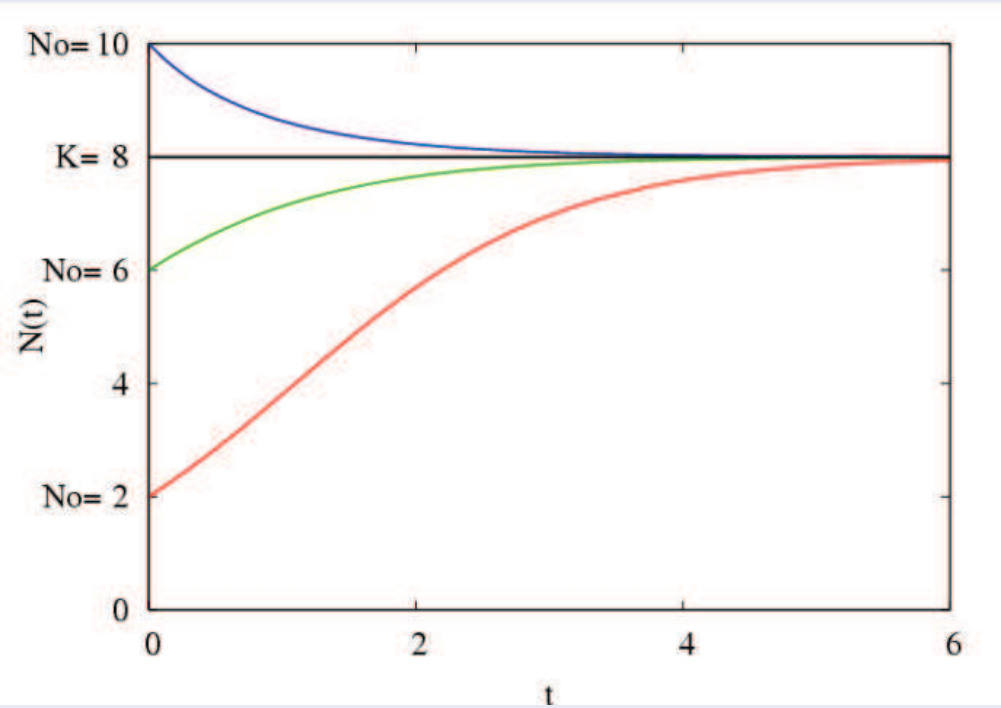
Le modèle malthusien est remis en cause vers 1838 par Pierre François Verhulst (la population augmente de manière exponentielle n'est pas biologiquement satisfaisant) qui propose un modèle dit logistique
Le modèle de Verhulst s'écrit sous la forme $\frac{dN(t)}{dt} = rN(t)\left(1 - \frac{N(t)}{K}\right)$
Avec $r > 0$ *le taux de croissance de la population et* $K > 0$ *la capacité d'accueil du milieu, (c'est-à-dire le nombre d'individus maximal que le milieu peut accueillir en tenant compte de l'espace, des ressources, etc*$\cdots$ *).*

Exemple (suite)

La solution de cette équation est donnée par

$$N(t) = N_0 \frac{Ke^{rt}}{K + N_0(e^{rt} - 1)}, \quad N_0 = N(0)$$

Une population à croissance logistique tend toujours vers K, la capacité d'accueil, quel que soit la densité de population d'origine ($N_0 > 0$).



2- Analyse d'erreurs

Une partie importante de l'analyse numérique consiste à contenir les effets des erreurs ainsi introduites, qui proviennent de trois sources principales :

- Les erreurs de modélisation : proviennent de l'étape de mathématisation d'un phénomène ;
- Les erreurs sur les données générée par la différence entre la valeur exacte x et la valeur mesurée x^* ou au fait que les données proviennent elle-même d'un calcul approché.
- Les erreurs de représentation sur ordinateur (Erreur d'arrondi) : la représentation des nombres sur ordinateur introduit souvent des erreurs (0,333 pour 1/3 par exemple), qui peuvent s'accumuler lorsqu'on effectue un très grand nombre d'opérations. Ces erreurs se propagent au fil des calculs et peuvent compromettre la précision des résultats.

- Les erreurs d'approximation ou de discrétisation (Les erreurs de troncature) : Ce sont les erreurs qu'on commet, par exemple, lorsqu'on calcule 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 (on parle alors quelquefois d'erreur de troncature).

Définition

Soit x un réel et x^* une valeur approchée de x.

