Représentation des nombres en machine

Said EL HAJJI

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

Laboratoire de Mathématiques, Informatique et Applications
http://www.fsr.ac.ma/mia/

Outline

Outline

Outline

Outline

Introduction

- 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
  - ...

Introduction

- 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
  - ...

Introduction

- 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
  - ...

Introduction

- 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
  - ...

Outline

Evaluation de l'erreur

Si X est une quantité à calculer et X^* la valeur calculée:

- $X - X^*$: l'erreur.
- $\mid E \mid = \mid X - X^* \mid$: l'erreur absolue.
- $E_r = \left| \frac{X - X^*}{X_r} \right|$: l'erreur relative.

Exemple :

$X = 2.224$, $X^* = 2.223$:

$$\mid E \mid = \mid X - X^* \mid = 2.224 - 2.223 = 0.001$$

On prend $X_r = X$:

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

Evaluation de l'erreur

Si X est une quantité à calculer et X^* la valeur calculée:

- $X - X^*$: l'erreur.
- $\mid E \mid = \mid X - X^* \mid$: l'erreur absolue.
- $E_r = \left| \frac{X - X^*}{X_r} \right|$: l'erreur relative.

Exemple :

$X = 2.224$, $X^* = 2.223$:

$$\mid E \mid = \mid X - X^* \mid = 2.224 - 2.223 = 0.001$$

On prend $X_r = X$:

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

Evaluation de l'erreur

Exemple :

Calculer la valeur de $(11111111)^2$

une petite calculatrice : $1,2345x10^{14}$

Réponse exacte est 123456787654321

L'erreur relative : 0.0005

Objectif:

- Nous voulons un bon ordre de grandeur (ici 10^{14}) et avoir le maximum de décimales exactes,
- Ce maximum ne peut excéder la longueur des mots permis par la machine et dépend donc de la machine

Evaluation de l'erreur

Exemple :

Calculer la valeur de $(11111111)^2$

une petite calculatrice : $1,2345x10^{14}$

Réponse exacte est 123456787654321

L'erreur relative : 0.0005

Objectif:

- Nous voulons un bon ordre de grandeur (ici 10^{14}) et avoir le maximum de décimales exactes,
- Ce maximum ne peut excéder la longueur des mots permis par la machine et dépend donc de la machine

Outline

Les nombres entiers :

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

$$-32768 \text{ et } 32767$$

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

$$-2147483648 \text{ et } 2147483647$$

Les nombres entiers :

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

$$-32768 \text{ et } 32767$$

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

$$-2147483648 \text{ et } 2147483647$$

Représentation :

Les nombres réels sont représentés en notation flottante:

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

1. b est la base.
2. Y est la mantisse
3. e est l'exposant.

Les nombres entiers :

Exemple :

- En base 10 : $x = 1/15 = 0.066666666.....$
- Représentation tronquée $s = 5$:

$$fl(x) = 0.66666 * 10^{-1}.$$

- L'erreur absolue :$X - fl(X) = 6 \times 10^{-7}$.
- 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}

Les nombres entiers :

Exemple :

- En base 10 : $x = 1/15 = 0.066666666.....$
- Représentation tronquée $s = 5$:

$$fl(x) = 0.66666 * 10^{-1}.$$

- L'erreur absolue :$X - fl(X) = 6 \times 10^{-7.}$.
- 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 :

$x = 1/15 = 0.066666666.$

- $fl(x) = 0.66667 \times 10^{-1}$
- L'erreur absolue : 3.333×10^{-7}
- L'erreur relative 5×10^{-6}

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

Les régles de base du modèle

l'addition flottante

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

Les régles de base du modèle

la soustraction flottante

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

Les régles de base du modèle

la multiplication flottante

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

Les régles de base du modèle

la division flottante

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

Les régles de base du modèle

- 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 .

Les régles de base du modèle

Remarque

$x + (y + z)$ peut être différent de $(x + y) + z$.

Exemple :

$s = 4$ on a :

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

car

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

et

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

La distributivité de la multiplication par rapport à l'addition

Exemple :

$$122 \times (333 + 695) = (122 \times 333) + (122 \times 695) = 125416$$

$s = 3$:

$$\begin{gathered} 122 \times (333 + 695) = fl(122) \times fl(1028) \\ = 122 \times 103 \times 10^1 = fl(125660) = 126 \times 10^3 \\ (122 \times 333) + (122 \times 695) = fl(40626) + fl(84790) \\ 406 \times 10^2 + 848 \times 10^2 = fl(406 + 848) \times 10^2 = \\ fl(1254 \times 10^2) = 125 \times 10^3 \end{gathered}$$

Propagation des erreurs.

