Arithmétique des ordinateurs

Polytech'Paris-UPMC

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque

Représentation des entiers en machine

Les Réels

Erreur d'arrondi

Représentation des nombres

On représente les nombres grâce à des symboles :

Représentation des nombres

On représente les nombres grâce à des symboles :

- *Représentation unaire* : un symbole de valeur unique

I = 1 II = 2

III = 3 IIIIIIIIII = 10

Représentation des nombres

On représente les nombres grâce à des symboles :

- *Représentation unaire* : un symbole de valeur unique

  I = 1 II = 2

  III = 3 IIIIIIIIII = 10

  Le calcul est facile :

  $$\text{I} + \text{III} = \text{IIII} \qquad \text{II} \times \text{III} = \text{II II II} = \text{IIIIII}$$

  mais cela devient vite incompréhensible

Représentation des nombres

On représente les nombres grâce à des symboles :

- *Représentation unaire* : un symbole de valeur unique

  I = 1 II = 2

  III = 3 IIIIIIIIII = 10

  Le calcul est facile :

  I + III = IIII II $\times$ III = II II II = IIIIII

  mais cela devient vite incompréhensible

- *Les chiffres Romains* : plusieurs symboles ayant des valeurs différentes

  I = 1 V = 5

  X = 10 L = 50

  Mais le nombre de symboles est théoriquement infini donc le calcul est impossible.

Représentation des nombres

On représente les nombres grâce à des symboles :

- *Représentation unaire* : un symbole de valeur unique

  I = 1   II = 2

  III = 3   IIIIIIIIII = 10

  Le calcul est facile :

  I + III = IIII   II $\times$ III = II II II = IIIIII

  mais cela devient vite incompréhensible
- *Les chiffres Romains* : plusieurs symboles ayant des valeurs différentes

  I = 1   V = 5

  X = 10   L = 50

  Mais le nombre de symboles est théoriquement infini donc le calcul est impossible.
- *Le système positionnel* : un petit nombre de symboles les *chiffres* dont la valeur dépend de la place

  999 = 900 + 90 + 9

  On date souvent le début de l'informatique de l'invention de cette représentation et du zéro. (« Al-jabr wa'l-muqâbalah » Muhammad ibn Müsä al-Khuwärizmï - env. 825)

Utilisation d'une base

Le plus souvent la valeur d'un chiffre dépend du nombre de symbole : *la base*.
En base b, les nombres sont représentés à l'aide de b symboles distincts.
Un nombre x est représenté par une suite de symboles :

$$x = a_n a_{n-1} ... a_1 a_0.$$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Le plus souvent la valeur d'un chiffre dépend du nombre de symbole : *la base*.
En base b, les nombres sont représentés à l'aide de b symboles distincts.
Un nombre x est représenté par une suite de symboles :

$$x = a_n a_{n-1} ... a_1 a_0.$$

- En décimal,
  $b = 10,\ \ a_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$
- En binaire,
  $b = 2,\ \ a_i \in \{0, 1\}$
- En hexadécimal,
  $b = 16,\ \ a_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$

Utilisation d'une base

Le plus souvent la valeur d'un chiffre dépend du nombre de symbole : *la base*.
En base b, les nombres sont représentés à l'aide de b symboles distincts.
Un nombre x est représenté par une suite de symboles :

$$x = a_n a_{n-1} \ldots a_1 a_0.$$

- En décimal,
  $b = 10, \ \ a_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$
- En binaire,
  $b = 2, \ \ a_i \in \{0, 1\}$
- En hexadécimal,
  $b = 16, \ \ a_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$

Les bases les plus utilisées sont : 10, 2, 2^k, 12, 60, 3 et $\frac{\sqrt{5}-1}{2}$ ...

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Représentation sur ordinateur

Les informations traitées sur ordinateur sont en général représentées et manipulées sous forme binaire.

L'unité d'information est le chiffre binaire ou bit (binary digit).

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque

Représentation des entiers en machine

Les Réels

Erreur d'arrondi

Représentation sur ordinateur

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Les informations traitées sur ordinateur sont en général représentées et manipulées sous forme binaire.

L'unité d'information est le chiffre binaire ou bit (binary digit).

- Les opérations arithmétiques de base sont faciles à exprimer en base 2.
  Par ex. les tables de multiplication se résument à :
  $0 \times 0 = 0$, $1 \times 0 = 0$, $0 \times 1 = 0$ et $1 \times 1 = 1$.

Représentation sur ordinateur

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Les informations traitées sur ordinateur sont en général représentées et manipulées sous forme binaire.

L'unité d'information est le chiffre binaire ou bit (binary digit).

- Les opérations arithmétiques de base sont faciles à exprimer en base 2.
  Par ex. les tables de multiplication se résument à :
  $0 \times 0 = 0$, $1 \times 0 = 0$, $0 \times 1 = 0$ et $1 \times 1 = 1$.
- La représentation binaire est facile à réaliser (systèmes à deux états obtenus à l'aide de transistors).

Représentation des nombres entiers

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

En base b,

$$x = a_n a_{n-1} ... a_1 a_0 = \sum_{i=0}^{n} a_i b^i.$$

a_0 est le chiffre de poids faible, a_n est le chiffre de poids fort.

Par ex. en base 10,
$1998 = 1 \times 10^3 + 9 \times 10^2 + 9 \times 10^1 + 8 \times 10^0$.

Par ex. en base 2,
$(101)_2 = 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 4 + 0 + 1 = 5$.

Nombres fractionnaires

La formule est la même mais il existe des exposants négatifs. En base b,

$$\begin{aligned} x &= a_n a_{n-1} \dots a_1 a_0, a_{-1} a_{-2} \dots a_{-k} \\ &= \sum_{\mathbf{i=-k}}^{n} a_i b^i. \end{aligned}$$

En base 10,

$$12,346 = 1 \times 10^1 + 2 \times 10^0 + 3 \times 10^{-1} + 4 \times 10^{-2} + 6 \times 10^{-3}.$$

En base 2,

$$\begin{aligned} 101,01 &= 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 0 \times \frac{1}{2^1} + 1 \times \frac{1}{2^2} \\ &= 5,25 \end{aligned}$$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Traduction depuis un base quelconque

Passage d'une base quelconque à la base 10

On utilise la formule :

$$
\begin{aligned}
(AB,C)_{16} &= 10 \times 16^1 + 11 \times 16^0 + 12 \times 16^{-1} \\
&= 160 + 11 + \frac{12}{16} = (171,75)_{10}
\end{aligned}
$$

- Cela revient donc à une simple somme.
- Cela fonctionne pour les nombres entiers ou les nombres fractionnaires

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque

Représentation des entiers en machine

Les Réels

Erreur d'arrondi

Pour les nombres entiers

on procède par divisions euclidiennes successives :

traduction vers une base quelconque

Pour les nombres entiers

on procède par divisions euclidiennes successives :

- On divise le nombre par la base,

traduction vers une base quelconque

Pour les nombres entiers

on procède par divisions euclidiennes successives :

- On divise le nombre par la base,
- puis le quotient obtenu par la base,

traduction vers une base quelconque

Pour les nombres entiers

on procède par divisions euclidiennes successives :

- On divise le nombre par la base,
- puis le quotient obtenu par la base,
- . . . ainsi de suite jusqu'à obtenir un quotient nul.

La suite des restes obtenus correspond aux chiffres dans la base visée.

traduction vers une base quelconque

Pour les nombres entiers

on procède par divisions euclidiennes successives :

- On divise le nombre par la base,
- puis le quotient obtenu par la base,
- ... ainsi de suite jusqu'à obtenir un quotient nul.

La suite des restes obtenus correspond aux chiffres dans la base visée.

Données : A, b : A est le nombre à convertir dans la base b

début

  $n \leftarrow 0$

  répéter

    [q, r] = div(A, b) // c'est-à-dire $A = q \times b + r$ et $0 \leq r < A$.

    $A \leftarrow q$

    $a_n \leftarrow r$

    $n \leftarrow n + 1$

  jusqu'à $q = 0$;

fin

Résultat : $(a_i)_{0 \leq i < n}$ les chiffres et n le nombre de chiffres

Ex : convertir $(44)_{10}$ vers la base 2

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Exemple

Ex : convertir $(44)_{10}$ vers la base 2

$$44 = 22 \times 2 + 0 \;\Rightarrow\; a_0 = 0$$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Exemple

Ex : convertir $(44)_{10}$ vers la base 2

$44 = 22 \times 2 + 0 \;\Rightarrow\; a_0 = 0$

$22 = 11 \times 2 + 0 \;\Rightarrow\; a_1 = 0$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Exemple

Ex : convertir $(44)_{10}$ vers la base 2

$$44 = 22 \times 2 + 0 \;\Rightarrow\; a_0 = 0$$

$$22 = 11 \times 2 + 0 \;\Rightarrow\; a_1 = 0$$

$$11 = 5 \times 2 + 1 \;\Rightarrow\; a_2 = 1$$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Ex : convertir $(44)_{10}$ vers la base 2

$44 = 22 \times 2 + 0 \;\Rightarrow\; a_0 = 0$

$22 = 11 \times 2 + 0 \;\Rightarrow\; a_1 = 0$

$11 = 5 \times 2 + 1 \;\Rightarrow\; a_2 = 1$

$5 = 2 \times 2 + 1 \;\Rightarrow\; a_3 = 1$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Exemple

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Ex : convertir $(44)_{10}$ vers la base 2

$44 = 22 \times 2 + 0 \;\Rightarrow\; a_0 = 0$

$22 = 11 \times 2 + 0 \;\Rightarrow\; a_1 = 0$

$11 = 5 \times 2 + 1 \;\Rightarrow\; a_2 = 1$

$5 = 2 \times 2 + 1 \;\Rightarrow\; a_3 = 1$

$2 = 1 \times 2 + 0 \;\Rightarrow\; a_4 = 0$

Exemple

Ex : convertir $(44)_{10}$ vers la base 2

$$44 = 22 \times 2 + 0 \;\Rightarrow\; a_0 = 0$$
$$22 = 11 \times 2 + 0 \;\Rightarrow\; a_1 = 0$$
$$11 = 5 \times 2 + 1 \;\Rightarrow\; a_2 = 1$$
$$5 = 2 \times 2 + 1 \;\Rightarrow\; a_3 = 1$$
$$2 = 1 \times 2 + 0 \;\Rightarrow\; a_4 = 0$$
$$1 = 0 \times 2 + 1 \;\Rightarrow\; a_5 = 1$$

donc $(44)_{10} = (101100)_2$.