- L'erreur absolue e est défini par $e = |x - x^*|$.

- L'erreur relative est $\left|\frac{x-x^*}{x}\right|$.

- Le pourcentage d'erreur est l'erreur relative multipliée par 100 ($\left|\frac{x-x^*}{x}\right| \times 100$).

Définition

Si l'erreur absolue vérifie

$$|x - x^*| \leq 0.5 \times 10^m$$

Alors le chiffre correspondant à la $m^{ième}$ puissance de 10 est dit significatif et tous ceux à sa gauche sont dit significatifs.

Exemple

Si $x = \pi$ *et* $x^* = 22/7 = 3.142857\cdots$ *donc*
$|x - x^*| = 0.00126\cdots \Longrightarrow |x - x^*| \leq 0.5 \times 10^{-2}$
Donc le chiffre des centième est significatif et on a en tout 3 chiffres significatif (3.14)

Exemple (Aspect fini de l'ordinateur)

Soit $a = 10^{20}, b = -10^{20}, c = 1$

- *Lorsque l'ordinateur effectue* $(a + b) + c$*, on obtient le résultat suivant : 1*
- *Lorsque l'ordinateur effectue* $a + (b + c)$*, on obtient le résultat suivant : 0*

Ceci est dû au fait que le stockage de a*,* b *et* c *sur l'ordinateur ne permet pas d'avoir une assez grande précision pour éviter dans le second cas que* $b + c = -10^{20}$*.*

Théorème

Soit b un entier strictement supérieur à 1. Tout nombre réel x non nul peut se représenter sous la forme

$$x = sgn(x)b^e \sum_{k=1}^{n} \frac{d_k}{b^k};$$

avec :

- $sgn(x)$ *est le signe de* x,
- *les* d_k *sont des entiers tels que* $0 \leq d_k \leq b-1$ *pour* $k \geq 1$ *et* $d_1 \neq 0$,
- $e \in \mathbb{Z}$.

De plus, cette écriture est unique

Définition

Le système de représentation machine des nombres réels le plus utilisé en calcul scientifique est celui de la représentation en virgule flottante. Un nombre $x \neq 0$ s'écrit $x = \pm m b^e$ avec :

- *b est la base de numération (entier supérieur ou égal à 2),*
- *$m = \sum_{k=1}^{n} d_k b^{-k}$ est la mantisse,*
- *n le nombre de chiffres de la mantisse (la précision),*
- *e est l'exposant, un entier relatif tel que $e_{min} \leq e \leq e_{max}$.*

Remarque

- *Numération décimale* $b = 10$ *et* $d_k \in \{0;1;2;3;4;5;6;7;8;9\}$
- *Numération binaire* $b = 2$ *et* $d_k \in \{0;1\}$
- *Numération octale* $b = 8$ *et* $d_k \in \{0;1;2;3;4;5;6;7\}$
- *Numération hexadécimale* $b = 16$ *et* $d_k \in \{0;1;\cdots;15\}$ *on utilise les lettres* $A, B, \cdots, F$ *pour représenter les chiffres hexadécimaux* $10, 11, \cdots, 15$.

Définition

l'ensemble $\mathbb{F} \subset \mathbb{R}$ *donné par :*

$$\begin{aligned} \mathbb{F} &= \left\{ x \in \mathbb{R} / x = \pm b^e \sum_{k=1}^{n} d_k b^{-k}, e_{min} \leq e \leq e_{max} \right\} \\ &= \{ x \in \mathbb{R} / x = \pm m b^e, e_{min} \leq e \leq e_{max} \} \end{aligned}$$

est un système de nombres à virgule flottante (floating point number system) noté par $\mathbb{F}(b, n, e_{min}, e_{max})$

3.2- Erreur d'arrondi

On suppose que $b = 10$, $x = sig(x)\ 0.d_1d_2\cdots d_t \times 10^n$ deux procédures permettent d'obtenir la représentation machine :

1. la troncature : $fl(x) = sig(x)\ 0.d_1d_2\cdots d_s \times 10^n$ avec $s < t$
2. L'arrondi :
$$\begin{cases} fl(x) = sig(x)\ 0.d_1d_2\cdots d_s \times 10^n & \text{si } 0 \leq d_{s+1} < 5 \\ fl(x) = (sig(x)\ 0.d_1d_2\cdots d_s + 10^{-s}) \times 10^n & \text{si } 5 \leq d_{s+1} < 10 \end{cases}$$
avec $s < t$

