

# Méthodes numériques et programmation

Souad EL Bernoussi et Awatif Sayah

Université Mohammed V - Agdal.  
Faculté des Sciences  
Département de Mathématiques

# Analyse numérique ?

Etude et la construction d'algorithmes (du nom du mathématicien Al Khawarizmi) de résolution numérique d'un problème donné.

En pratique, l'Analyse Numérique se propose d'étudier les propriétés mathématiques des algorithmes et leur mise en oeuvre (programmation).

# Analyse numérique ?

Etude et la construction d'algorithmes (du nom du mathématicien Al Khawarizmi) de résolution numérique d'un problème donné.

En pratique, l'Analyse Numérique se propose d'étudier les propriétés mathématiques des algorithmes et leur mise en oeuvre (programmation).

# Objectifs ?

L'objectif de l'analyse numérique est de :

concevoir

et d'étudier

des méthodes de résolution de certains

problèmes mathématiques (en général issus de la modélisation de problèmes 'réels'), et dont on

cherche à calculer la solution ou son

approximation à l'aide d'un ordinateur.

# Objectifs ?

L'objectif de l'analyse numérique est de :  
concevoir

et d'étudier

des méthodes de résolution de certains problèmes mathématiques (en général issus de la modélisation de problèmes 'réels'), et dont on cherche à calculer la solution ou son approximation à l'aide d'un ordinateur.

# Objectifs ?

L'objectif de l'analyse numérique est de :  
concevoir  
et d'étudier

des méthodes de résolution de certains problèmes mathématiques (en général issus de la modélisation de problèmes 'réels'), et dont on cherche à calculer la solution ou son approximation à l'aide d'un ordinateur.

# Objectifs ?

L'objectif de l'analyse numérique est de :  
concevoir  
et d'étudier  
des méthodes de résolution de certains  
problèmes mathématiques (en général issus de  
la modélisation de problèmes 'réels'), et dont on  
cherche à calculer la solution ou son  
approximation à l'aide d'un ordinateur.

# Enjeux de l'analyse numérique?

Résoudre des problèmes :

- que l'on ne sait pas résoudre autrement
- 'mieux' qu'on ne le faisait avant :
  - plus précisément,
  - moins cher...

# Enjeux de l'analyse numérique?

## Etique ( 'Objectifs' ) de l'analyse numérique

- Plus vite :
  - complexité des algorithmes
  - complexité des problèmes
- Plus précis :
  - erreur d'arrondi (liées à la machine)
  - erreur d'approximation (liées à l'algorithme)
- Plus fiable :
  - stabilité d'un algorithme
- Facile à programmer :
  - comprendre pour mieux réutiliser

# Enjeux de l'analyse numérique?

## Etique ( 'Objectifs' ) de l'analyse numérique

- Plus vite :
  - complexité des algorithmes
  - complexité des problèmes
- Plus précis :
  - erreur d'arrondi (liées à la machine)
  - erreur d'approximation (liées à l'algorithme)
- Plus fiable :
  - stabilité d'un algorithme
- Facile à programmer :
  - comprendre pour mieux réutiliser

# Enjeux de l'analyse numérique?

## Etique ( 'Objectifs' ) de l'analyse numérique

- Plus vite :
  - complexité des algorithmes
  - complexité des problèmes
- Plus précis :
  - erreur d'arrondi (liées à la machine)
  - erreur d'approximation (liées à l'algorithme)
- Plus fiable :
  - stabilité d'un algorithme
- Facile à programmer :
  - comprendre pour mieux réutiliser

# Enjeux de l'analyse numérique?

## Etique ( 'Objectifs' ) de l'analyse numérique

- Plus vite :
  - complexité des algorithmes
  - complexité des problèmes
- Plus précis :
  - erreur d'arrondi (liées à la machine)
  - erreur d'approximation (liées à l'algorithme)
- Plus fiable :
  - stabilité d'un algorithme
- Facile à programmer :
  - comprendre pour mieux réutiliser

# 1ère Conclusion

'La confiance aveugle dans ce que l'on appelle les résultats fournis par l'ordinateur peut être la cause d'erreurs qui peuvent coûter très chères'

Alors que faire ?