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

traduction vers une base quelconque

Pour les nombres fractionnaires :

traduction vers une base quelconque

Pour les nombres fractionnaires :

- On décompose le nombre en partie entière et fractionnaire si $x > 0$,
  $x = E[x] + Fract[x]$

Pour les nombres fractionnaires :

- On décompose le nombre en partie entière et fractionnaire si $x > 0$, $x = E[x] + Fract[x]$
- On convertit la partie entière par la méthode précédente. Alors $E[x] = \sum_{i=0}^{p} a_i.b^i = a_p a_{p-1}...a_0$ en base b.

traduction vers une base quelconque

Pour les nombres fractionnaires :

- On décompose le nombre en partie entière et fractionnaire si $x > 0$, $x = E[x] + Fract[x]$
- On convertit la partie entière par la méthode précédente. Alors $E[x] = \sum_{i=0}^{p} a_i.b^i = a_p a_{p-1} ... a_0$ en base b.
- On convertit la partie fractionnaire :

Pour les nombres fractionnaires :

- On décompose le nombre en partie entière et fractionnaire si $x > 0$, $x = E[x] + Fract[x]$
- On convertit la partie entière par la méthode précédente. Alors $E[x] = \sum_{i=0}^{p} a_i.b^i = a_p a_{p-1} ... a_0$ en base b.
- On convertit la partie fractionnaire :
  - on multiplie $Fract[x]$ par b. Soit s_1 la partie entière de ce produit,

traduction vers une base quelconque

Pour les nombres fractionnaires :

- On décompose le nombre en partie entière et fractionnaire si $x > 0$, $x = E[x] + Fract[x]$
- On convertit la partie entière par la méthode précédente. Alors $E[x] = \sum_{i=0}^{p} a_i.b^i = a_p a_{p-1}...a_0$ en base b.
- On convertit la partie fractionnaire :
  - on multiplie $Fract[x]$ par b. Soit s_1 la partie entière de ce produit,
  - on recommence avec la partie fractionnaire du produit pour obtenir s_2,

traduction vers une base quelconque

Pour les nombres fractionnaires :

- On décompose le nombre en partie entière et fractionnaire si $x > 0$, $x = E[x] + Fract[x]$
- On convertit la partie entière par la méthode précédente. Alors $E[x] = \sum_{i=0}^{p} a_i.b^i = a_p a_{p-1} ... a_0$ en base b.
- On convertit la partie fractionnaire :
  - on multiplie $Fract[x]$ par b. Soit s_1 la partie entière de ce produit,
  - on recommence avec la partie fractionnaire du produit pour obtenir s_2,
  - . . . ainsi de suite,

traduction vers une base quelconque

Pour les nombres fractionnaires :

- On décompose le nombre en partie entière et fractionnaire si $x > 0$, $x = E[x] + Fract[x]$
- On convertit la partie entière par la méthode précédente. Alors $E[x] = \sum_{i=0}^{p} a_i.b^i = a_p a_{p-1} ... a_0$ en base b.
- On convertit la partie fractionnaire :
  - on multiplie $Fract[x]$ par b. Soit s_1 la partie entière de ce produit,
  - on recommence avec la partie fractionnaire du produit pour obtenir s_2,
  - . . . ainsi de suite,
  - on stoppe l'algorithme si la partie fractionnaire devient nulle.

Conversion vers une base quelconque

Pour les nombres fractionnaires :

- On décompose le nombre en partie entière et fractionnaire si $x > 0$, $x = E[x] + Fract[x]$
- On convertit la partie entière par la méthode précédente. Alors
$$E[x] = \sum_{i=0}^{p} a_i.b^i = a_p a_{p-1} ... a_0 \text{ en base b.}$$
- On convertit la partie fractionnaire :
  - on multiplie $Fract[x]$ par b. Soit s_1 la partie entière de ce produit,
  - on recommence avec la partie fractionnaire du produit pour obtenir s_2,
  - . . . ainsi de suite,
  - on stoppe l'algorithme si la partie fractionnaire devient nulle.

  Alors :
$$Fract[x] = \sum_{i=1}^{+\infty} s_i.b^{-i} = 0, s_1 s_2 ... s_i ... \quad \text{en base b.}$$

Pour les nombres fractionnaires :

- On décompose le nombre en partie entière et fractionnaire si $x > 0$, $x = E[x] + Fract[x]$
- On convertit la partie entière par la méthode précédente. Alors $$E[x] = \sum_{i=0}^{p} a_i.b^i = a_p a_{p-1} ... a_0$$ en base b.
- On convertit la partie fractionnaire :
  - on multiplie $Fract[x]$ par b. Soit s_1 la partie entière de ce produit,
  - on recommence avec la partie fractionnaire du produit pour obtenir s_2,
  - . . . ainsi de suite,
  - on stoppe l'algorithme si la partie fractionnaire devient nulle.

  Alors : $$Fract[x] = \sum_{i=1}^{+\infty} s_i.b^{-i} = 0, s_1 s_2 ... s_i ...$$ en base b.
- Finalement $x = a_p a_{p-1} ... a_0, s_1 s_2 ... s_i ...$ en base b.

- Ex : écrire 63,734375 en base 16

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

exemple

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

exemple

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \;\Rightarrow\; s_1 = B$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

exemple

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \Rightarrow s_1 = B$
  - $0,75 * 16 = 12 \Rightarrow s_2 = C$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \ \Rightarrow \ s_1 = B$
  - $0,75 * 16 = 12 \ \Rightarrow \ s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque

Représentation des entiers en machine

Les Réels

Erreur d'arrondi

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \Rightarrow s_1 = B$
  - $0,75 * 16 = 12 \Rightarrow s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.
- Exemple 2 : écrire 0,3 en base 2

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

exemple

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \Rightarrow s_1 = B$
  - $0,75 * 16 = 12 \Rightarrow s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.
- Exemple 2 : écrire 0,3 en base 2
  - $0,3 * 2 = 0,6 \Rightarrow s_1 = 0$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque

Représentation des entiers en machine

Les Réels

Erreur d'arrondi

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \;\Rightarrow\; s_1 = B$
  - $0,75 * 16 = 12 \;\Rightarrow\; s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.
- Exemple 2 : écrire 0,3 en base 2
  - $0,3 * 2 = 0,6 \;\Rightarrow\; s_1 = 0$
  - $0,6 * 2 = 1,2 \;\Rightarrow\; s_2 = 1$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

exemple

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \ \Rightarrow \ s_1 = B$
  - $0,75 * 16 = 12 \ \Rightarrow \ s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.
- Exemple 2 : écrire 0,3 en base 2
  - $0,3 * 2 = 0,6 \ \Rightarrow \ s_1 = 0$
  - $0,6 * 2 = 1,2 \ \Rightarrow \ s_2 = 1$
  - $0,2 * 2 = 0,4 \ \Rightarrow \ s_3 = 0$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \;\Rightarrow\; s_1 = B$
  - $0,75 * 16 = 12 \;\Rightarrow\; s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.
- Exemple 2 : écrire 0,3 en base 2
  - $0,3 * 2 = 0,6 \;\Rightarrow\; s_1 = 0$
  - $0,6 * 2 = 1,2 \;\Rightarrow\; s_2 = 1$
  - $0,2 * 2 = 0,4 \;\Rightarrow\; s_3 = 0$
  - $0,4 * 2 = 0,8 \;\Rightarrow\; s_4 = 0$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

exemple

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \;\Rightarrow\; s_1 = B$
  - $0,75 * 16 = 12 \;\Rightarrow\; s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.
- Exemple 2 : écrire 0,3 en base 2
  - $0,3 * 2 = 0,6 \;\Rightarrow\; s_1 = 0$
  - $0,6 * 2 = 1,2 \;\Rightarrow\; s_2 = 1$
  - $0,2 * 2 = 0,4 \;\Rightarrow\; s_3 = 0$
  - $0,4 * 2 = 0,8 \;\Rightarrow\; s_4 = 0$
  - $0,8 * 2 = 1,6 \;\Rightarrow\; s_5 = 1$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

