Codage des nombres non entiers

M. Combacau
combacau@laas.fr

12 novembre 2024

Objectif

Savoir coder un nombre non entier en base 2
Connaître la précision d'un code
Comprendre les arrondis et les erreurs de codage

Propriétés générales

- Appartient à un ensemble **dense** ($\mathbb{Q}$ ou $\mathbb{R}$)
- n bits : 2^n codes

  Tous les nombres non entiers ne sont pas associés à un code !
- Nombre est codé par la valeur la plus proche codable exactement
- Nombre $v \in [d, u]$ avec d et u 2 nombres successifs codables
  - $Pa = |d - u|$: précision absolue
  - $Ea = min(|v - d|, |v - u|)$: erreur de codage absolue
  - $Pr = \frac{min(|d-v|,|u-v|)}{|d-u|}$: précision relative pour la valeur v
  - $Er = \frac{min(|v-d|,|v-u|)}{|v|}$: erreur de codage relative
- il existe toujours un code pour la valeur 0

Deux types de codage

Comme pour la représentation que l'on manipule en arithmétique, deux types de codage existent :

1. le codage en virgule fixe
   $\rightarrow$ traduction de l'écriture habituelle $\pi = 3,1415926...$
2. le codage en vigule flottante
   $\rightarrow$ traduction de la notation scientifique $y = 1,23 \times 10^{-12}$

Principe du codage en virgule fixe (1)

Voici tout d'abord un exemple

Soit un mot codant B en virgule fixe avec le format suivant sur 16 bits

b_{15}	b_{14}	b_{13}	b_{12}	b_{11}	b_{10}	b_9	b_8	b_7	b_6	b_5	b_4	b_3	b_2	b_1	b_0

partie entière , partie fractionnaire

Un tel code est caractérisé par la paire (m,f)=(10,6)
Si le nombre codé est positif, sa valeur est :

$$b_{15}.2^9 + \ldots + b_6.2^0 + b_5.2^{-1} + \ldots + b_0.2^{-6}$$

Principe du codage en virgule fixe (2)

Le codage est caractérisé par la paire (m, f) avec m+f = n

b_{n-1}	$\ldots$	b_f	b_{f-1}	$\ldots$	b_0
partie entiere			partie fractionnaire		

- Voyons ce code comme celui d'un nombre positif

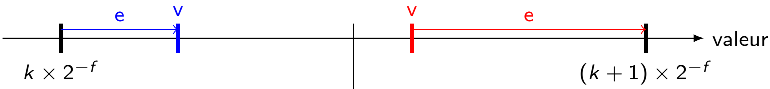
$$
\begin{aligned}
B &= b_{n-1}.2^{n-1-f} + \ldots + b_f.2^{f-f} + b_{f-1}.2^{f-1-f} + \ldots + b_0.2^{0-f} \\
&= (b_{n-1}.2^{n-1} + \ldots + b_f.2^{f} + b_{f-1}.2^{f-1} + \ldots + b_0) \times 2^{-f}
\end{aligned}
$$

- Correspond au code de la partie entière de $(2^f \times B)$
- Codage des nombres négatifs en ca2
- Décodage par une des deux expressions de B ci-dessus

Intervalle de codage, précision

- plus petite valeur codable (la plus négative)
  $-2^{n-1} \times 2^{-f}$ de code $[10 \dots 0]$
- code de $0 \rightarrow [0 \dots 0]$
- plus grande valeur codable (positive)
  $(2^{n-1} - 1) \times 2^{-f}$ de code $[01 \dots 1]$
- Précision absolue : valeur de $b_0 = 2^{-f}$
- Erreur maximale du codage : 2^{-f-1}
  pour cela, on code la partie entière de $((2^f \times B) + \frac{1}{2})$

Erreur relative (1)



$$\boxed{E_r = \frac{e}{v} = \frac{v-(k\times 2^{-f})}{v}}$$

1. $v = k \times 2^{-f} \Rightarrow e = 0$

$$\boxed{\Rightarrow E_{rmin} = 0}$$

2. $v = k \times 2^{-f} + 2^{-f-1} \Rightarrow e = 2^{-f-1}$

$$\Rightarrow E_{rmax} = \frac{2^{-f-1}}{k\times 2^{-f}+2^{-f-1}}$$

$$\Rightarrow E_{rmax} = \frac{2^{-1}}{k+2^{-1}}$$

$$\boxed{\Rightarrow E_{rmax} = \frac{1}{2k+1}}$$

$$\boxed{E_r = \frac{e}{v} = \frac{v-(k+1)\times 2^{-f}}{v}}$$

1. $v = (k+1) \times 2^{-f} \Rightarrow e = 0$

$$\boxed{\Rightarrow E_{rmin} = 0}$$

2. $v = k \times 2^{-f} + 2^{-f-1} \Rightarrow e = 2^{-f-1}$

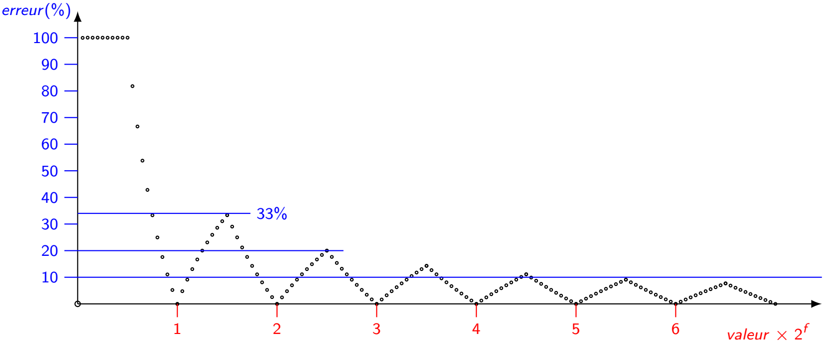
$$\Rightarrow E_{rmax} = \frac{2^{-f-1}}{k\times 2^{-f}+2^{-f-1}}$$