# 1ère Conclusion

'La confiance aveugle dans ce que l'on appelle les résultats fournis par l'ordinateur peut être la cause d'erreurs qui peuvent coûter très chères'

Alors que faire ?

# Sources d'Erreurs

Il y a en 3 catégories :

- Erreurs liés à la machine,
- Erreurs à la méthode (algorithme),
- Erreurs sur les donnés (résultat d'un calcul approché, d'une mesure physique,...)

# Arithmétique des calculateurs et Sources d'erreurs

Si sophistiqué qu'il soit , un calculateur ne peut fournir que des réponses approximatives.

Les approximations utilisées dépendent à la fois des contraintes physiques (espace mémoire, vitesse de l'horloge...) et du choix des méthodes retenues par le concepteur du programme .

# Arithmétique des calculateurs et Sources d'erreurs

Le but de ce chapitre est de prendre connaissance de l'impact de ces contraintes et de ces choix méthodologiques.

Dans certains cas il doit être pris en compte dans l'analyse des résultats dont une utilisation erronée pourrait être coûteuse.

# Arithmétique des calculateurs et Sources d'erreurs

Le but de ce chapitre est de prendre connaissance de l'impact de ces contraintes et de ces choix méthodologiques.

Dans certains cas il doit être pris en compte dans l'analyse des résultats dont une utilisation erronée pourrait être coûteuse.

# Arithmétique des calculateurs et Sources d'erreurs

La première contrainte est que le système numérique de l'ordinateur est discret, c'est à dire qu'il ne comporte qu'un nombre fini de nombres;

Il en découle que tous les calculs sont entachés d'erreurs.

# Arithmétique des calculateurs et Sources d'erreurs

La première contrainte est que le système numérique de l'ordinateur est discret, c'est à dire qu'il ne comporte qu'un nombre fini de nombres;

Il en découle que tous les calculs sont entachés d'erreurs.

# Arithmétique des calculateurs et Sources d'erreurs

La première contrainte est que le système numérique de l'ordinateur est discret, c'est à dire qu'il ne comporte qu'un nombre fini de nombres;  
Il en découle que tous les calculs sont entachés d'erreurs.

# Evaluation de l'erreur

Rappelons d'abord quelques notions de base ;  
Si  $X$  est une quantité à calculer et  $X^*$  la valeur calculée, on dit que :

- $X - X^*$  est l'erreur et  $|E| = |X - X^*|$  est l'erreur absolue.

**Exemple :**

*Si  $X = 2.224$  et  $X^* = 2.223$  alors l'erreur absolue*

$$|E| = |X - X^*| = 2.224 - 2.223 = 0.001$$

# Evaluation de l'erreur

Rappelons d'abord quelques notions de base ;  
Si  $X$  est une quantité à calculer et  $X^*$  la valeur calculée, on dit que :

- $X - X^*$  est l'erreur et  $|E| = |X - X^*|$  est l'erreur absolue.

**Exemple :**

*Si  $X = 2.224$  et  $X^* = 2.223$  alors l'erreur absolue*

$$|E| = |X - X^*| = 2.224 - 2.223 = 0.001$$

# Evaluation de l'erreur

Rappelons d'abord quelques notions de base ;  
Si  $X$  est une quantité à calculer et  $X^*$  la valeur calculée, on dit que :

- $X - X^*$  est l'erreur et  $|E| = |X - X^*|$  est l'erreur absolue.

## **Exemple :**

*Si  $X = 2.224$  et  $X^* = 2.223$  alors l'erreur absolue*

$$|E| = |X - X^*| = 2.224 - 2.223 = 0.001$$

- $E_r = \left| \frac{X - X^*}{X_r} \right|$  est l'erreur relative,

$X_r \neq 0$ .  $X_r$  est une valeur de référence pour  $X$ . En général, on prend  $X_r = X$ .

