EXERCICE 1: ENCODAGE DES NOMBRES ENTIERS

On considère la base $\beta = 2$.

1. Écrire sur 8 bits, en numération simple à position, les valeurs décimales suivantes : 31, 247
2. Écrire sur 8 bits, en complément à 2, les valeurs décimales suivantes : 44, -9, 0, -31.
3. Définir la plage des valeurs entières représentables avec un variable du type :
   - `short` : 2 octets
   - `int` : 4 octets
   - `long` : 8 octets

EXERCICE 2: ENCODAGE DES NOMBRES FLOTTANTS

1. Rappeler la définition d'un nombre flottant ? Qu'est-ce qu'un nombre flottant normalisé ?
2. On considère la base $\beta = 2$. Convertir en décimal les représentations (signe, mantisse, exposant) suivantes :

$$x = (-1)^s \left(\sum_{i=1}^{r} m_i \beta^{-i} \right) \beta^e$$

	s	m	e
a =	1	10 1101	3
b =	0	1100 0001	-5
c =	0	1 0000 0001	8

3. On considère une base β et soit $x \in \mathbb{R}$. Définir l'arrondi de x, noté $A(x)$, en utilisant r chiffres significatifs pour la mantisse. Définir l'erreur de représentation et la précision machine ε.
4. On considère la base $\beta = 10$. Représenter le réel $231,345$ en virgule flottante avec 4 chiffres significatifs. Calculer les erreurs absolue et relative de représentation.

EXERCICE 3: UN FAUX PARADOXE (FACULTATIF)

1. Montrer qu'en base 10, on a : $1 = 0,9999\ldots$
2. Montrer que, pour toute base β, si $b = \beta - 1$, on a $1 = 0,bbb\ldots$

EXERCICE 4: (FACULTATIF)

1. Soit ε un nombre dont la taille est de l'ordre de la précision machine. Justifier l'approximation :

$$1 - \frac{1}{\varepsilon} = -\frac{1}{\varepsilon}$$

EXERCICE 5: CONDITIONNEMENT D'UN PROBLÈME

Calculer le nombre de conditions absolues et relatives des problèmes suivants :

1. $f : x \to \beta x, \quad \beta \in \mathbb{R}$
2. $f : (x_1, x_2) \to x_1 + x_2$
3. $f : x \to \sqrt{x}$
4. $f : (x_1, x_2) \to x_1 - x_2$

EXERCICE 6: ÉVALUATION D'UNE FONCTION POLYNÔME PAR SCHÉMA DE HORNER

Soit une fonction polynôme de degré n définie par $p(x) = \sum_{i=0}^{n} a_i x^i$. La fonction $p(x)$ peut être écrite sous la forme :

$$p(x) = \sum_{i=0}^{n} a_i x^i = a_0 + x\left\{a_1 + x\left[a_2 + x + \left(\ldots(a_{n-1} + x a_n)\right)\right]\right\} \quad \text{(Schéma de Horner)}$$

Exo 1.

1) 31:

31/2 15 -- 1
15/2 7 -- 1
7/2 3 --- 1
3/2 1 --- 1
1/2 0 ----- 1

31_d = 00011111b

247:

247/2 123 --- 1
123/2 61 --- 1
61/2 30 -- 1
30/2 15 --- 0
15/2 7 --- 1
7/2 3 -- 1
3/2 1 --- 1
1/2 0 -- 1

247_d = 1111 0111 b

256 = 1 0000 0000
247 = 256 − 9
8 4 1

```
  1 1 1 1 1 1 1 1
  1 0000 0000
− 0000 1001
-------------
  0 1111 0111
```

2) 44:

44/2 22 .. 0
22/2 11 -- 0
11/2 5 -- 1
5/2 2 --- 1
2/2 1 --- 0
1/2 0 -- 1

00101100

−9: 9 ⇒

9/2 4 -- 1
4/2 2 --- 0
2/2 1 --- 0
1/2 0 --- 1

9_{10} = 00001001b ⇒ 1111 0110
\+ 1
11110111b $= (-9)_{10}$

$(0)_{10} = 0000\ 0000b.$

-31 :

$31 \Rightarrow$

31/2	15 ... 1	00011111
15/2	7 ... 1	↓
7/2	3 .. 1	11100000
3/2	1 ... 1	↓
1/2	0 ... 1	$11100001b = (-31)_{10}$.