Exemple

pour $s = 3$

- *par troncature :* $\frac{1}{3}$ *devient* $0,333 \times 10^0$ *et* $\frac{2}{3}$ *devient* $0,666 \times 10^0$
- *par arrondi :* $\frac{1}{3}$ *devient* $0,333 \times 10^0$ *et* $\frac{2}{3}$ *devient* $0,667 \times 10^0$

4- Norme IEEE 754

C'est l'une des représentations standard des nombre flottants les plus utilisés

1. simple précision
   - Les nombres sont représentés sur des blocs de mémoire de taille 32 bits.
   - Chaque nombre se décompose de la manière suivante :
     - ✓ 24 bits pour la mantisse (dont 1 bit de signe)
     - ✓ 8 bits pour l'exposant (dont 1 bit de signe)
   - Le plus grand nombre que l'on puisse représenter est $x_{max} = 1,7 \times 10^{38}$
   - Le plus petit nombre positif que l'on puisse représenter est $x_{min} = 7,0 \times 10^{-46}$
   - Ce standard correspond à $\mathbb{F}(2, 24, -126, 127)$.

2. double précision
   - Les nombres sont représentés sur des blocs de mémoire de taille 64 bits.
   - Chaque nombre se décompose de la manière suivante :
     - ✓ 53 bits pour la mantisse (dont 1 bit de signe)
     - ✓ 11 bits pour l'exposant (dont 1 bit de signe)
   - Le plus grand nombre que l'on puisse représenter est $x_{max} = 9 \times 10^{307}$
   - Le plus petit nombre positif que l'on puisse représenter est $x_{min} = 2,5 \times 10^{-324}$
   - Ce standard correspond à $\mathbb{F}(2, 52, -1022, 1023)$.

5- Arithmétique flottante

Les opérations élémentaires (addition, soustraction, multiplication et division) en arithmétique flottante sont effectuées de la façon suivante :

$$\begin{aligned} x+y &\rightarrow fl(fl(x)+fl(y)) \\ x-y &\rightarrow fl(fl(x)-fl(y)) \\ x\div y &\rightarrow fl(fl(x)\div fl(y)) \\ x\times y &\rightarrow fl(fl(x)\times fl(y)) \end{aligned}$$

Quelques conséquences de cette représentation :

- $a+b=a$ si b est plus petit . Par exemple, soit une machine avec $t=2$, $a=0.63\times 10^{1}$ et $b=0.82\times 10^{-4}$.
  Pour faire l'opération, on réduit au même exposant, soit $a+b=0.63\times 10^{1}+0.0000082\times 10^{1}=0.6300082\times 10^{1}$, et ce dernier nombre est représenté par $fl(a+b)=0.63\times 10^{1}$ car t $=2$.
  Conclusion : $a+b=a$ et $b\neq 0$.

5- Arithmétique flottante

Les opérations élémentaires (addition, soustraction, multiplication et division) en arithmétique flottante sont effectuées de la façon suivante :

$$\begin{aligned} x+y &\rightarrow fl(fl(x)+fl(y)) \\ x-y &\rightarrow fl(fl(x)-fl(y)) \\ x\div y &\rightarrow fl(fl(x)\div fl(y)) \\ x\times y &\rightarrow fl(fl(x)\times fl(y)) \end{aligned}$$

Quelques conséquences de cette représentation :

- $a+b=a$ si b est plus petit . Par exemple, soit une machine avec $t=2$, $a=0.63\times 10^1$ et $b=0.82\times 10^{-4}$.
  Pour faire l'opération, on réduit au même exposant, soit $a+b=0.63\times 10^1+0.0000082\times 10^1=0.6300082\times 10^1$, et ce dernier nombre est représenté par $fl(a+b)=0.63\times 10^1$ car t $=2$.
  Conclusion : $a+b=a$ et $b\neq 0$.