exemple

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \;\Rightarrow\; s_1 = B$
  - $0,75 * 16 = 12 \;\Rightarrow\; s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.
- Exemple 2 : écrire 0,3 en base 2
  - $0,3 * 2 = 0,6 \;\Rightarrow\; s_1 = 0$
  - $0,6 * 2 = 1,2 \;\Rightarrow\; s_2 = 1$
  - $0,2 * 2 = 0,4 \;\Rightarrow\; s_3 = 0$
  - $0,4 * 2 = 0,8 \;\Rightarrow\; s_4 = 0$
  - $0,8 * 2 = 1,6 \;\Rightarrow\; s_5 = 1$
  - $0,6 * 2 = 1,2 \;\Rightarrow\; s_6 = 1$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \;\Rightarrow\; s_1 = B$
  - $0,75 * 16 = 12 \;\Rightarrow\; s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.
- Exemple 2 : écrire 0,3 en base 2
  - $0,3 * 2 = 0,6 \;\Rightarrow\; s_1 = 0$
  - $0,6 * 2 = 1,2 \;\Rightarrow\; s_2 = 1$
  - $0,2 * 2 = 0,4 \;\Rightarrow\; s_3 = 0$
  - $0,4 * 2 = 0,8 \;\Rightarrow\; s_4 = 0$
  - $0,8 * 2 = 1,6 \;\Rightarrow\; s_5 = 1$
  - $0,6 * 2 = 1,2 \;\Rightarrow\; s_6 = 1$
  - $0,2 * 2 = 0,4 \;\Rightarrow\; s_7 = 0$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \ \Rightarrow \ s_1 = B$
  - $0,75 * 16 = 12 \ \Rightarrow \ s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.
- Exemple 2 : écrire 0,3 en base 2
  - $0,3 * 2 = 0,6 \ \Rightarrow \ s_1 = 0$
  - $0,6 * 2 = 1,2 \ \Rightarrow \ s_2 = 1$
  - $0,2 * 2 = 0,4 \ \Rightarrow \ s_3 = 0$
  - $0,4 * 2 = 0,8 \ \Rightarrow \ s_4 = 0$
  - $0,8 * 2 = 1,6 \ \Rightarrow \ s_5 = 1$
  - $0,6 * 2 = 1,2 \ \Rightarrow \ s_6 = 1$
  - $0,2 * 2 = 0,4 \ \Rightarrow \ s_7 = 0$
  - $\ldots$

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

exemple

- Ex : écrire 63,734375 en base 16
  - $63 = 3 * 16 + 15 = (3F)_{16}$
  - $0,734375 * 16 = 11,75 \;\Rightarrow\; s_1 = B$
  - $0,75 * 16 = 12 \;\Rightarrow\; s_2 = C$
  - Alors $(63,734375)_{10} = (3F,BC)_{16}$.
- Exemple 2 : écrire 0,3 en base 2
  - $0,3 * 2 = 0,6 \;\Rightarrow\; s_1 = 0$
  - $0,6 * 2 = 1,2 \;\Rightarrow\; s_2 = 1$
  - $0,2 * 2 = 0,4 \;\Rightarrow\; s_3 = 0$
  - $0,4 * 2 = 0,8 \;\Rightarrow\; s_4 = 0$
  - $0,8 * 2 = 1,6 \;\Rightarrow\; s_5 = 1$
  - $0,6 * 2 = 1,2 \;\Rightarrow\; s_6 = 1$
  - $0,2 * 2 = 0,4 \;\Rightarrow\; s_7 = 0$
  - $\ldots$
  - Alors $(0,3)_{10} = (0,01001100110011\ldots)_2$.

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Remarque

- La conversion d'un nombre entier s'arrête toujours
- La conversion d'un nombre fractionnaire ne s'arrête pas toujours.
- Certains nombres fractionnaires ont une représentation finie dans une base et infinie dans une autre.

  $\Rightarrow$ il faudra arrondir

Représentation des nombres
Représentation des nombres
Utilisation d'une base
Représentation sur ordinateur
Représentation des nombres entiers
Nombres fractionnaires
Traduction depuis un base quelconque
traduction vers une base quelconque
Exemple
traduction vers une base quelconque
exemple
Remarque
Représentation des entiers en machine
Les Réels
Erreur d'arrondi

Représentation des nombres
Représentation des entiers en machine
Codage des nombres entiers
Complément à 2
Complément à 2
Autres représentations
Dépassement de capacité
Les Réels
Erreur d'arrondi

Représentation des entiers en machine

Codage des nombres entiers

Représentation des nombres
Représentation des entiers en machine
Codage des nombres entiers
Complément à 2
Complément à 2
Autres représentations
Dépassement de capacité
Les Réels
Erreur d'arrondi

1. Entiers naturels :

Les entiers naturels sont codés en général sur 16 ou 32 ou 64 bits.
Un codage sur n bits permet de représenter tous les entiers compris entre 0 et $2^n - 1$.
Par ex., sur 2 octets on peut coder les entiers de 0 à $65535 = 2^{16} - 1$.

Codage des nombres entiers

Représentation des nombres
Représentation des entiers en machine
Codage des nombres entiers
Complément à 2
Complément à 2
Autres représentations
Dépassement de capacité
Les Réels
Erreur d'arrondi

1. Entiers naturels :

Les entiers naturels sont codés en général sur 16 ou 32 ou 64 bits.
Un codage sur n bits permet de représenter tous les entiers compris entre 0 et $2^n - 1$.
Par ex., sur 2 octets on peut coder les entiers de 0 à $65535 = 2^{16} - 1$.

2. Entiers relatifs : Il y a *deux façons* de faire :

- Garder un bit pour coder le signe :
  Un codage sur n bits permet de représenter tous les entiers compris entre $-2^{\mathbf{n-1}} - 1$ et $2^{\mathbf{n-1}} - 1$.

Codage des nombres entiers

Représentation des nombres
Représentation des entiers en machine
Codage des nombres entiers
Complément à 2
Complément à 2
Autres représentations
Dépassement de capacité
Les Réels
Erreur d'arrondi

1. Entiers naturels :
Les entiers naturels sont codés en général sur 16 ou 32 ou 64 bits.
Un codage sur n bits permet de représenter tous les entiers compris entre 0 et $2^n - 1$.
Par ex., sur 2 octets on peut coder les entiers de 0 à $65535 = 2^{16} - 1$.

2. Entiers relatifs : Il y a *deux façons* de faire :

- Garder un bit pour coder le signe :
  Un codage sur n bits permet de représenter tous les entiers compris entre $-2^{\mathbf{n-1}} - 1$ et $2^{\mathbf{n-1}} - 1$.
  + inverser un nombre ou tester son signe est facile,
  + l'ensemble des nombres représentables est symétrique
  - la somme de deux nombres non signés est différente de celle de deux nombres signés
  - il y a +0 et -0

Complément à 2

- Le complément à 2 :
  Le poids du chiffre de poids fort *est négatif*

$$
\begin{aligned}
x &= a_{n-1}...a_1a_0 \\
&= \quad a_{n-1} \times 2^{n-1} + a_{n-2} \times 2^{n-2} + \cdots + a_1 \times 2^1 + a_0 \times 2^0.
\end{aligned}
$$

- Le complément à 2 :
  Le poids du chiffre de poids fort *est négatif*

$$
\begin{aligned}
x &= a_{n-1}...a_1 a_0 \\
&= -a_{n-1} \times 2^{n-1} + a_{n-2} \times 2^{n-2} + \cdots + a_1 \times 2^1 + a_0 \times 2^0.
\end{aligned}
$$

Complément à 2

- Le complément à 2 :
  Le poids du chiffre de poids fort *est négatif*

$$
\begin{aligned}
x &= a_{n-1}...a_1a_0 \\
&= -a_{n-1} \times 2^{n-1} + a_{n-2} \times 2^{n-2} + \cdots + a_1 \times 2^1 + a_0 \times 2^0.
\end{aligned}
$$

  - Les entiers positifs sont identiques aux précédents, leur bit de poids fort est à 0.
  - Pour coder un entier négatif, on code sa valeur absolue, on *complémente* (on inverse) ses bits puis on ajoute 1.

Complément à 2

- Le complément à 2 :
  Le poids du chiffre de poids fort *est négatif*

$$\begin{aligned} x &= a_{n-1} ... a_1 a_0 \\ &= -a_{n-1} \times 2^{n-1} + a_{n-2} \times 2^{n-2} + \cdots + a_1 \times 2^1 + a_0 \times 2^0. \end{aligned}$$

  - Les entiers positifs sont identiques aux précédents, leur bit de poids fort est à 0.
  - Pour coder un entier négatif, on code sa valeur absolue, on *complémente* (on inverse) ses bits puis on ajoute 1.

  Un codage sur n bits permet de représenter tous les entiers compris entre $\mathbf{-2^{n-1}}$ et $2^{\mathbf{n-1}} - 1$.

  Il y a un nombre négatif supplémentaire

Complément à 2

- Le complément à 2 :
  Le poids du chiffre de poids fort *est négatif*

$$
\begin{aligned}
x &= a_{n-1}...a_1a_0 \\
&= -a_{n-1} \times 2^{n-1} + a_{n-2} \times 2^{n-2} + \cdots + a_1 \times 2^1 + a_0 \times 2^0.
\end{aligned}
$$

  - Les entiers positifs sont identiques aux précédents, leur bit de poids fort est à 0.
  - Pour coder un entier négatif, on code sa valeur absolue, on *complémente* (on inverse) ses bits puis on ajoute 1.

  Un codage sur n bits permet de représenter tous les entiers compris entre $-\mathbf{2^{n-1}}$ et $2^{\mathbf{n-1}} - 1$.

Il y a un nombre négatif supplémentaire

Exemple : coder -2 sur 8 bits.

$2_{10} = (00000010)_2$.

Le complément est : 11111101

On ajoute 1 et on obtient le résultat : 11111110.

Complément à 2

Sur 16 bits :

0111111111111111 = 32767 c'est le plus grand nombre représentable

32767+1 = 1000000000000000 = -32768 c'est le plus petit nombre négatif

1111111111111111 = -1 c'est le plus grand nombre négatif

Cela revient à faire du calcul modulo 2^n

Complément à 2

Sur 16 bits :

0111111111111111 = 32767 c'est le plus grand nombre représentable

32767+1 = 1000000000000000 = -32768 c'est le plus petit nombre négatif

1111111111111111 = -1 c'est le plus grand nombre négatif

Cela revient à faire du calcul modulo 2^n

Remarques :

+ Le bit de poids fort d'un nombre négatif est toujours 1 $\Rightarrow$ le test du signe est facile.

