Correction des exercices du TD 2

Exercice 2.1 et 2.3

1. Calculer pour le standard IEEE-754 en simple précision
   (a) le biais,caractérisant l'intervalle de variation de la valeur de l'exposant $B = 2^{q-1} - 1$ où q représente le nombre de bits de l'exposant ;
   (b) les valeurs l_- et l_+, respectivement égales à $l_- = B + 1$ et $l_+ = B$;
   (c) la valeur de la plus grande mantisse ;
   (d) le plus grand nombre que nous pouvons représenter ;
   (e) le plus petit nombre positif que nous pouvons représenter ;
   (f) la précision des valeurs numériques représentées par ce standard
2. Répéter pour la double précision les calculs précédents.

Comparaison de différentes quantités dans le standard IEEE-754 suivant la simple ou double précision.
On rappelle les données :

donnée et nombre de bits	simple précision	double précision
s	1	1
p	23	52
q	8	11

question	quantité calculée	Simple précision	Double précision
1	Biais $B = 2^{q-1} - 1$	$2^{8-1} - 1 = 127$	$2^{11-1} - 1 = 1023$
2	$l_- = -biais + 1$	$-127 + 1 = -126$	$-1023 + 1 = 1022$
	$l_+ = biais$	127	1023
3	plus grande mantisse	$\underbrace{111...1}_{23 \text{ fois}}$	$\underbrace{111...1}_{52 \text{ fois}}$
4	plus grand nombre représentable	$\underbrace{11111110}_{8 \text{ bits}}\ \underbrace{111...1}_{23 \text{ fois}}$	$\underbrace{11111111110}_{11 \text{ bits}}\ \underbrace{111...1}_{52 \text{ fois}}$
5	plus petit nombre positif représentable	$\underbrace{00000001}_{8 \text{ bits}}\ \underbrace{000...0}_{23 \text{ fois}}$	$\underbrace{00000001}_{11 \text{ bits}}\ \underbrace{000...0}_{52 \text{ fois}}$
6	précision = epsilon machine $\varepsilon = 2^{-p}$	$\varepsilon = 2^{-p} = 2^{-23} = 1,192.10^{-7}$	$\varepsilon = 2^{-p} = 2^{-52} = 2,22.10^{-16}$

Le plus grand nombre représentable peut se calculer. En simple précision, il correspond à $1,\underbrace{111...1}_{23 \text{ fois}} \times 10^{11111110}_{(2)} = \left(1 + \sum_{i=1}^{23} 2^{-i}\right).2^{2^8-2-127} = 1,9999998808.2^{127} = 3,4.10^{38}$. En double précision, on obtient $1,\underbrace{111...1}_{52 \text{ fois}} \times 10^{11111111110} = \left(1 + \sum_{i=1}^{52} 2^{-i}\right).2^{2^{11}-2-1023} = 1,7976931348.10^{308}$.

Le plus petit nombre positif représentable peut se calculer. En simple précision, il correspond à $1,\underbrace{000...0}_{23 \text{ fois}} \times 10^{00000001}_{(2)} = 2^{1-127} = 2^{-126} = 1,17549435.10^{-38}$. En double précision, on obtient $1,\underbrace{000...0}_{52 \text{ fois}} \times 10^{00000001}_{(2)} = 2^{1-1023} = 2^{-1022} = 2,2250738.10^{-308}$

Exercice 2.2 et 2.4

Représentation au standard IEEE-754 de $5,75$ et $0,1$, en simple et double précision.

3. Cas de $5,75$

   Partie entière : $5 = 4 + 1 = 2^2 + 2^0 = 101_{(2)}$.
   Partie décimale (en reprenant les résultats de l'ex. 1.5). $0,75_{(10)} = 2^{-1} + 2^{-2} = 0,11_{(2)}$
   Nombre : $5,75 = 101,11_{(2)} = 1,0111.10^{10}_{(2)} = 1,0111_{(2)}.2^2_{(10)}$ représenté en simple précision, en décalant l'exposant de 127 par $2^{2+127}_{(10)} = 2^{129}_{(10)} = 10000001_{(2)}$
   Le même nombre est représenté en double précision, en décalant l'exposant de 1023 donc avec $2^{2+1023}_{(10)} = 2^{1025}_{(10)} = 10000000001_{(2)}$
   On en déduit :
   - en simple précision : $\underbrace{10000001}_{8 \text{ bits}}\ \underbrace{01110...0}_{23 \text{ bits}}$
   - en double précision $\underbrace{10000000001}_{11 \text{ bits}}\ \underbrace{01110...0}_{52 \text{ bits}}$

   Il n'y a aucune erreur de représentation, ni en simple ni en double précision.

4. Cas de $0,1$

   Partie entière : nulle
   Partie décimale (en reprenant les résultats de l'ex. 1.5). $0,1_{(10)} = 0,000\underline{1100}_{(2)} = 1,\underline{10011}_{(2)} \times 2^{-4}_{(10)}$ représenté en simple précision, en décalant l'exposant de 127 par $2^{-4+127}_{(10)} = 2^{123}_{(10)} = 2^6+2^5+2^4+2^3+2+1$.
   Le même nombre est représenté en double précision, en décalant l'exposant de 1023 donc avec $2^{-4+1023}_{(10)} = 2^{1019}_{(10)} = 2^9 + 2^8 + 2^7 + 2^6 + 2^5 + 2^4 + 2^3 + 2 + 1$.On en déduit :
   - en simple précision : $\underbrace{01111011}_{8 \text{ bits}}\ \underbrace{10011001100110011001101}_{23 \text{ bits}}$ (le chiffre 1 final est dû à l'arrondi puisque le zéro initial aurait été suivi d'un 1)
   - en double précision $\underbrace{01111111011}_{11 \text{ bits}}\ \underbrace{1001...1001}_{52 \text{ bits}}$ (la séquence 1001 est répétée 13 fois dans la mantisse)

   L'erreur de représentation en simple précision est :
   $\Delta(x) = m(x) - x = 1,10011001100110011001101 \times 2^{-4}_{(10)} - 0,1 = 0,099999940395355225 - 0,1 \simeq 5,96.10^{-9}$

Exercice 2.5 laissé à faire à la maison