- L'addition des nombres flottants n'est pas associative :
  Pour $t = 4$ (quatre chiffres significatifs), soit
  $x = 0.6724 \times 10^3, y = 0.7215 \times 10^{-1}$ et $z = 0.5345 \times 10^1$, on a

$$\begin{aligned} fl(x+y) &= fl(0.6724 \times 10^3 + 0.7215 \times 10^{-1}) \\ &= fl(0.6724 \times 10^3 + 0.00007215 \times 10^3) \\ &= fl(0.67247215 \times 10^3) \\ &= 0.6724 \times 10^3 \end{aligned}$$

donc

$$\begin{aligned} fl((x+y)+z) &= fl(0.6724 \times 10^3 + 0.5345 \times 10^1) \\ &= fl(0.6724 \times 10^3 + 0.005345 \times 10^3) \\ &= fl(0.677745 \times 10^3) \\ &= 0.6777 \times 10^3 \end{aligned}$$

- L'addition des nombres flottants n'est pas associative :
  Pour $t = 4$ (quatre chiffres significatifs), soit
  $x = 0.6724 \times 10^3, y = 0.7215 \times 10^{-1}$ et $z = 0.5345 \times 10^1$, on a

$$\begin{aligned} fl(x+y) &= fl(0.6724 \times 10^3 + 0.7215 \times 10^{-1}) \\ &= fl(0.6724 \times 10^3 + 0.00007215 \times 10^3) \\ &= fl(0.67247215 \times 10^3) \\ &= 0.6724 \times 10^3 \end{aligned}$$

  donc

$$\begin{aligned} fl((x+y)+z) &= fl(0.6724 \times 10^3 + 0.5345 \times 10^1) \\ &= fl(0.6724 \times 10^3 + 0.005345 \times 10^3) \\ &= fl(0.677745 \times 10^3) \\ &= 0.6777 \times 10^3 \end{aligned}$$

$$\begin{aligned} fl(y+z) &= fl(0.7215 \times 10^{-1} + 0.5345 \times 10^{1}) \\ &= fl(0.007215 \times 10^{1} + 0.5345 \times 10^{1}) \\ &= fl(0.541715 \times 10^{1}) = 0.5417 \times 10^{1} \end{aligned}$$

donc

$$\begin{aligned} fl(x+(y+z)) &= fl(0.6724 \times 10^{3} + 0.5417 \times 10^{1}) \\ &= fl(0.6724 \times 10^{3} + 0.005417 \times 10^{3}) \\ &= fl(0.677817 \times 10^{3}) \\ &= 0.6778 \times 10^{3} \end{aligned}$$

D'où

$$fl((x+y)+z) \neq fl(x+(y+z))$$

$$\begin{aligned} fl(y+z) &= fl(0.7215 \times 10^{-1} + 0.5345 \times 10^{1}) \\ &= fl(0.007215 \times 10^{1} + 0.5345 \times 10^{1}) \\ &= fl(0.541715 \times 10^{1}) = 0.5417 \times 10^{1} \end{aligned}$$

donc

$$\begin{aligned} fl(x+(y+z)) &= fl(0.6724 \times 10^{3} + 0.5417 \times 10^{1}) \\ &= fl(0.6724 \times 10^{3} + 0.005417 \times 10^{3}) \\ &= fl(0.677817 \times 10^{3}) \\ &= 0.6778 \times 10^{3} \end{aligned}$$

D'où

$$fl((x+y)+z) \neq fl(x+(y+z))$$

$$\begin{aligned} fl(y+z) &= fl(0.7215 \times 10^{-1} + 0.5345 \times 10^{1}) \\ &= fl(0.007215 \times 10^{1} + 0.5345 \times 10^{1}) \\ &= fl(0.541715 \times 10^{1}) = 0.5417 \times 10^{1} \end{aligned}$$

donc

$$\begin{aligned} fl(x+(y+z)) &= fl(0.6724 \times 10^{3} + 0.5417 \times 10^{1}) \\ &= fl(0.6724 \times 10^{3} + 0.005417 \times 10^{3}) \\ &= fl(0.677817 \times 10^{3}) \\ &= 0.6778 \times 10^{3} \end{aligned}$$

D'où

$$fl((x+y)+z) \neq fl(x+(y+z))$$

- Pour $t = 3$ (trois chiffres significatifs), considérons l'opération

$$\begin{aligned} fl(854 \times (251 + 852)) &= fl(0.854 \times 10^3) \\ &\quad \times fl((0.251 \times 10^3 + 0.852 \times 10^3)) \\ &= fl(0.854 \times 10^3) \times fl((1.103 \times 10^3)) \\ &= fl(0.854 \times 10^3) \times fl((0.110 \times 10^4)) \\ &= fl(0.09394 \times 10^7) \\ &= fl(0.9394 \times 10^6) \\ &= 0.939 \times 10^6 \end{aligned}$$

$$\begin{aligned} fl(854 \times 251 + 854 \times 852)) &= fl(0.854 \times 10^3 \times 0.251 \times 10^3) \\ &\quad + fl(0.854 \times 10^3 \times 0.852 \times 10^3) \\ &= fl(0.214354 \times 10^6) \\ &\quad + fl(0.727608 \times 10^6) \\ &= fl(0.214 \times 10^6) + fl(0.727 \times 10^6) \\ &= fl(0.941 \times 10^6) \\ &= 0.941 \times 10^6 \end{aligned}$$

