MP2I TD flottants

Dans tout ce TP, les *nombres flottants* respectent le standard IEE754 et sont définis par une séquence de $1 + E + M$ bits $(1, E, M)$: quantité des bits de signe, d'exposant et de mantisse. Sauf mention du contraire, les applications numériques se font au format dit *du TP* avec $\mathbf{E = 4, M = 8}$.

Lorsqu'il y a un risque de confusion, on met en indice la base dans laquelle est écrite un nombre. Pour un nombre $x \in \mathbb{R}$, l'écriture en base 10 de x est notée x_{10} et x_2 est son écriture binaire. Par exemple 101_2 désigne cinq et 101_{10} représente cent un.

Dans ce TD, pour un nombre x on note s_x, e_x, m_x son signe, sa mantisse et son exposant (ou s, e, m s'il n'y a a pas de risque de confusion). Si $x \neq 0$, le triplet (s, e, m) correspond à l'écriture scientifique de x : $(-1)^s(1 + m)2^e$.

On note $f_{E,M}$ la fonction qui associe au réel x son écriture en machine dans le standard IEEE754 avec E bit d'exposant et M bits de mantisse (ou $f(x)$ s'il n'y a pas de risque de confusion). On note $g_{E,M}$ la fonction qui passe des flottants aux reels (ou plus simplement g s'il n'y pas de risque de confusion).

Un nombre réel x est dit *représentable en machine*, si on a l'égalité $g(f(x)) = x$.

Pour $x \in \mathbb{R}$, Le nombre réel $g(f(x))$ représentable en machine qui a la même écriture que x comme flottant au format E, M est noté plus simplement $\bar{x}$ ou $\bar{x}_{E,M}$ s'il y a un risque de confusion. En d'autres termes, $\bar{x}$ est le nombre représentable en machine le plus proche de x et vérifiant $f(x) = f(\bar{x})$.

Les additions, multiplications, soustractions et divisions entre réels sont notées $\{+, \times, -, \div\}$ et les opérations correspondantes entre flottants $\{\oplus, \otimes, \ominus, \oslash\}$. La division du réel a par le réel b est notée indifféremment $a \div b$ ou $\frac{a}{b}$. L'écriture a/b désigne dans ce devoir la division euclidienne entre deux entiers.

Le mode d'arrondi considéré dans ce sujet est l'arrondi *au plus proche pair* qui considère les 3 chiffres après le dernier bit maintenu.

Q.1 Rappeler les valeurs E, M pour les flottants dits *simple précision* (les `float` de C).

Q.2 Pour le format $(1, E, M)$, quelle est la valeur du décalage $d_{E,M}$ de l'exposant ? Application au format du TP $d_{4,8}$.

Q.3 Passage des réels aux flottants. Calculer $f_{4,8}(\frac{1}{5})$.

Q.4 Représenter $-\infty$ au format du TP puis traduire ce flottant en réel.

Q.5 Donner un exemple de NaN au format du TP en binaire IEE754 puis en base 10.

Q.6 Passage des flottants aux réels. Voici un nombre flottant au format du TP :

1 1010 10101010

Le transformer en réel.

Q.7 Autour de la notion de *successeur*.

Soit un format de flottant $(1, E, M)$ et $M' \in \mathbb{R}^+$ le plus grand nombre représentable de ce format.

Soit $x \in [0, M'[$ représentable en machine. Montrer qu'il existe un nombre y représentable en machine tel que $y > x$ et y est le plus petit dans ce cas.

Dorénavant, on appelle *successeur de x* et on note $s(x)$ le nombre y ci-dessus.

Q.8 Autour du nombre 1

(a) i. Donner l'écriture scientifique de 1.

ii. Exprimer $f_{4,8}(1)$.

iii. Décrire $f_{E,M}(1)$ puis $f_{E,M}(2)$ pour un format quelconque.

(b) Pour un format $(1, E, M)$ de nombres à virgule flottante, donner, en fonction de E, M, l'expression du successeur de 1 (le nombre noté s(1)).

Application numérique : donner ce nombre en base dix pour le format du TP.

