Correction des exercices du TD 1

Exercice 1.1

Soit un système de représentation flottant avec base β, mantisse à p places, exposant E avec $E_m \leq E \leq E_M$. Calculer le nombre de valeurs normalisées qui peuvent être représentées par ce système.
Application : $\beta = 10, p = 3, E_m = -15, E_M = 16$.

Un nombre normalisé en virgule flottante est un nombre dans lequel le chiffre qui précède la virgule est un 0 et le chiffre qui suit la virgule n'est pas un 0. Ce 0 est omis dans l'écriture de la mantisse et le nombre est stocké comme un entier.

Un nombre type n codé en représentation normalisée s'exprime :

$$n = (-1)^s 0.a_1 a_2 ... a_p x \beta^E \text{ avec } E_m \leq E \leq E_M$$

nombre d'exposants disponibles : $E_M - E_m + 1$

nombre de nombres de la forme $0.a_1 a_2 ... a_p$: $(\beta - 1)\beta^{p-1}$ puisque chaque chiffre a_i évolue entre 0 et $\beta - 1$ sauf a_1 qui ne peut être nul et évolue donc entre 1 et $\beta - 1$

nombre de signes possibles : 2

Il faut y ajouter 2 zéros (le positif et le négatif), soit au total :

$$2(\beta - 1)\beta^{p-1}(E_M - E_m + 1) + 2$$

Application :

$$\begin{aligned} 2(\beta - 1)\beta^{p-1}(E_M - E_m + 1) + 2 &= 2.9.10^2(16 + 15 + 1) + 2 \\ &= 8100.32 + 2 = 259202 \end{aligned}$$

Exercice 1.2

Soient trois réels $x = 0.125 \times 10^6, z = 0.437 \times 10^{12}, w = 0.215 \times 10^{-10}$. En utilisant le système de numération de l'exercice précédent pour le stockage de ces nombres, calculer :

1. La somme $x + z$ et commenter le résultat
2. Le produit $x \times z$ et commenter le résultat
3. La division w/x indiquant qu'il y a underflow

Solution :

1. $x + z = 0.125 \times 10^6 + 0.437 \times 10^{12} = (0.000000125 + 0.437)\, 10^{12}$ (alignement)
   $= 0.437000125 \times 10^{12} = 0.437 \times 10^{12}$ (standardisation avec $m = 3$)$= z$
   Conclusion : $x + z = z$ et pourtant $x \neq 0$!
2. $x \times z = 0.125 \times 10^6 \times 0.437 \times 10^{12} = 0.125 \times 0.437 \times 10^{18} = 0.054625 \times 10^{18} = 0.54625 \times 10^{17}$ (normalisation)$=$ 0.546×10^{17} (standardisation)
   Conclusion : l'exposant est supérieur à l'exposant maximal utilisable $E_M = 16$ donc le nombre $x \times z$ est supérieur au plus grand nombre représentable ; il y a donc dépassement. Ceci se traduira par un message d'erreur affiché par le calculateur indiquant qu'il y a overflow.
3. $w/z = (0.215 \times 10^{-10}) / (0.437 \times 10^{12}) = 0.215/0.437 \times 10^{-22} = 0.49199084668192219 \times 10^{-22} = 0.491 \times 10^{-22}$ (standardisation)

   Conclusion : l'exposant est inférieur à l'exposant minimla utilisable E_m et le quotient w/z est donc plsu petit que le plus petit nombre positif représentable (1.100×10^{-15}) ; il y a donc dépassement. Ceci se traduira par un message d'erreur affiché par le calculateur indiquant qu'il y a underflow.

Exercice 1.3

Soient trois réels $x = 0.400 \times 10^0 = y, z = 0.100 \times 10^3$. En utilisant le système de numération de l'exercice précédent pour le stockage de ces nombres, calculer les deux sommes $(x+y)+z$ et $x+(y+z)$.

Solution :

$(x+y)+z = 0.800\times10^0+0.100\times10^3 = 0.0008\times10^3+0.100\times10^3 = 0.001\times10^3(\text{standardisation})+0.100\times10^3 = 0.101 \times 10^3$

$y+z = 0.400 \times 10^0 + 0.100 \times 10^3 = 0.000400 \times 10^3 + 0.100 \times 10^3 = 0.100 \times 10^3(\text{standardisation})$ donc $x+(y+z) = 0.400 \times 10^0 + 0.100 \times 10^3 = 0.100 \times 10^3 = z$

Conclusion : l'ordre dans lequel sont effectuées les opérations peut avoir une incidence sur le résultat final. Lorsqu'on additionne des nombres positifs, il faut commencer par les plus petits.