- Pour $t = 3$ (trois chiffres significatifs), considérons l'opération

$$\begin{aligned}
fl(854 \times (251 + 852)) &= fl(0.854 \times 10^3) \\
&\quad \times fl((0.251 \times 10^3 + 0.852 \times 10^3)) \\
&= fl(0.854 \times 10^3) \times fl((1.103 \times 10^3)) \\
&= fl(0.854 \times 10^3) \times fl((0.110 \times 10^4)) \\
&= fl(0.09394 \times 10^7) \\
&= fl(0.9394 \times 10^6) \\
&= 0.939 \times 10^6 \\
fl(854 \times 251 + 854 \times 852)) &= fl(0.854 \times 10^3 \times 0.251 \times 10^3) \\
&\quad + fl(0.854 \times 10^3 \times 0.852 \times 10^3) \\
&= fl(0.214354 \times 10^6) \\
&\quad + fl(0.727608 \times 10^6) \\
&= fl(0.214 \times 10^6) + fl(0.727 \times 10^6) \\
&= fl(0.941 \times 10^6) \\
&= 0.941 \times 10^6
\end{aligned}$$

Donc la distributivité de la multiplication par rapport à l'addition n'est pas respectée en arithmétique flottante.

- Même phénomène pour la soustraction, division, multiplication $\cdots$
- Phénomènes de compensation :

Exemple

soit les deux fonctions suivantes :

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

on a

$$f = g$$

Mais sous MATLAB on a obtenu les résultats suivants :

Donc la distributivité de la multiplication par rapport à l'addition n'est pas respectée en arithmétique flottante.

- Même phénomène pour la soustraction, division, multiplication $\cdots$
- Phénomènes de compensation :

Exemple

soit les deux fonctions suivantes :

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

on a

$$f = g$$

Mais sous MATLAB on a obtenu les résultats suivants :

```
>>  f=inline('(sqrt(x+1)-sqrt(x))','x')

f =

     Inline function:
     f(x) = (sqrt(x+1)-sqrt(x))

>> g=inline('(1./(sqrt(x+1)+sqrt(x)))','x')

g =

     Inline function:
     g(x) = (1./(sqrt(x+1)+sqrt(x)))

>> sprintf('f(10^9)=%e       g(10^9)=%e',f(10^9),g(10^9))

ans =

    'f(10^9)=1.581139e-05       g(10^9)=1.581139e-05'

>> sprintf('f(10^16)=%e       g(10^16)=%e',f(10^16),g(10^16))

ans =

    'f(10^16)=0.000000e+00       g(10^16)=5.000000e-09'
```

6- Exercices

Exercice

En arithmétique flottante illustrer la non-validité de :

1. *La loi d'associativité pour*
   $x = 0,23371258 \times 10^{-4}$, $y = 0,33678429 \times 10^{2}$ *et*
   $z = -0,33677811 \times 10^{2}$ *avec 8 chiffres significatifs*
2. *La loi de distributivité* $x = 122, y = 333$ *et* $z = 395$
   avec 3 chiffres significatifs et arrondi

Corrigé

1. On a $x = 0,23371258 \times 10^{-4}; y = 0,33678429 \times 10^{2}$
   $z = -0,33677811 \times 10^{2}$

$$\begin{aligned} fl(x+y) &= fl(0,23371258 \times 10^{-4} + 0,33678429 \times 10^{2}) \\ &= fl(0,0000023371258 \times 10^{2} + 0,33678429 \times 10^{2}) \\ &= 0,33678452 \times 10^{2} \end{aligned}$$

6- Exercices

Exercice

En arithmétique flottante illustrer la non-validité de :

1. *La loi d'associativité pour* $x = 0,23371258 \times 10^{-4}$, $y = 0,33678429 \times 10^{2}$ *et* $z = -0,33677811 \times 10^{2}$ *avec 8 chiffres significatifs*
2. *La loi de distributivité* $x = 122, y = 333$ *et* $z = 395$ *avec 3 chiffres significatifs et arrondi*