(c) On appelle *epsilon machine* et on note $\varepsilon_{E,M}$ (ou ε s'il n'y pas de risque de confusion), la distance entre 1 et son successeur pour le format $(1, E, M)$.

i. Calculer $\varepsilon_{4,8}$ et $\dfrac{\varepsilon_{4,8}}{2}$ en base 10.

ii. Existe-t-il un format E', M' où $\varepsilon_{E,M}$ est représentable en machine comme nombre normalisé ? Pourquoi ?

iii. Est-ce que $\varepsilon_{4,8}$ est représentable en machine :

A. comme nombre normalisé au format du TP ? Justifier.

B. Comme nombre dénormalisé au format du TP ? Justifier.

Si $\varepsilon_{4,8}$ est représentable en machine, donner $f_{4,8}(\varepsilon_{4,8})$.

(d) Pour une machine dont les flottants suivent le format $(1, E, M)$, comparer :

i. $f(1) \oplus f(\varepsilon)$ avec $f(1)$

ii. $f(1) \oplus f(\dfrac{\varepsilon}{2})$ avec $f(1)$

iii. $\left(f(1) \oplus f(\dfrac{\varepsilon}{2})\right) \oplus f(\dfrac{\varepsilon}{2})$ avec $f(1) \oplus \left(f(\dfrac{\varepsilon}{2}) \oplus f(\dfrac{\varepsilon}{2})\right)$

Q.9 Écart avec le sucesseur.

Pour le format $(1, E, M)$ calculer $s(2) - 2$ en fonction de $\varepsilon_{E,M}$.

Pour $n \in [\![0, E-2]\!]$, exprimer $s(2^n) - 2^n$ en fonction de $\varepsilon_{E,M}$. Que constate-ton ?

Q.10 A propos des erreurs. Soit un format $(1, E, M)$.

(a) Erreur absolue : Soit $x = (-1)^s(1+m)2^e$ (représentable ou non) qui n'est pas le maximum du format. Donner une majoration de $|x - \bar{x}|$ par une expression en fonction de $\varepsilon_{E,M}$ et e.

(b) Erreur relative : Soit $x = (-1)^s(1+m)2^e$ (représentable ou non) qui n'est pas le maximum du format. Majorer $\dfrac{|x-\bar{x}|}{|x|}$ par une constante.

Q.11 Autour du plus grand nombre positif.

(a) Exprimer en fonction de E, M et du décalage $d_{E,M}$ la valeur exacte de $g_{E,M}$, le plus grand nombre positif représentable en machine au format $(1, E, M)$. L'expression doit être la plus simplifiée possible.

(b) Calculer $g_{4,8}$.

(c) Quel est le plus petit nombre positif qui, ajouté à $g_{4,8}$, donne l'infini au format du TP ?

Q.12 Autour du plus petit nombre positif.

(a) Exprimer en fonction de E, M et du décalage $d_{E,M}$ la valeur exacte de $\ell_{E,M}$, le plus petit nombre positif normalisé représentable en machine au format $(1, E, M)$. L'expression doit être la plus simplifiée possible.

(b) Exprimer en fonction de E, M et du décalage $d_{E,M}$ la valeur exacte de $\ell'_{E,M}$, le plus petit nombre positif dé-normalisé représentable en machine au format $(1, E, M)$. L'expression doit être la plus simplifiée possible.

On rappelle que l'exposant des nombres dénormalisés est égal au plus petit exposant de nombre normalisé.

(c) Calculer le quotient $\dfrac{\ell_{E,M}}{\ell'_{E,M}}$

(d) Calculer $\ell'_{4,8}$ et $\ell_{4,8}$

Q.13 Dans un arrondi d'un nombre à la i-ème décimale, on garde i décimales. Dans ce TP, pour arrondir, on regarde les 3 bits qui suivent le dernier bit maintenu.

Soit x un nombre réel d'exposant 0. Donner une expression de $f_{4,8}(x)$ pour que l'arrondi à la seconde décimale de l'arrondi à la cinquième décimale de $f_{4,8}(x)$ soit différent de l'arrondi direct à la seconde décimale de $f_{4,8}(x)$ (plusieurs réponses sont possibles). Expliquer.

Q.14 **Application** Un développeur utilise un type `triple` dans son code. Il s'agit d'un format de flottants respectant la norme IEEE754 mais mal documenté.

(a) Écrire en `Ocaml` un programme permettant de calculer le nombre M de bits de mantisse pour ce format.

Expliquer en quelques mots pourquoi votre méthode est correcte.

(b) Le développeur connaît maintenant M, le nombre de bits de la mantisse.

Écrire en `Ocaml`, un programme permettant de calculer E, le nombre de bits d'exposant pour ce format.

Expliquer en quelques mots pourquoi votre méthode est correcte.