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, si $\beta > 2$, 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([1]).

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

$$n = (-1)^s 0.a_1a_2...a_p \times \beta^E \text{ avec } E_m \leq E \leq E_M$$

On a

- nombre d'exposants disponibles : $E_M - E_m + 1$. En réalité, pour les nombres normalisés il faut ne pas tenir compte de l'exposant 00...0 (nombres sous-normalisés) et de l'exposant 11...1 (NaN). Donc exposants disponibles pour le codage : $E_M - E_m - 1$.
- nombre de nombres de la forme $0.a_1a_2...a_p$: $(\beta - 1)\beta^{p-1}$ puisque chaque chiffre a_i évolue entre 0 et $\beta - 1$ sauf si 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 :

  $$\begin{aligned} 2(\beta - 1)\beta^{p-1}(E_M - E_m + 1) + 2 &= 2.9.10^2(16 + 15 - 1) + 2 \\ &= 54000 + 2 = 54002 \end{aligned}$$

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 $x \neq 0$!

[1] Rappelons que si $\beta = 2$, le nombre qui précède la virgule est un 1 qui ne sera pas stocké mais toujours pris en compte lors du décodage (bit caché), sauf si l'exposant est 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)$= 0.546 \times 10^{17}$ (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. $w/z = (0.215 \times 10^{-10})/\left(0.437 \times 10^{12}\right) = 0.215/0.437 \times 10^{-22} = 0.49199084668192219 \times 10^{-22} = 0.491 \times 10^{-22}$ (standardisation)

   Conclusion : l'exposant est inférieur à l'exposant minimal utilisable E_m et le quotient w/z est donc plus 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 :

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

$y + z = 0.400 \times 10^0 + 0.100 \times 10^3 = 0.000400 \times 10^3 + 0.100 \times 10^3 = 0.100 \times 10^3$(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 1.4

Soit le nombre 387.62000_{10}.

1. Calculer sa valeur en hexadécimal
   Rappelons la correspondance entre chiffres hexadécimaux :

10	11	12	13	14	15
A	B	C	D	E	F

   La partie entière est 387. .
   $387_{10} = 24 \times 16 + 3 = (1 \times 16 + 8) \times 16 + 3 = 1 \times 16^2 + 8 \times 16^1 + 3 \times 16^0 = 183_{16}$.
   La partie décimale est 0.62000.
   $0.62000 \times 16 = 9.92$ dont la partie entière est 9.
   $0.92 \times 16 = 14.72$ dont la partie entière est 14.
   $0.72 \times 16 = 11.52$ dont la partie entière est 11.
   $0.52 \times 16 = 8.32$ dont la partie entière est 8.
   $0.32 \times 16 = 5.12$ dont la partie entière est 5.
   $0.12 \times 16 = 1.92$ dont la partie entière est 1.
   On s'arrête car on a atteint le nombre de chiffres décimaux fournis dans le nomre initial, plus un chiffre qui permet de connaître l'arrondi à effecter.
   On a donc $0.62000 = \frac{9}{16} + \frac{14}{16^2} + \frac{11}{16^3} + \frac{8}{16^4} + \frac{5}{16^5} + \frac{0.12}{16^6}$ qui s'écrit en se limitant à 5 chiffres par arrondi : $0,9EB85_{16}$.
   Donc 387.62000_{10} se convertit en $183,9EB85_{16}$.
2. Convertir la valeur héxadécimale en décimale.
   Conversion de la partie entière :
   $183_{16} = 1 \times 16^2 + 8 \times 16^1 + 3 \times 16^0 = 387_{10}$
   Convertion de la partie décimale :
   $0,9EB85_{16} = \frac{9}{16} + \frac{14}{16^2} + \frac{11}{16^3} + \frac{8}{16^4} + \frac{5}{16^5} = 0,61999_{10}$ en se limitant aux 5 chiffres décimaux présents au départ (i.e. en gardant la même précision).
   Donc $183,9EB85_{16}$ se convertit en $387,61999_{10}$.
3. Calculer l'erreur absolue de représentation.

$$|\Delta x| = |m(x) - x| = |387.62000 - 387,61999| = 10^{-5}$$

4. Calculer l'erreur relative absolue de précision.

$$\eta(x) = \frac{|\Delta x|}{|x|} = \frac{10^{-5}}{387.62000} \simeq 2,58.10^{-8}$$

Exercice 1.5

Considérons une machine décimale avec mantisse à 4 chiffres. Calculer l'erreur de représentation et l'erreur relative de représentation pour les nombres :

1. $a = 9,023506$
2. $b = 158,26$
3. $c = 0,001588946$

SOLUTION

On suppose que la machine transforme en notation normalisée décimale avec arrondi et non troncature.