Exercice 1.7

Rappel des données :

- $x = s.m.b^e$ avec $\frac{1}{b} \leq m < 1$, où :
  - s est le signe
  - m est la mantisse sans limitation de bits
  - b est la base
  - e est l'exposant, entier.
- la représentation de x en machine est $m(x) = s.M.b^E$ avec $m(x) \in M(\mathbb{R})$, où :
  - M est la mantisse limitée à p digits
  - E est l'exposant, limité à q digits.

Le codage en nombres flottants requiert ainsi $N = p+q+1$ digits.
On considère $b = 2$, on travaille en binaire.
Le problème consiste à calculer l'erreur de représentation pour différentes opérations arithmétiques.

4. Cas de l'affectation

   L'erreur est donnée par

$$\begin{aligned} e &= x - m(x) \\ &= s.m.2^e - s.M.2^E \\ &= s.2^E(m.2^{e-E} - M) \end{aligned}$$

   Il peut y avoir troncature de l'exposant donc on a toujours $E \leq e$, soit $e - E \geq 0$ donc $2^{e-E} \geq 1$. Dans ce cas, l'exposant étant toujours supposé positif, on obtiendra un nombre d'exposant plus faible que x, l'erreur sera donc plus importante que celle sans troncature d'exposant, mais impossible à minorer. On parle dans ce cas de dépassement de capacité.

   On suppose désormais qu'il n'y a pas de troncature sur l'exposant.

   L'erreur qui se produit sur la mantisse est due à la troncature appliquée au $(p+1)^{ième}$ digit, ou à l'arrondi effectué sur le $p^{ième}$ digit, elle peut donc se formaliser comme

$$e_m = s.\alpha.2^E.2^{-p}$$

   où : - $\alpha \in [0, 1[$ s'il y a troncature

   - $\alpha \in [-0.5, 0.5[$ s'il y a arrondi.

   En supposant qu'il n'y a pas d'approximation sur l'exposant et en utilisant la notation $\Delta(x)$ pour l'erreur de représentation, on obtient :

$$m(x) = x - s.\Delta(x).2^{E-p}$$

5. Erreur de l'opération d'addition

   Soient deux nombres x_1 et x_2 dont les représentations machine sont respectivement $m(x_1)$ et $m(x_2)$. La somme des nombres en machine est calculée par :

$$\begin{aligned} m\left[m(x_1)+m(x_2)\right] &= \left[x_1 - s_1.\Delta(x_1).2^{E_1-p} + x_2 - s_2.\Delta(x_2).2^{E_2-p}\right] - s_s.\Delta(x_s).2^{E_s-p} \\ &= x_1 + x_2 - s_1.\Delta(x_1).2^{E_1-p} - s_2.\Delta(x_2).2^{E_2-p} - s_s.\Delta(x_s).2^{E_s-p} \end{aligned}$$

   Remarque : L'erreur globale est l'accumulation de chacune des erreurs de représentation.

6. Erreur de l'opération de soustraction

Comme précédemment pour l'addition, on obtient pour la différence :

$$\begin{aligned} m\left[m(x_1) - m(x_2)\right] &= \left[x_1 - s_1.\Delta(x_1).2^{E_1-p} - x_2 + s_2.\Delta(x_2).2^{E_2-p}\right] - s_s.\Delta(x_s).2^{E_s-p} \\ &= x_1 - x_2 - s_1.\Delta(x_1).2^{E_1-p} + s_2.\Delta(x_2).2^{E_2-p} - s_s.\Delta(x_s).2^{E_s-p} \end{aligned}$$

Remarque : L'erreur se comporte comme pour la somme.