Corrigé

1. On a $x = 0,23371258 \times 10^{-4}; y = 0,33678429 \times 10^{2}$ $z = -0,33677811 \times 10^{2}$

$$\begin{aligned} fl(x+y) &= fl(0,23371258 \times 10^{-4} + 0,33678429 \times 10^{2}) \\ &= fl(0,00000023371258 \times 10^{2} + 0,33678429 \times 10^{2}) \\ &= 0,33678452 \times 10^{2} \end{aligned}$$

Corrigé

$$\begin{aligned} fl(fl(x+y)+z) &= fl(0,33678452 \times 10^2 - 0,33677811 \times 10^2) \\ &= fl(0,00000641 \times 10^2) \\ &= 0,641 \times 10^{-3}. \end{aligned}$$

Par ailleurs :

$$\begin{aligned} fl(y+z) &= fl(0,33678429 \times 10^2 - 0,33677811 \times 10^2) \\ &= fl(0,00000618 \times 10^2) \\ &= 0,618 \times 10^{-3} \end{aligned}$$

donc

$$\begin{aligned} fl(x+(y+z)) &= fl(0,023371258 \times 10^{-3} + 0,61800000 \times 10^{-3}) \\ &= 0,64137126 \times 10^{-3} \end{aligned}$$

Corrigé

$$\begin{aligned} fl(fl(x+y)+z) &= fl(0,33678452 \times 10^2 - 0,33677811 \times 10^2) \\ &= fl(0,00000641 \times 10^2) \\ &= 0,641 \times 10^{-3}. \end{aligned}$$

Par ailleurs :

$$\begin{aligned} fl(y+z) &= fl(0,33678429 \times 10^2 - 0,33677811 \times 10^2) \\ &= fl(0,00000618 \times 10^2) \\ &= 0,618 \times 10^{-3} \end{aligned}$$

donc

$$\begin{aligned} fl(x+(y+z)) &= fl(0,023371258 \times 10^{-3} + 0,61800000 \times 10^{-3}) \\ &= 0,64137126 \times 10^{-3} \end{aligned}$$

Corrigé

$$\begin{aligned} fl(fl(x+y)+z) &= fl(0,33678452 \times 10^2 - 0,33677811 \times 10^2) \\ &= fl(0,00000641 \times 10^2) \\ &= 0,641 \times 10^{-3}. \end{aligned}$$

Par ailleurs :

$$\begin{aligned} fl(y+z) &= fl(0,33678429 \times 10^2 - 0,33677811 \times 10^2) \\ &= fl(0,00000618 \times 10^2) \\ &= 0,618 \times 10^{-3} \end{aligned}$$

donc

$$\begin{aligned} fl(x+(y+z)) &= fl(0,023371258 \times 10^{-3} + 0,61800000 \times 10^{-3}) \\ &= 0,64137126 \times 10^{-3} \end{aligned}$$

2. Pour $t = 3$ (trois chiffres significatifs), considérons l'opération

$$\begin{aligned}
fl(122 \times (333 + 695)) &= fl(0.122 \times 10^3) \\
&\quad \times fl((0.333 \times 10^3 + 0.695 \times 10^3)) \\
&= fl(0.122 \times 10^3) \times fl((1.028 \times 10^3)) \\
&= fl(0.122 \times 10^3) \times fl((0.103 \times 10^4)) \\
&= fl(0.012566 \times 10^7) \\
&= fl(0.12566 \times 10^6) \\
&= 0.126 \times 10^6
\end{aligned}$$

$$\begin{aligned}
fl(122 \times 333 + 122 \times 695)) &= fl(0.122 \times 10^3 \times 0.333 \times 10^3) \\
&\quad + fl(0.122 \times 10^3 \times 0.695 \times 10^3) \\
&= fl(0,040626 \times 10^6) + fl(0,08479 \times 10^6) \\
&= fl(0,40626 \times 10^5) + fl(0,8479 \times 10^5) \\
&= fl(0.406 \times 10^5) + fl(0.848 \times 10^5) \\
&= fl(1,254 \times 10^5) \\
&= 0.125 \times 10^6
\end{aligned}$$