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + ...$$

pour x positif : On obtient des tres bons resultats

Pour x négatif:

x	e^x	S
-10	$4.54.10^{-5}$	$4.54.10^{-5}$
-15	$3.06.10^{-7}$	$3.05.10^{-7}$
-20	$2.06.10^{-9}$	$-1.55.10^{-7}$
-25	$1.39.10^{-11}$	$1.87.10^{-5}$
-30	$9.36.10^{-14}$	$6.25.10^{-4}$

On voit que pour $x \leq (-20)$ les résultats obtenus sont dépourvus de sens.
L'explication de ce phénomène est la suivante: pour $x = -30$ les termes de la serie vont en croissant jusqu'à $\frac{x^{30}}{30!} = 8.10^{11}$ puis ils décroissent et $\frac{x^{107}}{107!} \tilde{} - 9.19.10^{-15}$.

Outline

Conditionnement et stabilité numérique.

Le fait que certains nombres ne soient pas représentés de façon exacte dans un ordinateur entraine que l'introduction même de donnée d'un problème en machine modifie quelque peu le problème initial; Il se peut que cette petite variation des données entraine une variation importante des résultats.
C'est la notion de conditionnement d'un problème.

Conditionnement et stabilité numérique.

On dit qu'un problème est bien (ou mal) conditionné, si une petite variation des données entraine une petite (une grande) variation sur les résultats.
Cette notion de conditionnement est liée au problème mathématique lui même et est indépendante de la méthode utilisée pour le résoudre.

Conditionnement et stabilité numérique.

Une autre notion importante en pratique est celle de stabilité numérique cád une propagation des erreurs numériques.
Ces notions de conditionnement d'un problème et de stabilité numérique d'une méthode de résolution sont fondamentales en analyse numérique.

Exemple simple:

$\begin{cases} x + \frac{1}{2}y = \frac{3}{2} \\ \frac{1}{2}x + \frac{1}{3}y = \frac{5}{6} \end{cases}$, la solution est : $[x = 1, y = 1]$

$\begin{cases} x + \frac{1}{2}y = \frac{3}{2} \\ \frac{1}{2}x + \frac{1}{3}y = 1 \end{cases}$, la solution est: $[x = 0, y = 3]$.

Instabilité numérique :

Si les erreurs introduites dans les étapes intermédiaires ont un effet négligeable sur le résultat final,on dira que le calcul ou l'algorithme est numériquement stable. Sinon , on dira que l'algorithme est numériquement instable.

Instabilité numérique :

Exemple :

$$I_n = \int_0^1 \frac{x^n}{a+x} dx$$

On a:

$$I_n = \int_0^1 \frac{x^{n-1}(x+a-a)}{a+x} dx =$$

$$\int_0^1 x^{n-1} dx - a \int_0^1 \frac{x^{n-1}}{a+x} dx = \frac{1}{n} - a I_{n-1} =$$

$$\sum_{i=0}^{n-1} \frac{(-a)^i}{n-i} + (-a)^n I_0$$

$$I_0 = ln(\frac{1+a}{a})$$

On peut calculer I_n pour toutes les valeurs de n.
Mais l'algorithme est numériquement instable car toute erreur dans le calcul de $I_0 = ln(\frac{1+a}{a})$ va se propager.

En effet si on note par I_0^* la valeur approchée de I_0 et si $I_0^* = I_0 + \epsilon$ alors

$$\begin{aligned} I_n^* &= \sum_{i=0}^{n-1} \frac{(-a)^i}{n-i} + (-a)^n I_0^* \\ &= \sum_{i=0}^{n-1} \frac{(-a)^i}{n-i} + (-a)^n (I_0 + \epsilon) \end{aligned}$$

donc $|I_n - I_n^*| \geq a^n \epsilon$.

Instabilité numérique :

Remarque

Sources d'erreur:

- Les erreurs liées à l'imprécision des mesures physiques ou au résultat d'un calcul approché
- Les erreurs de méthodes liées à l'algorithme utilisé
- Les erreurs de calcul liées à la machine

En général, pour l'objet de notre cours, si le premier chapitre met l'accent sur les erreurs liées à la machine, nous nous interresserons beaucoup plus aux erreurs liées aux méthodes ou encore aux algorithmes utilisés.