Sur 16 bits :

0111111111111111 = 32767 c'est le plus grand nombre représentable

32767+1 = 1000000000000000 = -32768 c'est le plus petit nombre négatif

1111111111111111 = -1 c'est le plus grand nombre négatif

Cela revient à faire du calcul modulo 2^n

Remarques :

+ Le bit de poids fort d'un nombre négatif est toujours 1 $\Rightarrow$ le test du signe est facile.

- L'inversion est plus complexe.

Complément à 2

Sur 16 bits :

0111111111111111 = 32767 c'est le plus grand nombre représentable

32767+1 = 1000000000000000 = -32768 c'est le plus petit nombre négatif

1111111111111111 = -1 c'est le plus grand nombre négatif

Cela revient à faire du calcul modulo 2^n

Remarques :

+ Le bit de poids fort d'un nombre négatif est toujours 1 $\Rightarrow$ le test du signe est facile.

- L'inversion est plus complexe.

- L'ensemble des nombres représentables n'est pas symétrique.

Complément à 2

Sur 16 bits :

0111111111111111 = 32767 c'est le plus grand nombre représentable

32767+1 = 1000000000000000 = -32768 c'est le plus petit nombre négatif

1111111111111111 = -1 c'est le plus grand nombre négatif

Cela revient à faire du calcul modulo 2^n

Remarques :

+ Le bit de poids fort d'un nombre négatif est toujours 1 $\Rightarrow$ le test du signe est facile.

- L'inversion est plus complexe.

- L'ensemble des nombres représentables n'est pas symétrique.

+ Le codage n'est pas redondant.

Complément à 2

Sur 16 bits :

0111111111111111 = 32767 c'est le plus grand nombre représentable

32767+1 = 1000000000000000 = -32768 c'est le plus petit nombre négatif

1111111111111111 = -1 c'est le plus grand nombre négatif

Cela revient à faire du calcul modulo 2^n

Remarques :

+ Le bit de poids fort d'un nombre négatif est toujours 1 $\Rightarrow$ le test du signe est facile.

- L'inversion est plus complexe.

- L'ensemble des nombres représentables n'est pas symétrique.

+ Le codage n'est pas redondant.

+ L'addition et la soustraction sont identiques.

Autres représentations

- représentations redondantes
  - Une représentation est dite redondante s'il y a plusieurs moyens de représenter un même nombre.
  - Souvent, on représente les entiers en base 2 avec les chiffres $0, 1$ et 2 ou avec les chiffres $0, 1$ et -1.
  - Inconvénients :
    - perte de place
    - la comparaison est difficile
  - Avantage :
    - Addition sans retenue
- représentations modulaires
  - Avantage :
    - Addition et multiplication sans retenue
  - Inconvénients :
    - la comparaison est impossible
    - le retour vers une autre représentation est très difficile

Représentation des nombres
Représentation des entiers en machine
Codage des nombres entiers
Complément à 2
Complément à 2
Autres représentations
Dépassement de capacité
Les Réels
Erreur d'arrondi

Dépassement de capacité

Un ordinateur ne calcule pas bien !

Représentation des nombres

Représentation des entiers en machine
Codage des nombres entiers
Complément à 2
Complément à 2
Autres représentations
Dépassement de capacité

Les Réels

Erreur d'arrondi

Dépassement de capacité

Un ordinateur ne calcule pas bien !

- Pour un ordinateur le nombre de chiffres est fixé.

Représentation des nombres

Représentation des entiers en machine
Codage des nombres entiers
Complément à 2
Complément à 2
Autres représentations
Dépassement de capacité

Les Réels

Erreur d'arrondi

Dépassement de capacité

Un ordinateur ne calcule pas bien !

- Pour un ordinateur le nombre de chiffres est fixé.
- Pour un mathématicien le nombre de chiffres dépend de la valeur représentée.

Représentation des nombres
Représentation des entiers en machine
Codage des nombres entiers
Complément à 2
Complément à 2
Autres représentations
Dépassement de capacité
Les Réels
Erreur d'arrondi

Dépassement de capacité

Un ordinateur ne calcule pas bien !

- Pour un ordinateur le nombre de chiffres est fixé.
- Pour un mathématicien le nombre de chiffres dépend de la valeur représentée.
- L'ordinateur calcule toujours modulo 2^n.

Représentation des nombres
Représentation des entiers en machine
Codage des nombres entiers
Complément à 2
Complément à 2
Autres représentations
Dépassement de capacité
Les Réels
Erreur d'arrondi

Dépassement de capacité

Un ordinateur ne calcule pas bien !

- Pour un ordinateur le nombre de chiffres est fixé.
- Pour un mathématicien le nombre de chiffres dépend de la valeur représentée.
- L'ordinateur calcule toujours modulo 2^n.

Lorsque le résultat d'un calcul doit être représenté sur plus de chiffres que ceux disponibles, il y a *Dépassement de capacité* (Ariane 5).

Représentation des nombres
Représentation des entiers en machine
Codage des nombres entiers
Complément à 2
Complément à 2
Autres représentations
Dépassement de capacité
Les Réels
Erreur d'arrondi

Les Réels

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Nos formules mathématiques utilisent les réels.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Nos formules mathématiques utilisent les réels.

Mais les réels ne peuvent pas être représentés

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Nos formules mathématiques utilisent les réels.

Mais les réels ne peuvent pas être représentés

L'utilisation des nombres fractionnaires n'est pas adaptée

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Nos formules mathématiques utilisent les réels.

Mais les réels ne peuvent pas être représentés

L'utilisation des nombres fractionnaires n'est pas adaptée

- car on utilise des nombres d'ordre de grandeur très différent (population $\sim 10^9$, dimension de parois cellulaire $\sim 10^{-9}$)

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Nos formules mathématiques utilisent les réels.

Mais les réels ne peuvent pas être représentés

L'utilisation des nombres fractionnaires n'est pas adaptée

- car on utilise des nombres d'ordre de grandeur très différent (population $\sim 10^9$, dimension de parois cellulaire $\sim 10^{-9}$)
- car on souhaite obtenir une *précision relative*

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Les Réels

Nos formules mathématiques utilisent les réels.

Mais les réels ne peuvent pas être représentés

L'utilisation des nombres fractionnaires n'est pas adaptée

- car on utilise des nombres d'ordre de grandeur très différent (population $\sim 10^9$, dimension de parois cellulaire $\sim 10^{-9}$)
- car on souhaite obtenir une *précision relative*

$\Longrightarrow$*il faut utiliser la notation scientifique*

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Notation scientifique d'un réel en base b

$\forall x \in \mathbb{R},$ on écrit

$$x \quad = \quad \pm a_0, a_1 a_2 \ldots a_i \ldots E \pm e_k \ldots e_0$$

avec

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Notation scientifique d’un réel en base b

$\forall x \in \mathbb{R},$ on écrit

$$x = \pm a_0, a_1 a_2 \ldots a_i \ldots E \pm e_k \ldots e_0$$

avec

- *Le signe :* $\varepsilon \in \{-1, 1\}$

Représentation des nombres

Représentation des entiers en machine

Les Réels
Les Réels
Notation scientifique d’un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d’arrondi
Les modes d’arrondi IEEE 754
Les modes d’arrondi (suite)

Erreur d’arrondi

Notation scientifique d'un réel en base b

$\forall x \in \mathbb{R},$ on écrit

$$x = \pm a_0, a_1 a_2 \dots a_i \dots E \pm e_k \dots e_0$$

avec

- *Le signe :* $\varepsilon \in \{-1, 1\}$
- *La mantisse :* $m = a_0, a_1 a_2 ... a_i ... \in [0, b[$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Notation scientifique d'un réel en base b

$\forall x \in \mathbb{R},$ on écrit

$$x = \pm a_0, a_1 a_2 \ldots a_i \ldots E \pm e_k \ldots e_0$$

avec

- *Le signe :* $\varepsilon \in \{-1, 1\}$
- *La mantisse :* $m = a_0, a_1 a_2 ... a_i ... \in [0, b[$
- *L'exposant :* $e = \pm e_k \ldots e_0 \in \mathbb{Z}$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Notation scientifique d'un réel en base b

$\forall x \in \mathbb{R},$ on écrit

$$x \quad = \quad \pm a_0, a_1 a_2 \dots a_i \dots E \pm e_k \dots e_0$$

avec

- *Le signe :* $\varepsilon \in \{-1, 1\}$
- *La mantisse :* $m = a_0, a_1 a_2 ... a_i ... \in [0, b[$
- *L'exposant :* $e = \pm e_k \dots e_0 \in \mathbb{Z}$

Alors :

$$x = \varepsilon . b^e . m$$

Représentation des nombres

Représentation des entiers en machine

Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)

Erreur d'arrondi

Représentation normalisée

Cette notation est redondante par exemple en base 10 :

$$1 = 0.1E1 = 0.01E2 = \ldots$$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Représentation normalisée

Cette notation est redondante par exemple en base 10 :

$$1 = 0.1E1 = 0.01E2 = \dots$$

Pour éviter cela, on normalise la notation c'est à dire :

- on impose $m \in [1, b[$,
- ou encore « le chiffre de poids fort de m n'est pas nul ».

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

La norme IEEE-754

Coder un réel en arithmétique virgule flottante, c'est coder le triplet $\{\varepsilon, e, m\}$.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

La norme IEEE-754

Coder un réel en arithmétique virgule flottante, c'est coder le triplet $\{\varepsilon, e, m\}$.

La norme IEEE-754 (1985) prévoit entre autre les codages :

- *simple précision* sur 32 bits,
- *double précision* sur 64 bits.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

La norme IEEE-754

Coder un réel en arithmétique virgule flottante, c'est coder le triplet $\{\varepsilon, e, m\}$.