### Exemple :

Si  $X = 2.224$  et  $X^* = 2.223$

alors, si on prend  $X_r = X$ , l'erreur relative

$$E_r = \left| \frac{X - X^*}{X_r} \right| = \frac{|X - X^*|}{|X|} = \frac{0.001}{2.224} = 4.496 \times 10^{-4}$$

- $E_r = \left| \frac{X-X^*}{X_r} \right|$  est l'erreur relative,  
 $X_r \neq 0$ .  $X_r$  est une valeur de référence pour  $X$ . En général, on prend  $X_r = X$ .

### Exemple :

Si  $X = 2.224$  et  $X^* = 2.223$

alors, si on prend  $X_r = X$ , l'erreur relative

$$E_r = \left| \frac{X-X^*}{X_r} \right| = \frac{|X-X^*|}{|X|} = \frac{0.001}{2.224} = 4.496 \times 10^{-4}$$

- $E_r = \left| \frac{X - X^*}{X_r} \right|$  est l'erreur relative,  
 $X_r \neq 0$ .  $X_r$  est une valeur de référence pour  
 $X$ . En général, on prend  $X_r = X$ .

### Exemple :

Si  $X = 2.224$  et  $X^* = 2.223$

*alors, si on prend  $X_r = X$ , l'erreur relative*

$$E_r = \left| \frac{X - X^*}{X_r} \right| = \frac{|X - X^*|}{|X|} = \frac{0.001}{2.224} = 4.496 \times 10^{-4}$$

- $E_r = \left| \frac{X-X^*}{X_r} \right|$  est l'erreur relative,  
 $X_r \neq 0$ .  $X_r$  est une valeur de référence pour  
 $X$ . En général ,on prend  $X_r = X$ .

### Exemple :

Si  $X = 2.224$  et  $X^* = 2.223$

alors , si on prend  $X_r = X$  , l'erreur relative

$$E_r = \left| \frac{X-X^*}{X_r} \right| = \frac{|X-X^*|}{|X|} = \frac{0.001}{2.224} = 4.496 \times 10^{-4}$$

Cependant, si  $X$  est la valeur d'une fonction  $F(t)$  avec  $a \leq t \leq b$ , on choisira parfois une valeur de référence globale pour toutes les valeurs de  $t$ .

**Exemple :**

Si  $X = \sin(t)$  avec  $0 \leq t \leq \frac{\pi}{4}$ , on pourra prendre

$$X_r = \frac{\sqrt{2}}{2} = \sup_{0 \leq t \leq \frac{\pi}{4}} \sin(t).$$

En général , on ne connait pas le signe de l'erreur de sorte que l'on considère les erreurs absolues et les erreurs relatives absolues.

Les opérations élémentaires propagent des erreurs.

Dans la pratique, on considère que :

1) L'erreur absolue sur une somme est la somme des erreurs absolues.

2) L'erreur relative sur un produit ou un quotient est la somme des erreurs relatives.

En général , on ne connaît pas le signe de l'erreur de sorte que l'on considère les erreurs absolues et les erreurs relatives absolues. Les opérations élémentaires propagent des erreurs.

Dans la pratique, on considère que :

1) L'erreur absolue sur une somme est la somme des erreurs absolues.

2) L'erreur relative sur un produit ou un quotient est la somme des erreurs relatives.

En général , on ne connaît pas le signe de l'erreur de sorte que l'on considère les erreurs absolues et les erreurs relatives absolues. Les opérations élémentaires propagent des erreurs.

Dans la pratique, on considère que :

1) L'erreur absolue sur une somme est la somme des erreurs absolues.

2) L'erreur relative sur un produit ou un quotient est la somme des erreurs relatives.

En général , on ne connaît pas le signe de l'erreur de sorte que l'on considère les erreurs absolues et les erreurs relatives absolues. Les opérations élémentaires propagent des erreurs.

