

Algorithmique

Travaux dirigés, 22 octobre 2004

Louis Granboulan

1 Arithmétique flottante

1. Étude théorique

On se donne quatre entiers : une base B , une longueur de mantisse n et deux exposants extrémaux E_{\min} et E_{\max} . Habituellement, $B \geq 2$, $n \geq 2$ et $E_{\min} < 0 < E_{\max}$. L'écriture en virgule flottante d'un réel x est $\epsilon_x . x_0 x_1 \dots x_{n-1} B^{e_x}$ avec $\epsilon_x = \pm 1$, $E_{\min} \leq e_x \leq E_{\max}$ et $0 \leq x_i < B$, si on a $x = \epsilon_x (\sum_{i=0}^{n-1} x_i B^{-i}) B^{e_x}$. On appelle *mantisse* la valeur $f_x = \sum_{i=0}^{n-1} x_i B^{-i}$.

- (a) Bien évidemment, tous les réels ne sont pas représentables exactement. Calculer quels sont le plus petit et le plus grand réels positifs représentables exactement.
- (b) L'addition, soustraction, multiplication, ... de deux réels exactement représentables n'est pas toujours un réel exactement représentable. On décide par exemple d'arrondir au plus grand réel exactement représentable inférieur au résultat. Que donne l'algorithme suivant ? On suppose que E_{\max} est suffisamment grand pour éviter les *overflow*.

```

a ← 1.0      b ← 1.0
tant que ((a + 1.0) - a) - 1.0 == 0 faire a ← a + a
tant que ((a + b) - a) - b <> 0 faire b ← b + 1.0
imprimer b
    
```

2. Norme IEEE 754

Cette norme étend la notation précédente comme suit : la base $B = 2$ et on note $f_x = \sum_{i=1}^n x_i 2^{-i}$. L'exposant e_x peut varier dans l'intervalle $[E_{\min} - 1, E_{\max} + 1]$.

Exposant	Mantisse	Valeur		
$E_{\min} - 1$	$f_x = 0$	± 0	float	double
$E_{\min} - 1$	$f_x \neq 0$	$\pm f_x 2^{E_{\min}}$	32 bits	64 bits
$E_{\min} \leq e_x \leq E_{\max}$		$\pm (1 + f_x) 2^{e_x}$	$n = 23$	$n = 52$
$E_{\max} + 1$	$f_x = 0$	$\pm \infty$	$E_{\min} = -126$	$E_{\min} = -1022$
$E_{\max} + 1$	$f_x \neq 0$	NaN	$E_{\max} = +127$	$E_{\max} = +1023$

Les arrondis des opérations peuvent se faire de plusieurs façons : au plus proche, vers 0, vers le haut ou vers le bas.

- (a) Comparer cette représentation avec la représentation précédente.

2 Calcul d'une somme

1. Série harmonique

- (a) On définit la suite $u_0 = 0$ et $u_n = u_{n-1} + \frac{1}{n}$. Montrer que le calcul de cette suite en représentation flottante converge.
- (b) Trouver une astuce pour calculer assez précisément $\sum_{n=1}^N \frac{1}{n}$. Évaluer la précision du résultat.

2. Encadrement du résultat

- (a) Décrire des techniques de calcul d'un encadrement de $\sum_{n=1}^N a_n$.
- (b) En déduire quelques considérations sur la représentation d'un réel par un intervalle.

3. Double précision

On représente un réel par une somme de deux flottants. On veut normaliser $A + B$ sous la forme $X + x$ de telle sorte que la valeur absolue de x soit la plus petite possible. On suppose que les opérations sur les flottants sont arrondies au plus proche.

- (a) Exprimer la condition de normalisation sous la forme $|x| \leq \frac{1}{2} \alpha(X)$.
- (b) Trouver comment faire cette normalisation en six opérations dans le cas général et en trois opérations (additions et soustractions) si $|B|$ est suffisamment petit.
- (c) En déduire comment faire une addition en vingt opérations.
- (d) Étudier, dans cette représentation, les autres opérations habituelles.