La norme IEEE-754 (1985) prévoit entre autre les codages :

- *simple précision* sur 32 bits,
- *double précision* sur 64 bits.

Elle utilise la base 2 pour coder e et m :

$$e = \sum_{i=0}^{p} b_i \cdot 2^i \quad \text{et}$$

$$m = \sum_{i=0}^{\infty} a_i \cdot 2^{-i-1} \quad \text{avec} \quad (a_i, b_i) \in \{0, 1\}$$

Le codage de ε tient sur un bit et vaut 0 si $x \geq 0$ et 1 si $x < 0$.

Représentation des nombres

Représentation des entiers en machine

Les Réels

Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)

Erreur d'arrondi

La norme IEEE-754

Coder un réel en arithmétique virgule flottante, c'est coder le triplet $\{\varepsilon, e, m\}$.

La norme IEEE-754 (1985) prévoit entre autre les codages :

- *simple précision* sur 32 bits,
- *double précision* sur 64 bits.

Elle utilise la base 2 pour coder e et m :

$$
\begin{aligned}
e &= \sum_{i=0}^{p} b_i \cdot 2^i \quad \text{et} \\
m &= 1 + \sum_{i=1}^{\infty} a_i \cdot 2^{-i-1} \quad \text{avec} \quad (a_i, b_i) \in \{0, 1\}
\end{aligned}
$$

Le codage de ε tient sur un bit et vaut 0 si $x \geq 0$ et 1 si $x < 0$.

- On ne code pas a_0 qui vaut toujours 1 (bit caché).

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

La norme IEEE-754

Coder un réel en arithmétique virgule flottante, c'est coder le triplet $\{\varepsilon, e, m\}$.

La norme IEEE-754 (1985) prévoit entre autre les codages :

- *simple précision* sur 32 bits,
- *double précision* sur 64 bits.

Elle utilise la base 2 pour coder e et m :

$$\begin{aligned} e + 2^{p-1} - 1 &= \sum_{i=0}^{p} b_i \cdot 2^i \text{ et} \\ m &= 1 + \sum_{i=1}^{\infty} a_i \cdot 2^{-i-1} \text{ avec } (a_i, b_i) \in \{0, 1\} \end{aligned}$$

Le codage de ε tient sur un bit et vaut 0 si $x \geq 0$ et 1 si $x < 0$.

- On ne code pas a_0 qui vaut toujours 1 (bit caché).
- L'exposant n'est pas signé, il est biaisé c'est-à-dire que l'on représente $e + 2^{p-1} - 1$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

La norme IEEE-754 (suite)

1	2 ... 9	10 32
ε	e + $2^7 - 1$	a_1 .... a_{23}

Codage IEEE 754 Simple précision

1	2 ... 12	13 64
ε	e + $2^{10} - 1$	a_1 .... a_{52}

Codage IEEE 754 Double précision

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Exposants

Les exposants 00...00 et 11...11 sont interdits :

- l'exposant 00...00 signifie que le nombre est dénormalisé
- l'exposant 11...11 indique que l'on n'a pas affaire à un nombre. $\longrightarrow$ Notation `NaN` (Not a Number), résultat par exemple de la division de 0 par 0

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Exposant maximum, exposant minimum

En *simple précision*, l'exposant est codé sur 8 bits :

Au maximum $(11111110)_2$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Exposant maximum, exposant minimum

En *simple précision*, l'exposant est codé sur 8 bits :

$$\begin{aligned} \text{Au maximum } (11111110)_2 &= 2^7 + 2^6 + ... + 2 &&= e_{max} + 2^7 - 1 \\ e_{max} &= 2^6 + ... + 2 + 1 &&= 2^7 - 1 = 127 \end{aligned}$$

Représentation des nombres

Représentation des entiers en machine

Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)

Erreur d'arrondi

Exposant maximum, exposant minimum

En *simple précision*, l'exposant est codé sur 8 bits :

$$\begin{aligned} \text{Au maximum } (11111110)_2 &= 2^7 + 2^6 + ... + 2 &&= e_{max} + 2^7 - 1 \\ e_{max} &= 2^6 + ... + 2 + 1 &&= 2^7 - 1 = 127 \\ \text{avec} \quad 2^{127} &\simeq 1,7 \cdot 10^{38}. \end{aligned}$$

Représentation des nombres

Représentation des entiers en machine

Les Réels

Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)

Erreur d'arrondi

Exposant maximum, exposant minimum

En *simple précision*, l'exposant est codé sur 8 bits :

$$\begin{aligned} \text{Au maximum } (11111110)_2 &= 2^7 + 2^6 + \ldots + 2 &&= e_{max} + 2^7 - 1 \\ e_{max} &= 2^6 + \ldots + 2 + 1 &&= 2^7 - 1 = 127 \\ \text{avec} \quad 2^{127} &\simeq 1,7 \cdot 10^{38}. \end{aligned}$$

$$\text{Au minimum } (00000001)_2$$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Exposant maximum, exposant minimum

En *simple précision*, l'exposant est codé sur 8 bits :

$$
\begin{aligned}
\text{Au maximum } (11111110)_2 &= 2^7 + 2^6 + ... + 2 &&= e_{max} + 2^7 - 1 \\
e_{max} &= 2^6 + ... + 2 + 1 &&= 2^7 - 1 = 127 \\
\text{avec} \quad 2^{127} &\simeq 1,7 \cdot 10^{38}. && \\
\text{Au minimum } (00000001)_2 &= 1 &&= e_{min} + 2^7 - 1, \\
e_{min} &= 2 - 2^7 &&= -126 \\
2^{-126} &\simeq 1.2 \cdot 10^{-38}. &&
\end{aligned}
$$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Exposant maximum, exposant minimum

En *simple précision*, l'exposant est codé sur 8 bits :

$$
\begin{array}{rlll}
\text{Au maximum } (11111110)_2 & = 2^7 + 2^6 + \ldots + 2 & = e_{max} + 2^7 - 1 \\
e_{max} & = 2^6 + \ldots + 2 + 1 & = 2^7 - 1 = 127 \\
\text{avec} \quad 2^{127} & \simeq 1,7 \cdot 10^{38}. & \\
\text{Au minimum } (00000001)_2 & = 1 & = e_{min} + 2^7 - 1, \\
e_{min} & = 2 - 2^7 & = -126 \\
2^{-126} & \simeq 1.2 \cdot 10^{-38}. &
\end{array}
$$

En *double précision*, l'exposant est codé sur 11 bits :

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Exposant maximum, exposant minimum

En *simple précision*, l'exposant est codé sur 8 bits :

$$\begin{aligned}
\text{Au maximum } (11111110)_2 &= 2^7 + 2^6 + \ldots + 2 &&= e_{max} + 2^7 - 1 \\
e_{max} &= 2^6 + \ldots + 2 + 1 &&= 2^7 - 1 = 127 \\
\text{avec} \quad 2^{127} &\simeq 1,7 \cdot 10^{38}. \\
\text{Au minimum } (00000001)_2 &= 1 &&= e_{min} + 2^7 - 1, \\
e_{min} &= 2 - 2^7 &&= -126 \\
2^{-126} &\simeq 1.2 \cdot 10^{-38}.
\end{aligned}$$

En *double précision*, l'exposant est codé sur 11 bits :

$$\begin{aligned}
\text{Au maximum } (11111111110)_2 &= 2^{10} + \ldots + 2 &&= e_{max} + 2^{10} - 1 \\
e_{max} &= 2^9 + \ldots + 2 + 1 &&= 2^{10} - 1 = 1023 \\
2^{1023} &\simeq 9.0 \cdot 10^{307} \\
\text{Au minimum } (00000000001)_2 &= 1 &&= e_{min} + 2^{10} - 1 \\
e_{min} &= 2 - 2^{10} &&= -1022 \\
2^{-1022} &\simeq 2.2 \cdot 10^{-308}
\end{aligned}$$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Notion de mode d'arrondi

Soient X_{min} (resp. X_{max}) le plus petit (resp. le plus grand) nombre flottant.
Soit $\mathbb{F}$ l'ensemble des nombre flottant représentable.

Notion de mode d'arrondi

Soient X_{min} (resp. X_{max}) le plus petit (resp. le plus grand) nombre flottant.
Soit $\mathbb{F}$ l'ensemble des nombre flottant représentable.
Pour tout réel $x \in]X_{min}, X_{max}[$,
il existe $\{X^-, X^+\}$ dans $\mathbb{F}^2$ tels que

$$X^- < x < X^+ \quad \text{et} \quad]X^-, X^+[\cap \mathbb{F} = \emptyset$$

Notion de mode d'arrondi

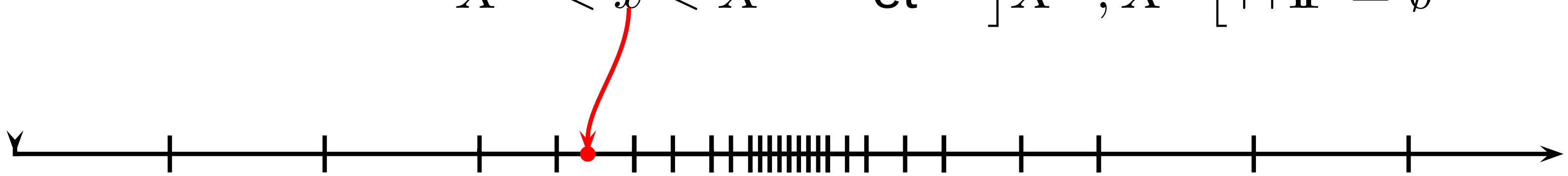
Soient X_{min} (resp. X_{max}) le plus petit (resp. le plus grand) nombre flottant.
Soit $\mathbb{F}$ l'ensemble des nombre flottant représentable.
Pour tout réel $x \in]X_{min}, X_{max}[$,
il existe $\{X^-, X^+\}$ dans $\mathbb{F}^2$ tels que

$$X^- < x < X^+ \quad \text{et} \quad]X^-, X^+[\cap \mathbb{F} = \emptyset$$

Notion de mode d’arrondi

Soient X_{min} (resp. X_{max}) le plus petit (resp. le plus grand) nombre flottant.
Soit $\mathbb{F}$ l’ensemble des nombre flottant représentable.
Pour tout réel $x \in]X_{min}, X_{max}[$,
il existe $\{X^-, X^+\}$ dans $\mathbb{F}^2$ tels que