Dans la pratique, on considère que :

- 1) L'erreur absolue sur une somme est la somme des erreurs absolues.
- 2) L'erreur relative sur un produit ou un quotient est la somme des erreurs relatives.

On peut estimer l'effet d'une erreur  $E$  sur l'argument  $x$  d'une fonction  $f(x)$  au moyen de la dérivée de  $f(x)$ .

En effet  $f(x + E) \simeq f(x) + Ef'(x)$

On peut estimer l'effet d'une erreur  $E$  sur l'argument  $x$  d'une fonction  $f(x)$  au moyen de la dérivée de  $f(x)$ .

En effet  $f(x + E) \simeq f(x) + Ef'(x)$

## Exemple :

*Calculer la valeur de  $(11111111)^2$*

*La valeur fournie par une petite calculatrice à cinq chiffres est  $1,2345 \times 10^{14}$*

*Mais la réponse exacte est 123456787654321.*

*La machine a donc tronqué le résultat à 5 chiffres et l'erreur absolue est de  $6 * 10^9$ .*

*L'erreur relative est de 0.005% .*

*Cet exemple montre qu'il faut établir clairement l'objectif visé.*

## Exemple :

*Calculer la valeur de  $(11111111)^2$*

*La valeur fournie par une petite calculatrice à cinq chiffres est  $1,2345 \times 10^{14}$*

*Mais la réponse exacte est 123456787654321.*

*La machine a donc tronqué le résultat à 5 chiffres et l'erreur absolue est de  $6 * 10^9$ .*

*L'erreur relative est de 0.005% .*

*Cet exemple montre qu'il faut établir clairement l'objectif visé.*

## Exemple :

*Calculer la valeur de  $(11111111)^2$*

*La valeur fournie par une petite calculatrice à cinq chiffres est  $1,2345 \times 10^{14}$*

*Mais la réponse exacte est 123456787654321.*

*La machine a donc tronqué le résultat à 5 chiffres et l'erreur absolue est de  $6 * 10^9$ .*

*L'erreur relative est de 0.005% .*

*Cet exemple montre qu'il faut établir clairement l'objectif visé.*

## Exemple :

*Calculer la valeur de  $(11111111)^2$*

*La valeur fournie par une petite calculatrice à cinq chiffres est  $1,2345 \times 10^{14}$*

*Mais la réponse exacte est 123456787654321.*

*La machine a donc tronqué le résultat à 5 chiffres et l'erreur absolue est de  $6 * 10^9$ .*

*L'erreur relative est de 0.005% .*

*Cet exemple montre qu'il faut établir clairement l'objectif visé.*

## Exemple :

*Calculer la valeur de  $(11111111)^2$*

*La valeur fournie par une petite calculatrice à cinq chiffres est  $1,2345 \times 10^{14}$*

*Mais la réponse exacte est 123456787654321.*

*La machine a donc tronqué le résultat à 5 chiffres et l'erreur absolue est de  $6 * 10^9$ .*

*L'erreur relative est de 0.005% .*

*Cet exemple montre qu'il faut établir clairement l'objectif visé.*

## Exemple :

*Calculer la valeur de  $(11111111)^2$*

*La valeur fournie par une petite calculatrice à cinq chiffres est  $1,2345 \times 10^{14}$*

*Mais la réponse exacte est 123456787654321.*

*La machine a donc tronqué le résultat à 5 chiffres et l'erreur absolue est de  $6 * 10^9$ .*

*L'erreur relative est de 0.005% .*

Cet exemple montre qu'il faut établir clairement l'objectif visé.

Cet objectif est double ;

1) Nous voulons un bon ordre de grandeur (ici  $10^{14}$ ) et avoir le maximum de décimales exactes,

2) Ce maximum ne peut excéder la longueur des mots permis par la machine et dépend donc de la machine

Cet objectif est double ;

1) Nous voulons un bon ordre de grandeur (ici  $10^{14}$ ) et avoir le maximum de décimales exactes,

2) Ce maximum ne peut excéder la longueur des mots permis par la machine et dépend donc de la machine

# La mémoire de l'ordinateur : le stockage des nombres

La mémoire d'un ordinateur est formée d'un certain nombre d'unités adressables appelées **OCTETS** .