7. Erreur de l'opération de multiplication

En réutilisant les mêmes notations, on obtient pour $x_1 \times x_2$:

$$\begin{aligned} m(m(x_1).m(x_2)) &= (x_1 - s_1.\Delta(x_1).2^{E_1-p})(x_2 - s_2.\Delta(x_2).2^{E_2-p}) - s_m.\Delta(x_m).2^{E_m-p} \\ &= x_1x_2 - s_1.x_2.\Delta(x_1).2^{E_1-p} - s_2.x_1.\Delta(x_2).2^{E_2-p} \\ &\quad + s_1.\Delta(x_1).2^{E_1-p}s_2.\Delta(x_2).2^{E_2-p} - s_m.\Delta(x_m).2^{E_m-p} \end{aligned}$$

En négligeant le terme d'ordre 2 en $\Delta(x_1).\Delta(x_2)$, on obtient :

$$m(m(x_1).m(x_2)) = x_1x_2 - s_1.x_2.\Delta(x_1).2^{E_1-p} - s_2.x_1.\Delta(x_2).2^{E_2-p} - s_m.\Delta(x_m).2^{E_m-p}$$

8. Erreur de l'opération de division

Pour x_1/x_2 il vient :

$$\begin{aligned} m(m(x_1)/m(x_2)) &= (x_1 - s_1.\Delta(x_1).2^{E_1-p})/(x_2 - s_2.\Delta(x_2).2^{E_2-p}) - s_d.\Delta(x_d).2^{E_d-p} \\ &= x_1/[x_2 - s_2.\Delta(x_2).2^{E_2-p}] - s_1.\Delta(x_1).2^{E_1-p}/[x_2 - s_2.\Delta(x_2).2^{E_2-p}] \\ &\quad - s_d.\Delta(x_d).2^{E_d-p} \end{aligned}$$

Utilisons les DL au voisinage de zéro : si on considère deux nombres x et α tels que $\alpha \ll x$, on a

$$\begin{aligned} (x+\alpha)^{-1} &= \left[x(1+\frac{\alpha}{x})\right]^{-1} \\ &= x^{-1}(1+\varepsilon)^{-1} \end{aligned}$$

où $\varepsilon = \frac{\alpha}{x}$ est petit devant 1, et on sait que $(1+\varepsilon)^{-1} \simeq 1-\varepsilon$. On en déduit que

$$(x+\alpha)^{-1} \simeq x^{-1}(1-\frac{\alpha}{x}) = x^{-1} - \alpha x^{-2}$$

Appliquons ce résultat au calcul de $x_1/[x_2 - s_2.\Delta(x_2).2^{E_2-p}]$:

$$\begin{aligned} x_1/[x_2 - s_2.\Delta(x_2).2^{E_2-p}] &= x_1.\left[x_2^{-1} + s_2.\Delta(x_2).2^{E_2-p}.x_2^{-2}\right] \\ &\simeq x_1.x_2^{-1} + s_2.x_1.\Delta(x_2).2^{E_2-p}.x_2^{-2} \end{aligned}$$

De même pour $s_1.\Delta(x_1).2^{E_1-p}/[x_2 - s_2.\Delta(x_2).2^{E_2-p}]$:

$$[x_2 - s_2.\Delta(x_2).2^{E_2-p}]^{-1} = x_2^{-1} + s_2.\Delta(x_2).2^{E_2-p}x_2^{-2}$$

d'où

$$\begin{aligned} s_1.\Delta(x_1).2^{E_1-p}/[x_2 - s_2.\Delta(x_2).2^{E_2-p}] &= s_1.\Delta(x_1).2^{E_1-p}.x_2^{-1} + s_1.s_2.\Delta(x_1).\Delta(x_2).2^{E_1-p}2^{E_2-p}x_2^{-2} \\ &\simeq s_1.\Delta(x_1).2^{E_1-p}.x_2^{-1} \end{aligned}$$

en négligeant le terme d'ordre 2. Soit :

$$\begin{aligned} m(m(x_1)/m(x_2)) &= x_1.x_2^{-1} - s_1.\Delta(x_1).2^{E_1-p}.x_2^{-1} + s_2.x_1.\Delta(x_2).2^{E_2-p}.x_2^{-2} \\ &\quad - s_d.\Delta(x_d).2^{E_d-p} \end{aligned}$$

Remarque : Au cas où le réel x_2 est petit et le réel x_1 grand, la quantité $s_2.x_1.\Delta(x_2).2^{E_2-p}.x_2^{-2}$ peut devenir importante et perturber la valeur obtenue. Il faut donc éviter d'effectuer de telles divisions.