$-256 \sim 255$

3) short 2 octets ⇒ 16 bits.

unsigned short : $[0; 2^{16}-1] = [0, 65535]$

short : $[-32768; 32767]$

int 4 octets ⇒ 32 bits

unsigned : $[0, 2^{32}-1]$

signed : $[-2^{31}, 2^{31}-1]$

Exo 2.

1) Nb flottant = (s, m, e) = $(-1)^s \cdot m \cdot \beta^e$

signe, mantisse, exposant

triplet

$1337 \Rightarrow$ $s = 0$, $m = 1337$, $e = 4$, $\beta 10$ $\quad 0{,}1337 \cdot 10^4$

pour garantir l'unicité,

$0{,}\boxed{1}337 \cdot 10^4$ $\leftarrow$ $s = 0$, $m = 01337$, $e = 5$, $\beta 10$

ce nombre doit être ≠ 0.

Normalisé : ⇒ 1er chiffre de la mantisse est ≠ de 0.
⇒ Garantir l'unicité de la représentation.

2) $a = \{ \overset{s}{1} \quad \overset{m}{101101} \quad \overset{e}{3} \}$ $\longrightarrow$ convertir en décimal

$= (-1)^1 \; 0{,}101101 \cdot 2^3$

$1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} \cdots$

$101101 = (2^{-1} + 2^{-3} + 2^{-4} + 2^{-6}) \cdot 2^3$

$= 2^2 + 2^0 + 2^{-1} + 2^{-3} = 4 + 1 + 0{,}5 + 0{,}125$

$a = -5{,}625$

$b = \{ 0 \quad 11000001 \quad -5 \}$

$b = (2^{-1} + 2^{-2} + 2^{-8}) \cdot 2^{-5}$ $\qquad 0{,}11000001 \quad 2^{-5}$

$= 2^{-6} + 2^{-7} + 2^{-13}$

$= +0{,}02355957$

$c = \{ 0 \quad 100000001 \quad 8 \}$

$c = (2^{-1} + 2^{-9}) \cdot 2^8 = 2^7 + 2^{-1} = 128{,}5$

3) Soit $x \in \mathbb{R}$

Sa représentation en flottant normalisé

$S \quad m \quad e$

avec $m = m_{-1} m_{-2} m_{-3} m_{-4} \cdots m_{-r} m_{-r-1} \cdots$

$A(x)$ = arrondi à r chiffres.

$\hookrightarrow A(x) = (S \; m' \; e)$

$m' = m_{-1} m_{-2} \cdots m_{-r+2} m_{-r+1} m'_{-r}$ et $m'_{-r} = \begin{cases} m_{-r} & \text{si } m_{-r-1} < \frac{\beta}{2} \\ m_{-r}+1 & \text{si } m_{-r-1} \geq \frac{\beta}{2} \end{cases}$

$0{,}13|3|7$

$0{,}13. \quad 0{,}134$

$\beta = 3 \qquad 0,122112$

$A(x)_4 = 0,1221.$ $\qquad 1 < \frac{\beta}{2} = \frac{3}{2}$

$A(x)_5 = 0,12212$ $\qquad 2 > \frac{\beta}{2}$

Erreur de représentation.

$|A(x) - x| =$ erreur absolue.