Un ordinateur moderne contient des millions voir des milliards d'octets.

Les nombres sont stockés dans un ordinateur comme **ENTIERS** ou **REELS**.

# La mémoire de l'ordinateur : le stockage des nombres

La mémoire d'un ordinateur est formée d'un certain nombre d'unités adressables appelées OCTETS .

Un ordinateur moderne contient des millions voir des milliards d'octets.

Les nombres sont stockés dans un ordinateur comme ENTIERS ou REELS.

# La mémoire de l'ordinateur : le stockage des nombres

La mémoire d'un ordinateur est formée d'un certain nombre d'unités adressables appelées OCTETS .

Un ordinateur moderne contient des millions voir des milliards d'octets.

Les nombres sont stockés dans un ordinateur comme ENTIERS ou REELS.

## Les nombres entiers :

Les nombres entiers sont ceux que l'on utilise d'habitude sauf que le plus grand nombre représentable dépend du nombre d'octets utilisés:

- avec deux (2) octets, on peut représenter les entiers compris entre

$-32768$  et  $32767$

- avec quatre (4) octets on peut représenter les entiers compris entre

$-2147483648$  et  $2147483647$

## Les nombres entiers :

Les nombres entiers sont ceux que l'on utilise d'habitude sauf que le plus grand nombre représentable dépend du nombre d'octets utilisés:

- avec deux (2) octets, on peut représenter les entiers compris entre

–32768 et 32767

- avec quatre (4) octets on peut représenterr les entiers compris entre

–2147483648 et 2147483647

## Les nombres entiers :

Les nombres entiers sont ceux que l'on utilise d'habitude sauf que le plus grand nombre représentable dépend du nombre d'octets utilisés:

- avec deux (2) octets, on peut représenter les entiers compris entre

–32768 et 32767

- avec quatre (4) octets on peut représenterr les entiers compris entre

–2147483648 et 2147483647

# Les nombres réels

Dans la mémoire d'un ordinateur, les nombres réels sont représentés en notation flottante.

Cette notation a été introduite pour garder une erreur relative à peu près constante; quelque soit l'ordre de grandeur du nombre qu'on manipule.

# Les nombres réels

Dans la mémoire d'un ordinateur, les nombres réels sont représentés en notation flottante.

Cette notation a été introduite pour garder une erreur relative à peu près constante; quelque soit l'ordre de grandeur du nombre qu'on manipule.

En notation flottante, un nombre a la forme :

$$x = \pm Y \times b^e$$

$b$  est la base du système numérique utilisé

$Y$  est la mantisse : une suite de  $s$  entier

$y_1 y_2 \dots y_s$  avec  $y_1 \neq 0$  si  $x \neq 0$  et  $0 \leq y_i \leq (b - 1)$

$e$  est l'exposant (un nombre entier relatif)

La norme choisie est celle où la mantisse est comprise entre 0 et 1 et où le premier chiffre après la virgule est différent de zéro.

En notation flottante, un nombre a la forme :

$$x = \pm Y \times b^e$$

$b$  est la base du système numérique utilisé

$Y$  est la mantisse : une suite de  $s$  entier

$y_1 y_2 \dots y_s$  avec  $y_1 \neq 0$  si  $x \neq 0$  et  $0 \leq y_i \leq (b - 1)$

$e$  est l'exposant (un nombre entier relatif)

La norme choisie est celle où la mantisse est comprise entre 0 et 1 et où le premier chiffre après la virgule est différent de zéro.

En notation flottante, un nombre a la forme :

$$x = \pm Y \times b^e$$

$b$  est la base du système numérique utilisé

$Y$  est la mantisse : une suite de  $s$  entier

$y_1 y_2 \dots y_s$  avec  $y_1 \neq 0$  si  $x \neq 0$  et  $0 \leq y_i \leq (b - 1)$

$e$  est l'exposant (un nombre entier relatif)

La norme choisie est celle où la mantisse est comprise entre 0 et 1 et où le premier chiffre après la virgule est différent de zéro.