2. Pour $t = 3$ (trois chiffres significatifs), considérons l'opération

$$\begin{aligned}
fl(122 \times (333 + 695)) &= fl(0.122 \times 10^3) \\
&\quad \times fl((0.333 \times 10^3 + 0.695 \times 10^3)) \\
&= fl(0.122 \times 10^3) \times fl((1.028 \times 10^3)) \\
&= fl(0.122 \times 10^3) \times fl((0.103 \times 10^4)) \\
&= fl(0.012566 \times 10^7) \\
&= fl(0.12566 \times 10^6) \\
&= 0.126 \times 10^6 \\
fl(122 \times 333 + 122 \times 695)) &= fl(0.122 \times 10^3 \times 0.333 \times 10^3) \\
&\quad + fl(0.122 \times 10^3 \times 0.695 \times 10^3) \\
&= fl(0,040626 \times 10^6) + fl(0,08479 \times 10^6) \\
&= fl(0,40626 \times 10^5) + fl(0,8479 \times 10^5) \\
&= fl(0.406 \times 10^5) + fl(0.848 \times 10^5) \\
&= fl(1,254 \times 10^5) \\
&= 0.125 \times 10^6
\end{aligned}$$

Exercice

Soit l'équation du second degré $ax^2 + bx + c = 0$ *avec* $a \neq 0$.
On suppose que le discriminant $\Delta > 0$

1. *Vérifier que* $x_1 \times x_2 = \frac{c}{a}$
2. *Supposons que les calculs soient effectués avec 10 chiffres significatifs* $(t = 10)$, *trouver les racines de* $x^2 - 1634x + 2 = 0$
3. *Calculer* x_2 *on fonction de* x_1.
4. *Que remarquez-vous ?*

Corrigé

1. On suppose que le discriminant $\Delta > 0$ donc
   $x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$ et $x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a}$
   $\Longrightarrow x_1 \times x_2 = \frac{(-b+\sqrt{b^2-4ac})(-b-\sqrt{b^2-4ac})}{4a^2} = \frac{b^2-(b^2-4ac)}{4a^2} = \frac{c}{a}$
2. on a $\Delta' = 817^2 - 2 = 667487 \Longrightarrow \sqrt{\Delta'} = 816;9987760$ d'où les solutions : $x_1 = 817 + 816,9987760 = 1633,998776$ et $x_2 = 817 - 816,9987760 = 0,0012240$
3. On a $x_1 x_2 = \frac{c}{a} = 2 \Longrightarrow x_2 = \frac{2}{x_1} = \frac{2}{1633,998776} = 0,001223991125$
4. Dans la méthode de la question 2 on a perte de 5 chiffres significatifs sur x_2

Corrigé

1. On suppose que le discriminant $\Delta > 0$ donc
   $x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$ et $x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a}$
   $\Longrightarrow x_1 \times x_2 = \frac{(-b+\sqrt{b^2-4ac})(-b-\sqrt{b^2-4ac})}{4a^2} = \frac{b^2-(b^2-4ac)}{4a^2} = \frac{c}{a}$
2. on a $\Delta' = 817^2 - 2 = 667487 \Longrightarrow \sqrt{\Delta'} = 816;9987760$ d'où les solutions : $x_1 = 817 + 816,9987760 = 1633,998776$ et $x_2 = 817 - 816,9987760 = 0,0012240$
3. On a $x_1 x_2 = \frac{c}{a} = 2 \Longrightarrow x_2 = \frac{2}{x_1} = \frac{2}{1633,998776} = 0,001223991125$
4. Dans la méthode de la question 2 on a perte de 5 chiffres significatifs sur x_2

Corrigé

1. On suppose que le discriminant $\Delta > 0$ donc
   $x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$ et $x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a}$
   $\Longrightarrow x_1 \times x_2 = \frac{(-b+\sqrt{b^2-4ac})(-b-\sqrt{b^2-4ac})}{4a^2} = \frac{b^2-(b^2-4ac)}{4a^2} = \frac{c}{a}$
2. on a $\Delta' = 817^2 - 2 = 667487 \Longrightarrow \sqrt{\Delta'} = 816;9987760$ d'où les solutions : $x_1 = 817 + 816,9987760 = 1633,998776$ et $x_2 = 817 - 816,9987760 = 0,0012240$
3. On a $x_1 x_2 = \frac{c}{a} = 2 \Longrightarrow x_2 = \frac{2}{x_1} = \frac{2}{1633,998776} = 0,001223991125$
4. Dans la méthode de la question 2 on a perte de 5 chiffres significatifs sur x_2