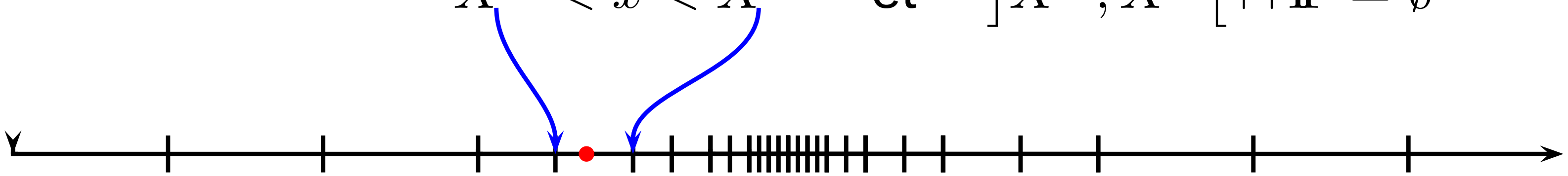
$$X^- < x < X^+ \quad \text{et} \quad]X^-, X^+[\cap \mathbb{F} = \emptyset$$



Notion de mode d'arrondi

Soient X_{min} (resp. X_{max}) le plus petit (resp. le plus grand) nombre flottant.
Soit $\mathbb{F}$ l'ensemble des nombre flottant représentable.
Pour tout réel $x \in]X_{min}, X_{max}[$,
il existe $\{X^-, X^+\}$ dans $\mathbb{F}^2$ tels que

$$X^- < x < X^+ \quad \text{et} \quad]X^-, X^+[\cap \mathbb{F} = \emptyset$$



Notion de mode d'arrondi

Soient X_{min} (resp. X_{max}) le plus petit (resp. le plus grand) nombre flottant.
Soit $\mathbb{F}$ l'ensemble des nombre flottant représentable.
Pour tout réel $x \in]X_{min}, X_{max}[$,
il existe $\{X^-, X^+\}$ dans $\mathbb{F}^2$ tels que

$$X^- < x < X^+ \quad \text{et} \quad]X^-, X^+[\cap \mathbb{F} = \emptyset$$

Un arrondi est dit *exact* si on choisit X^+ et X^- pour représenter x.

Notion de mode d'arrondi

Soient X_{min} (resp. X_{max}) le plus petit (resp. le plus grand) nombre flottant.
Soit $\mathbb{F}$ l'ensemble des nombre flottant représentable.
Pour tout réel $x \in \,]X_{min}, X_{max}[$,

il existe $\{X^-, X^+\}$ dans $\mathbb{F}^2$ tels que

$$X^- < x < X^+ \quad \text{et} \quad]X^-, X^+[\cap \mathbb{F} = \emptyset$$

Un arrondi est dit *exact* si on choisit X^+ et X^- pour représenter x.
Un *mode d'arrondi* est une règle qui, en fonction de x, fournit X^- ou X^+.
Par exemple pour la représentation simple précision :

x =

ε	$e + 2^7 - 1$	a_1	$\ldots$	a_{23}	a_{24}	....

X =

ε	$E + 2^7 - 1$	a'_1	$..$	a'_{23}

Notion de mode d'arrondi

Soient X_{min} (resp. X_{max}) le plus petit (resp. le plus grand) nombre flottant.
Soit $\mathbb{F}$ l'ensemble des nombre flottant représentable.
Pour tout réel $x \in]X_{min}, X_{max}[$,
il existe $\{X^-, X^+\}$ dans $\mathbb{F}^2$ tels que

$$X^- < x < X^+ \quad \text{et} \quad]X^-, X^+[\cap \mathbb{F} = \emptyset$$

Un arrondi est dit *exact* si on choisit X^+ et X^- pour représenter x.
Un *mode d'arrondi* est une règle qui, en fonction de x, fournit X^- ou X^+.
Par exemple pour la représentation simple précision :

x =	ε	$e + 2^7 - 1$	$a_1 \quad \ldots \quad a_{23} \quad a_{24}$ ....
X =	ε	$E + 2^7 - 1$	$a'_1 \quad \ldots \quad a'_{23}$

Le mode d'arrondi détermine le passage des a_i aux a'_i.

Les modes d'arrondi IEEE 754

Dans la norme IEEE 754 il y a 4 modes d'arrondi

L'arrondi vers zéro :

x est représenté par le flottant le plus près de x compris entre x et 0. Cela revient à *tronquer* les chiffres :

$$a_i = a'_i \text{ et } e = E.$$

L'arrondi au plus près :

x est représenté par le flottant le plus près de x. Ceci s'obtient en faisant

$$a'_{23} = a_{23} + a_{24}$$

avec une propagation éventuelle de la retenue. On a donc

$$E = e + \delta \text{ avec } \delta = 0 \text{ ou } 1.$$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

Les modes d'arrondi (suite)

Représentation des nombres
Représentation des entiers en machine
Les Réels
Les Réels
Notation scientifique d'un réel en base b
Représentation normalisée
La norme IEEE-754
La norme IEEE-754 (suite)
Exposants
Exposant maximum, exposant minimum
Notion de mode d'arrondi
Les modes d'arrondi IEEE 754
Les modes d'arrondi (suite)
Erreur d'arrondi

L'arrondi vers plus l'infini :

x est toujours représenté par X^+.
La relation devient

$$a'_{23} = a_{23} + \overline{\varepsilon}$$

plus les mêmes comportements que précédemment ($\overline{\varepsilon}$ représente la négation logique de ε).

L'arrondi vers moins l'infini :

x est toujours représenté par X^-.
On a

$$a'_{23} = a_{23} + \varepsilon.$$

Cet arrondi s'effectue à chaque entrée de données et après chaque opération arithmétique élémentaire.

⚠ *Dans tous les modes, si x est exactement représentable ($x \in \mathbb{F}$) alors*

$$X = x$$

Erreur d'arrondi

Représentation des nombres

Représentation des entiers en machine

Les Réels

Erreur d'arrondi
L'ensemble des nombres flottants

Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

L'ensemble des nombres flottants

Soit $\mathbb{F}$ l'ensemble des réels codables sur ordinateur.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

L’ensemble des nombres flottants

Soit $\mathbb{F}$ l’ensemble des réels codables sur ordinateur.
Pour être codable exactement en machine un réel doit :

- être un rationnel,

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d’arrondi
L’ensemble des nombres flottants
Deux notions d’erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d’accumulation d’erreur
Phénomène de cancellation
Un exemple d’erreur d’arrondi
Conditionnement
Conclusion

L'ensemble des nombres flottants

Soit $\mathbb{F}$ l'ensemble des réels codables sur ordinateur.
Pour être codable exactement en machine un réel doit :

- être un rationnel,
- avoir un exposant compris entre -126 et $+127$ en simple précision et entre -1022 et $+1023$ pour la double précision,

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

L'ensemble des nombres flottants

Soit $\mathbb{F}$ l'ensemble des réels codables sur ordinateur.
Pour être codable exactement en machine un réel doit :

- être un rationnel,
- avoir un exposant compris entre -126 et $+127$ en simple précision et entre -1022 et $+1023$ pour la double précision,
- avoir une mantisse dont le développement en base 2 ne contienne que des zéros à partir du bit 24 en simple précision et à partir du bit 53 en double précision.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

L'ensemble des nombres flottants

Soit $\mathbb{F}$ l'ensemble des réels codables sur ordinateur.
Pour être codable exactement en machine un réel doit :

- être un rationnel,
- avoir un exposant compris entre -126 et $+127$ en simple précision et entre -1022 et $+1023$ pour la double précision,
- avoir une mantisse dont le développement en base 2 ne contienne que des zéros à partir du bit 24 en simple précision et à partir du bit 53 en double précision.

Cet ensemble dépend de la précision utilisée mais c'est toujours un ensemble *fini* donc *borné* et *discret*.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

L'ensemble des réels $\mathbb{R}$ est *infini*, *indénombrable*, *non borné* et *complet*. Tout réel qui n'est pas un nombre flottant doit être **approché** par un élément de $\mathbb{F}$.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

L'ensemble des réels $\mathbb{R}$ est *infini*, *indénombrable*, *non borné* et *complet*. Tout réel qui n'est pas un nombre flottant doit être **approché** par un élément de $\mathbb{F}$.

Il y a donc :

- des phénomènes de sous-capacité et de sur-capacité dus au fait que $\mathbb{F}$ est borné,

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

L'ensemble des réels $\mathbb{R}$ est *infini*, *indénombrable*, *non borné* et *complet*. Tout réel qui n'est pas un nombre flottant doit être **approché** par un élément de $\mathbb{F}$.

Il y a donc :

- des phénomènes de sous-capacité et de sur-capacité dus au fait que $\mathbb{F}$ est borné,
- des phénomènes d'erreurs d'arrondi dûs au fait que $\mathbb{F}$ est discret.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

L'ensemble des réels $\mathbb{R}$ est *infini*, *indénombrable*, *non borné* et *complet*. Tout réel qui n'est pas un nombre flottant doit être **approché** par un élément de $\mathbb{F}$.