En notation flottante, un nombre a la forme :

$$x = \pm Y \times b^e$$

$b$  est la base du système numérique utilisé

$Y$  est la mantisse : une suite de  $s$  entier

$y_1 y_2 \dots y_s$  avec  $y_1 \neq 0$  si  $x \neq 0$  et  $0 \leq y_i \leq (b - 1)$

$e$  est l'exposant (un nombre entier relatif)

La norme choisie est celle où la mantisse est comprise entre 0 et 1 et où le premier chiffre après la virgule est différent de zéro.

En notation flottante, un nombre a la forme :

$$x = \pm Y \times b^e$$

$b$  est la base du système numérique utilisé

$Y$  est la mantisse : une suite de  $s$  entier

$y_1 y_2 \dots y_s$  avec  $y_1 \neq 0$  si  $x \neq 0$  et  $0 \leq y_i \leq (b - 1)$

$e$  est l'exposant (un nombre entier relatif)

La norme choisie est celle où la mantisse est comprise entre 0 et 1 et où le premier chiffre après la virgule est différent de zéro.

En notation flottante, un nombre a la forme :

$$x = \pm Y \times b^e$$

$b$  est la base du système numérique utilisé

$Y$  est la mantisse : une suite de  $s$  entier

$y_1 y_2 \dots y_s$  avec  $y_1 \neq 0$  si  $x \neq 0$  et  $0 \leq y_i \leq (b - 1)$

$e$  est l'exposant (un nombre entier relatif)

La norme choisie est celle où la mantisse est comprise entre 0 et 1 et où le premier chiffre après la virgule est différent de zéro.

## Calcul de l'erreur

Nous terminons ce chapitre en définissant les notions de troncature et d'arrondie.

## Exemple :

*En base 10,  $x = 1/15 = 0.066666666\dots$*

*Dans le cas d'une représentation tronquée nous aurons, pour  $s = 5$ ,  $fl(x) = 0.66666 * 10^{-1}$ .*

## Exemple :

*En base 10,  $x = 1/15 = 0.066666666\dots$*

*Dans le cas d'une représentation tronquée nous aurons, pour  $s = 5$ ,  $fl(x) = 0.66666 * 10^{-1}$ .*

## Exemple :

*En base 10,  $x = 1/15 = 0.066666666\dots$*

*Dans le cas d'une représentation tronquée nous aurons, pour  $s = 5$ ,  $fl(x) = 0.66666 * 10^{-1}$ .*

Remarquez comment nous avons modifié l'exposant afin de respecter la règle qui veut que le premier chiffre de la mantisse ne soit pas nul . Dans ce cas, l'erreur absolue  $X - fl(X)$  est de  $6 \times 10^{-7}$ . L'erreur relative est de l'ordre de  $10^{-5}$

Dans une représentation tronquée à  $s$  chiffres, l'erreur relative maximale est de l'ordre de  $10^{-s}$

Dans une représentation arrondie, lorsque la première décimale négligée est supérieure à 5, on ajoute 1 à la dernière décimale conservée.

## Exemple :

$$x = 1/15 = 0.0666666666.$$

*Nous écrivons  $fl(x) = 0.66667 \times 10^{-1}$*

*L'erreur absolue serait alors  $3.333 \times 10^{-7}$  et*

*l'erreur relative serait  $5 \times 10^{-6}$*

En général, l'erreur relative dans une représentation arrondie à  $s$  chiffres est de  $5 \times 10^{-(s+1)}$  soit la moitié de celle d'une représentation tronquée.

# Les règles de base du modèle

Pour effectuer une opération sur deux nombres réels, on effectue l'opération sur leurs représentations flottantes et on prend ensuite la représentation flottante du résultat.