$\frac{|A(x) - x|}{|x|} =$ erreur relative

Précision machine $\frac{\beta}{2} \cdot \beta^{-r}$ lié à m'_{-r}

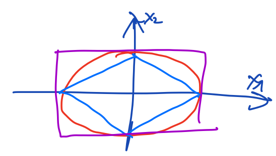
4) $\beta = 10 \qquad x = 231,345$

$x = (-1)^0 \cdot 0,231345 \cdot 10^3$

$A(x)_4 = (-1)^0 \cdot 0,2313 \cdot 10^3$

$\mathcal{E}_a = |A(x)_4 - x| = 0,045$

$\mathcal{E}_r = \frac{|A(x)_4 - x|}{|x|} = \frac{0,045}{231,345} = 2 \cdot 10^{-4}$



norme

$\mathcal{L}_1 : \|X\|_1 = \sum_{i=1}^{n} |X_i|$

$\mathcal{L}_2 : \|X\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$

$\mathcal{L}_\infty : \|X\|_\infty = \max_i (|X_i|)$

Exo 5. Conditionnement.

1) $f : x \to \beta x, \ \beta \in \mathbb{R}$

Soit $x' = x + \Delta x$ $\qquad$ x + une petit erreur.

erreur de troncature

capteur bruité

⋮

$y' = f(x') = \beta(x + \Delta x) = \beta x + \beta \Delta x$

$y' = y + \Delta y$

Comme $x \in \mathbb{R}$ (1 dimension) $\Rightarrow \mathcal{L}_1 = \mathcal{L}_\infty$

$\|x\|_1 = \|x\|_\infty = |x|$

$\|\Delta x\|_1 = \|\Delta x\|_\infty = |\Delta x|$

$\|y\| = \|\beta x\| = |\beta||x|$

$K_a = \frac{\|\Delta y\|}{\|\Delta x\|} = |\beta|$

$K_r = \frac{\|\Delta y\| \ \|x\|}{\|y\| \ \|\Delta x\|} = \frac{|\beta \Delta x| \ |x|}{|\beta x| \ |\Delta x|} = 1$

$$\|\Delta y\| = \|\beta \Delta x\| = |\beta| |\Delta x|$$

erreur reste inchangée en terme de %.

2) $y : (x_1, x_2) \mapsto x_1 + x_2$

$\Delta x = (\Delta x_1, \Delta x_2)$

4)

$y' = x_1 + \Delta x_1 \overset{-}{+} (x_2 + \Delta x_2) = x_1 \overset{-}{+} x_2 \overset{-}{+} \Delta x_1 + \Delta x_2$

$y' = y + \Delta y$

$\|x\|_1 = |x_1| + |x_2|$ $\qquad \|x\|_\infty = \max(|x_1|, |x_2|)$

$\|\Delta x\|_1 = |\Delta x_1| + |\Delta x_2|$ $\qquad \|\Delta x\|_\infty = \max(|\Delta x_1|, |\Delta x_2|)$

$\|y\|_1 = \|y\|_\infty = |x_1 \overset{-}{+} x_2|$

$\|\Delta y_1\| = \|\Delta y\|_\infty = |\Delta x_1 \overset{-}{+} \Delta x_2|$

en norme 1 :

$$k_a = \frac{\|\Delta y\|_1}{\|\Delta x\|_1} = \frac{|\Delta x_1 \overset{-}{+} \Delta x_2|}{|\Delta x_1| + |\Delta x_2|} \leq \frac{|\Delta x_1| + |\Delta x_2|}{|\Delta x_1| + |\Delta x_2|} \Rightarrow k_a \leq 1$$

erreur diminue en absolue

$$k_r = k_a \frac{\|x\|_1}{\|y\|_1} \leq \frac{|x_1| + |x_2|}{|x_1 \overset{-}{+} x_2|}$$

en norme ∞ :