Il y a donc :

- des phénomènes de sous-capacité et de sur-capacité dus au fait que $\mathbb{F}$ est borné,
- des phénomènes d'erreurs d'arrondi dûs au fait que $\mathbb{F}$ est discret.

! *les opérateurs arithmétiques ne sont pas des lois internes dans $\mathbb{F}$.*

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Tout résultat informatique est entaché d'erreur.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Deux notions d'erreurs

Tout résultat informatique est entaché d'erreur.

Supposons que $X \in \mathbb{F}$ est le résultat obtenu en voulant calculer la valeur $x \in \mathbb{R}$.

- on peut écrire

$$X \quad = \quad x + e_a \text{ avec } e_a \in \mathbb{R}$$

e_a est appelé *erreur absolue*. Par exemple, une montre est fiable à 2 minutes près.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Tout résultat informatique est entaché d'erreur.

Supposons que $X \in \mathbb{F}$ est le résultat obtenu en voulant calculer la valeur $x \in \mathbb{R}$.

- on peut écrire

$$X = x + e_a \text{ avec } e_a \in \mathbb{R}$$

  e_a est appelé *erreur absolue*. Par exemple, une montre est fiable à 2 minutes près.

- on peut écrire

$$X = x(1 + e_r) \text{ avec } e_r \in \mathbb{R}$$

  e_r est appelé *erreur relative*. Par exemple un appareil de mesure donne des résultats fiables à 5%.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Deux notions d'erreurs

Tout résultat informatique est entaché d'erreur.

Supposons que $X \in \mathbb{F}$ est le résultat obtenu en voulant calculer la valeur $x \in \mathbb{R}$.

- on peut écrire

$$X = x + e_a \text{ avec } e_a \in \mathbb{R}$$

e_a est appelé *erreur absolue*. Par exemple, une montre est fiable à 2 minutes près.

- on peut écrire

$$X = x(1 + e_r) \text{ avec } e_r \in \mathbb{R}$$

e_r est appelé *erreur relative*. Par exemple un appareil de mesure donne des résultats fiables à 5%.

▷ La notion la plus importante avec la représentation flottante est l'erreur relative. C'est la « qualité des chiffres » obtenus

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Représentation flottante et erreur relative

La notation flottant est liée à l'erreur relative : Soit $x \in \mathbb{R}$ que l'on écrit

$$x \quad = \quad \varepsilon.m.b^e \text{ avec } 1 \leq m < b$$

Comme $x \in \mathbb{R}$, m peut être de taille infinie.
Choisir une approximation flottante de x revient à choisir une mantisse de p chiffres telle que :

$$X \quad = \quad \varepsilon.M.b^e \text{ alors } X \quad = \quad x(1 + b^{-p}.\alpha)$$

$b^{-p}\alpha$ est l'erreur relative sur x. α est appelé *erreur d'arrondi normalisée* :

- en arrondi au plus près, $\alpha \in [-0.5, 0.5[$;
- en arrondi vers zéro, $\alpha \in [0, 1[$;
- en arrondi vers plus ou moins l'infini, $\alpha \in [-1, +1[$;

$$\begin{aligned}
\text{Par exemple } \pi &= 3.14159265358979323846264 \cdots \times 10^0 \\
3.141593 \times 10^0 &= \pi(1 + 10^{-6} * 0.110\ldots) \\
3.141592 \times 10^0 &= \pi(1 - 10^{-6} * 0.208\ldots)
\end{aligned}$$

Effet des opérations

En simplifiant, on peut dire que :

- Les additions et soustractions ajoutent les erreurs absolues

$$X + Y \quad = \quad x + e_a^x + y + e_a^y \quad = \quad (x + y) + (e_a^x + e_a^y)$$

Effet des opérations

En simplifiant, on peut dire que :

- Les additions et soustractions ajoutent les erreurs absolues

$$X + Y \quad = \quad x + e_a^x + y + e_a^y \quad = \quad (x+y) + (e_a^x + e_a^y)$$

- Les multiplications et divisions ajoutent les erreurs relatives

$$X \times Y \quad = \quad x(1+e_r^x) \times y(1+e_r^y) \quad = \quad (x \times y)(1 + e_r^x + e_a^y + e_r^x e_a^y)$$

Effet des opérations

En simplifiant, on peut dire que :

- Les additions et soustractions ajoutent les erreurs absolues

$$X + Y = x + e_a^x + y + e_a^y = (x + y) + (e_a^x + e_a^y)$$

- Les multiplications et divisions ajoutent les erreurs relatives

$$X \times Y = x(1 + e_r^x) \times y(1 + e_r^y) = (x \times y)(1 + e_r^x + e_a^y + e_r^x e_a^y)$$

Donc il est plus facile de gérer les erreurs absolues quand on fait uniquement des additions & soustractions

Effet des opérations

En simplifiant, on peut dire que :

- Les additions et soustractions ajoutent les erreurs absolues

$$X + Y = x + e_a^x + y + e_a^y = (x + y) + (e_a^x + e_a^y)$$

- Les multiplications et divisions ajoutent les erreurs relatives

$$X \times Y = x(1 + e_r^x) \times y(1 + e_r^y) = (x \times y)(1 + e_r^x + e_a^y + e_r^x e_a^y)$$

Donc il est plus facile de gérer les erreurs absolues quand on fait uniquement des additions & soustractions et plus facile de gérer les erreurs relatives quand on fait uniquement des multiplications & divisions.

Dépassement de capacité

Parfois le résultat d'un calcul ne peut pas être représenté.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Parfois le résultat d'un calcul ne peut pas être représenté.

- Un résultat trop grand $+\infty$ et $-\infty$.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Dépassement de capacité

Parfois le résultat d'un calcul ne peut pas être représenté.

- Un résultat trop grand $+\infty$ et $-\infty$.
- Un résultat trop petit $+0$ et -0.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Dépassement de capacité

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Parfois le résultat d'un calcul ne peut pas être représenté.

- Un résultat trop grand $+\infty$ et $-\infty$.
- Un résultat trop petit $+0$ et -0.
- Autre `NaN` ($\sqrt{-3}$, $\frac{0}{0}$).

! *Si on ne vérifie pas, ces valeurs sont sources de grandes erreurs (Ariane 5, USS Yorktown).*

Phénomène d'accumulation d'erreur

Lorsqu'on fait de nombreux calculs, les erreurs peuvent s'accumuler.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Phénomène d'accumulation d'erreur

Lorsqu'on fait de nombreux calculs, les erreurs peuvent s'accumuler.

Quand on calcule par la méthode des rectangles $\int_0^1 tdt = 0.5$. On obtient :

Nb d'itération	Résultat	Erreur
10	0.450000017881393	$< 10^{-1}$
1000	0.500496685504913	$< 10^{-3}$
10000	0.499953210353851	$< 10^{-4}$
100000	0.499371945858002	$< 10^{-3}$
1000000	0.493827700614929	$< 10^{-2}$
10000000	0.451817184686661	$< 10^{-1}$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Phénomène d'accumulation d'erreur

Lorsqu'on fait de nombreux calculs, les erreurs peuvent s'accumuler.

Quand on calcule par la méthode des rectangles $\int_0^1 tdt = 0.5$. On obtient :

Nb d'itération	Résultat	Erreur
10	0.450000017881393	$< 10^{-1}$
1000	0.500496685504913	$< 10^{-3}$
10000	0.499953210353851	$< 10^{-4}$
100000	0.499371945858002	$< 10^{-3}$
1000000	0.493827700614929	$< 10^{-2}$
10000000	0.451817184686661	$< 10^{-1}$

$\Rightarrow$ chaque calcul ajoute une petite erreur.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Phénomène d'accumulation d'erreur

Lorsqu'on fait de nombreux calculs, les erreurs peuvent s'accumuler.

Quand on calcule par la méthode des rectangles $\int_0^1 tdt = 0.5$. On obtient :

Nb d'itération	Résultat	Erreur
10	0.450000017881393	$< 10^{-1}$
1000	0.500496685504913	$< 10^{-3}$
10000	0.499953210353851	$< 10^{-4}$
100000	0.499371945858002	$< 10^{-3}$
1000000	0.493827700614929	$< 10^{-2}$
10000000	0.451817184686661	$< 10^{-1}$

$\Rightarrow$ chaque calcul ajoute une petite erreur.
$\triangleright$ Par exemple, le missile patriot : une erreur de 0.000000095, répétée 360 000 fois.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Phénomène de cancellation

Mais il peut y avoir de grandes erreurs même sur un petit nombre de calculs. Cela est souvent du au phénomène de cancellation.

$$X + Y \quad = \quad x(1 + e_r^x) + y(1 + e_r^y) \quad = \quad (x + y)(1 + \frac{x}{x + y} e_r^x + \frac{y}{x + y} e_r^y)$$

Si $x \simeq -y$, les termes d'erreur peuvent être arbitrairement grands.