# l'addition flottante

$$x \oplus y = fl(fl(x) + fl(y))$$

# la soustraction flottante

$$x \ominus y = fl(fl(x) - fl(y))$$

# la multiplication flottante

$$x \otimes y = fl(fl(x) \times fl(y))$$

# la division flottante

$$x \div y = fl(fl(x)/fl(y))$$

Chaque opération intermédiaire dans un calcul introduit une nouvelle erreur d'arrondi ou de troncature.

Dans la pratique, il faudra se souvenir du fait que deux expressions algébriquement équivalentes peuvent fournir des résultats différents et que l'ordre des opérations peut changer les résultats.

Chaque opération intermédiaire dans un calcul introduit une nouvelle erreur d'arrondi ou de troncature.

Dans la pratique, il faudra se souvenir du fait que deux expressions algébriquement équivalentes peuvent fournir des résultats différents et que l'ordre des opérations peut changer les résultats.

Pour l'addition et la soustraction on ne peut effectuer ces 2 opérations que si les exposants sont les mêmes. On transforme le plus petit exposant et donc on ne respecte plus la règle voulant que le premier chiffre de la mantisse ne soit pas nul.

## Quelques remarques sur ce modèle:

On constate une déviation importante par rapport aux lois habituelles de l'arithmétique.  
 $x + (y + z)$  peut être différent de  $(x + y) + z$ .

**Exemple :**

*Pour 4 chiffres significatifs ( $s = 4$ ) on a :*

$$(1 + 0.0005) + 0.0005 = 1.000$$

*car*

$$\begin{aligned} 0.1 \times 10^1 + 0.5. \times 10^{-3} &= \\ 0.1. \times 10^1 + 0.00005. \times 10^1 &= \\ 0.1 \times 10^1 + 0.0000. \times 10^1 &= 0.1 \times 10^1 \end{aligned}$$

**Exemple :**

*Pour 4 chiffres significatifs ( $s = 4$ ) on a :*

$$(1 + 0.0005) + 0.0005 = 1.000$$

*car*

$$\begin{aligned} 0.1 \times 10^1 + 0.5. \times 10^{-3} &= \\ 0.1. \times 10^1 + 0.00005. \times 10^1 &= \\ 0.1 \times 10^1 + 0.0000. \times 10^1 &= 0.1 \times 10^1 \end{aligned}$$

*et*

$$1 + (0.0005 + 0.0005) = 1.001$$

*Ainsi, l'addition flottante n'est pas associative  
(TD:Somme d'une série à termes positifs)*

*et*

$$1 + (0.0005 + 0.0005) = 1.001$$

*Ainsi, l'addition flottante n'est pas associative  
(TD:Somme d'une série à termes positifs)*

*On constate aussi que si  $y$  est très petit par rapport à  $x$ , l'addition de  $x$  et  $y$  donnera seulement  $x$ .*

## Exemple :

*L'équation  $1 + x = 1$  a  $x = 0$  comme unique solution. Mais dans un système à 10 chiffres significatifs, elle aura une infinité de solutions (il suffit de prendre  $|x| < 5 \times 10^{-11}$ )*

La distributivité de la multiplication par rapport à l'addition.

## Exemple :

*Considérons l'opération*

$$\begin{aligned}122 \times (333 + 695) &= \\(122 \times 333) + (122 \times 695) &= \\125416 &\end{aligned}$$

*Si nous effectuons ces deux calculs en arithmétique à 3 chiffres ( $s = 3$ ) et arrondi, nous obtenons:*

$$\begin{aligned}
 122 \times (333 + 695) &= \\
 fl(122) \times fl(1028) &= \\
 122 \times 103 \times 10^1 &= \\
 fl(125660) &= 126 \times 10^3
 \end{aligned}$$

$$\begin{aligned}
 (122 \times 333) + (122 \times 695) &= \\
 fl(40626) + fl(84790) &= \\
 406 \times 10^2 + 848 \times 10^2 &= \\
 fl(406 + 848) \times 10^2 &= \\
 fl(1254 \times 10^2) &= \\
 125 \times 10^3 &
 \end{aligned}$$

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