$|\Delta x_1| \leq \max(|\Delta x_1|, |\Delta x_2|)$

$|\Delta x_2| \leq \max(|\Delta x_1|, |\Delta x_2|)$

$$k_a = \frac{\|\Delta y\|_\infty}{\|\Delta x\|_\infty} = \frac{|\Delta x_1 \overset{-}{+} \Delta x_2|}{\max(|\Delta x_1|, |\Delta x_2|)} < \frac{2\max(|\Delta x_1|, |\Delta x_2|)}{\max(|\Delta x_1|, |\Delta x_2|)} \Rightarrow k_a \leq 2.$$

$$k_r = k_a \frac{\|x\|_\infty}{\|y\|_\infty} \leq \frac{2\max(|x_1|, |x_2|)}{|x_1 \overset{-}{+} x_2|}$$

$x_1 = x_2^{+10^{-5}}$ $\qquad y' = x_1 + \Delta x_1 - (x_2 + \Delta x_2)$

$\Delta x_1 = ?$

$\Delta x_2 = ?$ $\qquad y' = \Delta x_1 - \Delta x_2 \qquad y = x_1 - x_2 = 0. \ = 10^{-5}$

$e_a = |y - y'| = |\Delta x_1 - \Delta x_2|$

$$e_r = \frac{|y - y'|}{|y|} = \frac{|\Delta x_1 - \Delta x_2|}{\underset{10^{-5}}{0}}$$ phénomène d'annulation

→ erreur énorme

3) $y = \sqrt{x}$

$x' = x + \Delta x$

$y' = \sqrt{x + \underset{\sim}{\Delta x}}$ petit devant x => DL ordre 1.

$$y' = \sqrt{x} + \frac{d}{dx}(\sqrt{x})\,\Delta x = \underbrace{\sqrt{x}}_{y} + \underbrace{\frac{1}{2\sqrt{x}}\Delta x}_{\Delta y}.$$

$x, \Delta x, y, \Delta y$ en dimension 1 => toutes les normes équivalentes

$\|x\| = |x|$ $\qquad \|y\| = |\sqrt{x}| = \sqrt{x}$

$\|\Delta x\| = |\Delta x|$ $\qquad \|\Delta y\| = \left|\frac{\Delta x}{2\sqrt{x}}\right| = \frac{1}{2\sqrt{x}}|\Delta x|$

$$k_a = \frac{\|\Delta y\|}{\|\Delta x\|} = \frac{1}{2\sqrt{x}}\,\frac{|\Delta x|}{|\Delta x|} = \frac{1}{2\sqrt{x}}$$

$$k_r = k_a \frac{\|x\|}{\|y\|} = \frac{1}{2\sqrt{x}} \cdot \frac{|x|}{\sqrt{x}} = \frac{|x|}{2|x|} = \frac{1}{2}$$ => diviser l'erreur par 2.

$$\|x\|_p^n = \sqrt[p]{\sum_{i=1}^{n} x_i^p}$$

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

1. Écrire l'algorithme qui implémente le schéma de Horner
2. Démontrer que l'algorithme d'évaluation de Horner est linéaire par rapport au degré du polynôme évalué
3. Généraliser au cas :

$$p(x) = \sum_{i=0}^{n} a_i \prod_{j=1}^{i} (x - c_{j-1}) = a_0 + a_1(x - c_0) + a_2(x - c_0)(x - c_1) + \ldots + a_n(x - c_0)(x - c_1)\ldots(x - c_{n-1})$$

Exercice 7: Évaluation des fonctions complexes

Proposer une méthode pour calculer les fonctions suivantes en utilisant les fonctions élémentaires $(+, -, *, /)$

1. $\sqrt{x}$, $x > 0$
2. $\sin(x)$, $\cos(x)$
3. e^x
4. $\log(x)$, $x > 0$

Exercice 8: Calcul des fonctions et erreurs de représentation

Proposer une méthode pour éviter une perte de précision dans les calculs suivants :

1. $e^x - \sin(x) - \cos(x)$
2. $\log(x) - 1$
3. $\log(x) + \log(1/x)$