Phénomène de cancellation

Mais il peut y avoir de grandes erreurs même sur un petit nombre de calculs.
Cela est souvent du au phénomène de cancellation.

$$X + Y = x(1 + e_r^x) + y(1 + e_r^y) = (x + y)(1 + \frac{x}{x + y} e_r^x + \frac{y}{x + y} e_r^y)$$

Si $x \simeq -y$, les termes d'erreur peuvent être arbitrairement grands.
Par exemple :

$1,23456$ $\quad$ $1,234564992$ $\quad$ $1,234560143$

Phénomène de cancellation

Mais il peut y avoir de grandes erreurs même sur un petit nombre de calculs.
Cela est souvent du au phénomène de cancellation.

$$X + Y = x(1 + e_r^x) + y(1 + e_r^y) = (x + y)(1 + \frac{x}{x+y} e_r^x + \frac{y}{x+y} e_r^y)$$

Si $x \simeq -y$, les termes d'erreur peuvent être arbitrairement grands.
Par exemple :

$$\begin{array}{r} 1,23456 \\ - \quad 1,23449 \\ \hline \end{array} \qquad \begin{array}{r} 1,234564992 \\ - \quad 1,234490000 \\ \hline \end{array} \qquad \begin{array}{r} 1,234560143 \\ - \quad 1,234495000 \\ \hline \end{array}$$

Phénomène de cancellation

Mais il peut y avoir de grandes erreurs même sur un petit nombre de calculs.
Cela est souvent du au phénomène de cancellation.

$$X + Y = x(1 + e_r^x) + y(1 + e_r^y) = (x + y)(1 + \frac{x}{x+y} e_r^x + \frac{y}{x+y} e_r^y)$$

Si $x \simeq -y$, les termes d'erreur peuvent être arbitrairement grands.
Par exemple :

$$\begin{array}{r} 1,23456 \\ -\quad 1,23449 \\ \hline 7,0000E{-}5 \end{array} \qquad \begin{array}{r} 1,234564992 \\ -\quad 1,234490000 \\ \hline 7,4991E{-}5 \end{array} \qquad \begin{array}{r} 1,234560143 \\ -\quad 1,234495000 \\ \hline 6,5143E{-}5 \end{array}$$

Phénomène de cancellation

Mais il peut y avoir de grandes erreurs même sur un petit nombre de calculs. Cela est souvent du au phénomène de cancellation.

$$X + Y \quad = \quad x(1+e_r^x) + y(1+e_r^y) \quad = \quad (x+y)(1 + \frac{x}{x+y}e_r^x + \frac{y}{x+y}e_r^y)$$

Si $x \simeq -y$, les termes d'erreur peuvent être arbitrairement grands.
Par exemple :

$$\begin{array}{rr} & 1,23456 \\ - & 1,23449 \\ \hline & 7,0000E{-}5 \end{array} \qquad \begin{array}{rr} & 1,234564992 \\ - & 1,234490000 \\ \hline & 7,4991E{-}5 \end{array} \qquad \begin{array}{rr} & 1,234560143 \\ - & 1,234495000 \\ \hline & 6,5143E{-}5 \end{array}$$

La comparaison :
if $a == b$ **then**
␣ . . .
est un exemple très courant de cancellation !

L'algorithme du Gentleman

début

$A \leftarrow 1.0$

$B \leftarrow 1.0$

tant que $((A+1)-A)-1=0$ **faire**

$\quad A \leftarrow 2 \times A$

tant que $((B+A)-A)-B \neq 0$ **faire**

$\quad B \leftarrow B + 1.0$

fin

Résultat : B

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Un exemple d’erreur d’arrondi

L’algorithme du Gentleman

début

$A \leftarrow 1.0$

$B \leftarrow 1.0$

tant que $((A+1)-A)-1=0$ **faire**

$\quad A \leftarrow 2 \times A$

tant que $((B+A)-A)-B \neq 0$ **faire**

$\quad B \leftarrow B + 1.0$

fin

Résultat : B la base de calcul

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d’arrondi
L’ensemble des nombres flottants
Deux notions d’erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d’accumulation d’erreur
Phénomène de cancellation
Un exemple d’erreur d’arrondi
Conditionnement
Conclusion

Conditionnement

Un programme informatique donnera un résultat qui dépend des données.

Représentation des nombres

Représentation des entiers en machine

Les Réels

Erreur d'arrondi
L'ensemble des nombres flottants

Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conditionnement

Un programme informatique donnera un résultat qui dépend des données.

- Une erreur sur les données ou un arrondi au cours du calcul va entraîner une modification du résultat.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conditionnement

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Un programme informatique donnera un résultat qui dépend des données.

- Une erreur sur les données ou un arrondi au cours du calcul va entraîner une modification du résultat.
- Sous certaines conditions,
  si r est le résultat d'un problème de donnée a,
  $r + \Delta r$ sera celui avec la donnée $a + \Delta a$.

Conditionnement

Un programme informatique donnera un résultat qui dépend des données.

- Une erreur sur les données ou un arrondi au cours du calcul va entraîner une modification du résultat.
- Sous certaines conditions,
  si r est le résultat d'un problème de donnée a,
  $r + \Delta r$ sera celui avec la donnée $a + \Delta a$.
- Certains problèmes sont plus sensibles que d'autres aux erreurs

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conditionnement

Un programme informatique donnera un résultat qui dépend des données.

- Une erreur sur les données ou un arrondi au cours du calcul va entraîner une modification du résultat.
- Sous certaines conditions,
  si r est le résultat d'un problème de donnée a,
  $r + \Delta r$ sera celui avec la donnée $a + \Delta a$.
- Certains problèmes sont plus sensibles que d'autres aux erreurs
  - Calcul du point commun à deux droites perpendiculaires

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Un programme informatique donnera un résultat qui dépend des données.

- Une erreur sur les données ou un arrondi au cours du calcul va entraîner une modification du résultat.
- Sous certaines conditions,
  si r est le résultat d'un problème de donnée a,
  $r + \Delta r$ sera celui avec la donnée $a + \Delta a$.
- Certains problèmes sont plus sensibles que d'autres aux erreurs
  - Calcul du point commun à deux droites perpendiculaires
  - Calcul du point commun à deux droites presque parallèles

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conditionnement

Un programme informatique donnera un résultat qui dépend des données.

- Une erreur sur les données ou un arrondi au cours du calcul va entraîner une modification du résultat.
- Sous certaines conditions,
  si r est le résultat d'un problème de donnée a,
  $r + \Delta r$ sera celui avec la donnée $a + \Delta a$.
- Certains problèmes sont plus sensibles que d'autres aux erreurs
  - Calcul du point commun à deux droites perpendiculaires
  - Calcul du point commun à deux droites presque parallèles
- On définit donc le conditionnement

$$C_\alpha = \sup_{\|\Delta a\| \leq \alpha} \frac{\|\Delta r\|}{\|\Delta a\|}$$

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conclusion

Les ordinateurs calculent mal car ils essayent de représenter un ensemble continu, non-dénombrable et complet par un ensemble fini et discret.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conclusion

Les ordinateurs calculent mal car ils essayent de représenter un ensemble continu, non-dénombrable et complet par un ensemble fini et discret.

- Dépassement de capacité (overflow), sous capacité (underflow).

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conclusion

Les ordinateurs calculent mal car ils essayent de représenter un ensemble continu, non-dénombrable et complet par un ensemble fini et discret.

- Dépassement de capacité (overflow), sous capacité (underflow).
- Opérations non internes à l'ensemble $\mathbb{F}$.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conclusion

Les ordinateurs calculent mal car ils essayent de représenter un ensemble continu, non-dénombrable et complet par un ensemble fini et discret.

- Dépassement de capacité (overflow), sous capacité (underflow).
- Opérations non internes à l'ensemble $\mathbb{F}$.
- Perte de propriétés des opérations (associativité, commutativité, distributivité).

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conclusion

Les ordinateurs calculent mal car ils essayent de représenter un ensemble continu, non-dénombrable et complet par un ensemble fini et discret.

- Dépassement de capacité (overflow), sous capacité (underflow).
- Opérations non internes à l'ensemble $\mathbb{F}$.
- Perte de propriétés des opérations (associativité, commutativité, distributivité).
- Perte de propriétés des fonctions :
  - bornes $\cos(x) > 1$,
  - croissance.

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conclusion

Les ordinateurs calculent mal car ils essayent de représenter un ensemble continu, non-dénombrable et complet par un ensemble fini et discret.

- Dépassement de capacité (overflow), sous capacité (underflow).
- Opérations non internes à l'ensemble $\mathbb{F}$.
- Perte de propriétés des opérations (associativité, commutativité, distributivité).
- Perte de propriétés des fonctions :
  - bornes $\cos(x) > 1$,
  - croissance.
- Accumulation d'erreurs, cancellation ...

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Les ordinateurs calculent mal car ils essayent de représenter un ensemble continu, non-dénombrable et complet par un ensemble fini et discret.

- Dépassement de capacité (overflow), sous capacité (underflow).
- Opérations non internes à l'ensemble $\mathbb{F}$.
- Perte de propriétés des opérations (associativité, commutativité, distributivité).
- Perte de propriétés des fonctions :
  - bornes $\cos(x) > 1$,
  - croissance.
- Accumulation d'erreurs, cancellation ...

« It makes me nervous to fly on airplanes since I know they are designed using floating-point arithmetic. »
A. Householder

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Conclusion

Représentation des nombres
Représentation des entiers en machine
Les Réels
Erreur d'arrondi
L'ensemble des nombres flottants
Deux notions d'erreurs
Représentation flottante et erreur relative
Effet des opérations
Dépassement de capacité
Phénomène d'accumulation d'erreur
Phénomène de cancellation
Un exemple d'erreur d'arrondi
Conditionnement
Conclusion

Les ordinateurs calculent mal car ils essayent de représenter un ensemble continu, non-dénombrable et complet par un ensemble fini et discret.

- Dépassement de capacité (overflow), sous capacité (underflow).
- Opérations non internes à l'ensemble $\mathbb{F}$.
- Perte de propriétés des opérations (associativité, commutativité, distributivité).
- Perte de propriétés des fonctions :
  - bornes $\cos(x) > 1$,
  - croissance.
- Accumulation d'erreurs, cancellation . . .

« It makes me nervous to fly on airplanes since I know they are designed using floating-point arithmetic. »
A. Householder

Solutions possibles

arithmétiques sécurisés (arithmétique stochastique, arithmétique d'intervalle), calcul exact, calcul symbolique