$$\Rightarrow E_{rmax} = \frac{2^{-1}}{k+2^{-1}}$$

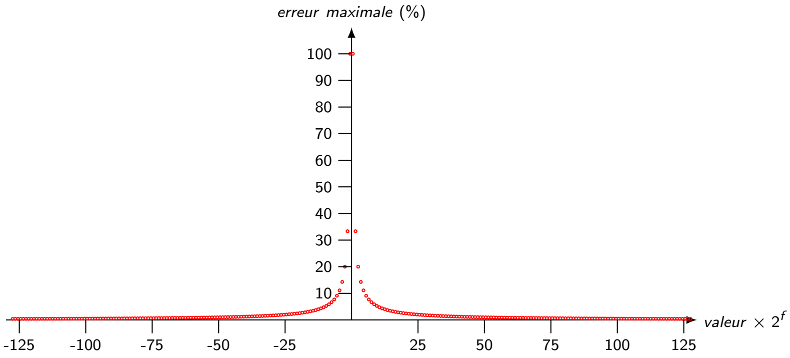
$$\boxed{\Rightarrow E_{rmax} = \frac{1}{2k+1}}$$

$k \times 2^{-f} + 2^{-f-1}$

Erreur relative (2)



Erreur relative (3)



Exemple ilustratif (1)

Soit à coder le nombre $12,125$ en (5,3).
Retour à un codage entier sur 8 bits

$$12,125 \times 2^3 = 97$$

97 est codé par 64+32+1 soit en binaire 01100001 qui est bien le code de 12,125 en virgule fixe (5,3). En effet

- 01100, le code de la partie fixe vaut bien $2^3 + 2^2 = 12$
- le code de la partie fractionnaire 001 vaut bien $2^{-3} = 0.125$

Exemple ilustratif (2)

Soit à coder en (5,3) le nombre $-12,125$

- Première technique (rappel)
  1. Retour à un codage entier : $12,125 \times 2^3 = 97$
  2. $97 + 0,5 = 97,5$ arrondi inférieur à 97
  3. Calcul du complément à deux : $2^8 - (97) = 159$
  4. Codage de 159 $\rightarrow$ 10011111
- Deuxième technique (rappel)
  1. Retour à un codage entier : $12,125 \times 2^3 = 97$
  2. $97 + 0,5 = 97,5$ arrondi inférieur à 97
  3. Codage de 97 $\rightarrow$ 01100001
  4. Complémenter les bits "à gauche du premier 1" $\rightarrow$ 10011111

Exemple ilustratif (3)

Soit à coder en (5,3) le nombre $-12,248$ (sur 8 bits)

- Première technique (rappel)
  1. Retour à un codage entier : $12,248 \times 2^3 = 97,984$
  2. $97,984 + 0,5 = 98,484$, partie entière : 98
  3. Calcul du complément à deux : $2^8 - (98) = 158$
  4. Codage de $158 \rightarrow 10011110$
- Deuxième technique (rappel)
  1. Retour à un codage entier : $12,248 \times 2^3 = 97,984$
  2. $97,984 + 0,5 = 98,484$ arrondi inférieur à $98(= 64 + 32 + 2)$
  3. Codage de $98 \rightarrow 01100010$
  4. Complémenter les bits "à gauche du premier 1" $\rightarrow 10011110$

Erreur absolue de codage : $Ea = |98 - 97,984| \times 2^{-3} = 0.002$

Erreur relative de codage : $Er = \frac{|98-97,984|}{97,984} = \frac{0.002}{12,248} \approx 1.63 \times 10^{-4}$

Exemple illustratif (4)

- Soit à coder 0.215 en (5,3) sur 8 bits
  1. Valeur entière : $0.215 \times 2^3 = 1,72$
     Arrondi :$1,72 + 0,5 = 2,22$, partie entière : 2
  2. Erreur absolue $Ea = |0.215 - 2 \times 2^{-3}| = 0.035$
  3. Erreur relative $Er = \frac{0.035}{0.215} \approx 16\%$
  4. Code de 2 $\rightarrow$ 00000010

- Soit à coder 0.250 en (5,3) sur 8 bits
  1. Valeur entière : $0.250 \times 2^3 = 2,$
     Arrondi : $2 + 0.5 = 2.5$ partie entière : 2
  2. Erreur absolue $Ea = 0$, erreur relative $Er = \frac{0}{2.00} = 0\%$
  3. Code de 2 $\rightarrow$ 00000010

Des erreurs très différentes pour des valeurs relativement proches.