1. Le nombre a est représenté en notation normalisée par $0,9024.10^1$ donc $\Delta a = 0,9024.10^1 - 9,023506 = -0,000494 = -4,94.10^{-4}$.

$$\iota(a) = \frac{\Delta a}{m(a)} = \frac{-4.94 \times 10^{-4}}{0.9024 \times 10^1} = -5.4743 \times 10^{-5}$$

2. Le nombre b est représenté en notation normalisée par $0,1583.10^3$ donc $\Delta b = 0,1583.10^3 - 158,26 = 0,04 = 4.10^{-2}$

$$\iota(b) = \frac{\Delta b}{m(b)} = \frac{4 \times 10^{-2}}{0.1583 \times 10^3} = 2.526\,8 \times 10^{-4}$$

3. Le nombre c est représenté en notation normalisée par $0,1589.10^{-2}$ donc $\Delta c = 0.1589 \times 10^{-2} - 0.001588946 = 5.4 \times 10^{-8}$

$$\iota(c) = \frac{\Delta c}{m(c)} = \frac{5.4 \times 10^{-8}}{0.1589 \times 10^{-2}} = 3.3984 \times 10^{-5}$$

Exercice 1.6

Soient les nombres $a = 15,2750$, $b = 358,937$ et $c = 5233,618$.

1. Convertir ces nombres en hexadécimal
   Cas de a.
   Partie entière : $15_{10} = F_{16}$
   Partie décimale :
   $0,2750 \times 16 = 4,4$ dont la partie entière est 4
   $0,4 \times 16 = 6,4$ dont la partie entière est 6
   $0,4 \times 16 = 6,4$ dont la partie entière est 6....
   Donc $a = 15,2750_{10} = F,46\underline{6}..._{16}$. Quand on convertit, on garde le même nombre de chiffres en hexadécimal qu'en décimal donc on a

$$15,2750_{10} = F,4666_{16}$$

   Cas de b.
   Partie entière : $358 = 22 \times 16 + 6 = 1 \times 16^2 + 6 \times 16 + 6 = 166_{16}$
   Partie décimale :
   $0,937 \times 16 = 14,992$ dont la partie entière est $14 = E$
   $0,992 \times 16 = 15,872$ dont la partie entière est $15 = F$
   $0,872 \times 16 = 13,952$ dont la partie entière est $13 = D$

Donc $b = 358,937$ se convertit en $166,EFD_{16}$

Cas de c

Partie entière : $5233 = 327 \times 16 + 1 = (20 \times 16 + 7) \times 16 + 1 = ((1 \times 16 + 4) \times 16 + 7) \times 16 + 1 = 1 \times 16^3 + 4 \times 16^2 + 7 \times 16 + 1 = 1471$

Partie décimale :

$0,618 \times 16 = 9,888$ dont la partie entière est 9

$0,888 \times 16 = 14,208$ dont la partie entière est $14 = E$

$0,208 \times 16 = 3,328$ dont la partie entière est 3

Donc $c = 5233,618$ se convertit en $1471,9E3_{16}$

2. Convertir ces nombres hexadécimaux en nombres décimaux avec exactitude de 8 (i.e. 8 chiffres significatifs).

   Cas de a.

   $F,4666_{16} = 15 + 4 \times 16^{-1} + 6 \times 16^{-2} + 6 \times 16^{-3} + 6 \times 16^{-4} = 15,274993_{10}$

   Cas de b.

   $166,EFD_{16} = 358 + 14 \times 16^{-1} + 15 \times 16^{-2} + 13 \times 16^{-3} = 358,93676_{10}$

   Cas de c.

   $1471,9E3_{16} = 5233 + 9 \times 16^{-1} + 14 \times 16^{-2} + 3 \times 16^{-3} = 5233,6179_{10}$

3. Calculer l'erreur relative de cette dernière conversion sous la forme $x.xx \times 10^{-7}$

   $\Delta a = m(a) - a = 15,274993 - 15,2750 = -7 \times 10^{-6}$

   $\eta(a) = \frac{\Delta a}{a} = -\frac{7\times 10^{-6}}{15,2750} \simeq -4,58 \times 10^{-7}$

   $\Delta b = m(b) - b = 358,93676 - 358,937 = -2,4 \times 10^{-4}$

   $\eta(b) = \frac{\Delta b}{b} = \frac{-2,4\times 10^{-4}}{358,937} \simeq -6,69 \times 10^{-7}$

   $\Delta c = m(c) - c = 5233,6179 - 5233,618 = -1 \times 10^{-4}$

   $\eta(c) = \frac{\Delta c}{c} = \frac{-1\times 10^{-4}}{5233,618} = -1,911 \times 10^{-8} \simeq -0,19 \times 10^{-7}$

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 $m(x) \in M(\ )$, 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.

1. 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 :

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

2. 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 :

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

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

3. Erreur de l'opération de soustraction

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

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

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

4. Erreur de l'opération de multiplication

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

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

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

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

5. Erreur de l'opération de division

   Pour x_1/x_2 il vient :

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

   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 :

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

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.