TD #8 — Nombres flottants

Exercice 1. *Racine carrée entière*

On suppose vouloir calculer des *racines carrées entières* $\lfloor \sqrt{x} \rfloor$ d'entiers $x \in [\![0, 2^{64} - 1]\!]$.

1. Expliquez pourquoi il n'est pas envisageable d'utiliser des nombres flottants « à double précision » (avec mantisse (effective) de 53 bits) pour effectuer ce calcul (par exemple comme la composition des fonctions `int_of_float`, `sqrt` et `float_of_int` en OCaml).
   Demandez-moi pour un indice.

Exercice 2. *Tri par base pour nombres flottants*

On rappelle quelques aspects de la représentation des nombres flottants sur 64 bits (« double précision ») IEEE754, qui sont les seuls « nombres flottants » que l'on considérera dans tous l'exercice. On utilisera pour cela (entre autres) des *masques* binaires (notés en hexadécimal) pour désigner des ensembles de bits.

— Le bit de signe (indiquant si le nombre est positif (bit à zéro) ou négatif (bit à un)) est le bit « le plus significatif », c'est à dire celui « le plus à gauche », ou d'indice 63. Il correspond au masque : `0x8000000000000000`.
— Les 11 bits d'exposant sont les bits les plus significatifs suivant (d'indices 62 à 52), qui correspondent au masque : `0x7FF0000000000000`. La valeur e de l'exposant quand non nul est donnée par l'entier que ces bits représentent en base deux (bit le plus significatif « à gauche ») auquel on soustrait 1023 (c'est une représentation *biaisée* de nombres relatifs).
— Les 52 bits restant représentent la mantisse, dont la valeur est donnée par 2^{-52} fois l'entier qu'ils représentent en base deux (bit le plus significatif « à gauche ») plus 2^{52} (le bit « gratuit »), hors cas particuliers que nous ignorerons. Le masque correspondant est : `0x000FFFFFFFFFFFFF`.

De plus on a que :

— Un nombre dont tous les bits de l'exposant sont à un représente $\pm\infty$ (en fonction du bit de signe) si tous les bits de la mantisse sont à zéro, et un `NaN` sinon.
— Un nombre dont tous les bits sont à zéro sauf éventuellement le bit de signe représente ± 0 en fonction de la valeur de ce dernier.
— Les représentations sont normalisées : l'unique représentation d'un nombre non nul (fini) représentable est celle d'exposant minimal, et sa valeur est $\pm 2^e \times m$.

Exemples

— Une mantisse `0x0000000000004` représente $2^{-52} \times (2^{52} + 2^2)$, soit $1 + 2^{-50}$.
— Un exposant `0x431` représente l'exposant $1073 - 1023 = 50$.
— Un nombre flottant de représentation binaire `0x4310000000000004` représente le nombre $2^{50} \times (1 + 2^{-50}) = 2^{50} + 1$.

Dans tout ce qui suit, soit x un nombre réel représentable exactement par un nombre flottant, on note $\rho(x)$ l'entier $\in [\![0, 2^{64} - 1]\!]$ obtenu en interprétant sa représentation binaire (comme ci-dessus) comme l'écriture en base deux d'un entier naturel. Dans le cas particulier de $x = 0$, on note $\rho^+(x)$ et $\rho^-(x)$ ses représentations positives et négatives.

1. Montrez que l'on peut avoir $x \in \mathbb{Z}$ et $x \neq \rho(x)$.
2. Soit x, y deux nombres réels représentables exactement par des flottants non négatifs (où l'on considère `-0.0` comme étant négatif), montrez que $\rho(x)$ et $\rho(y)$ se comparent identiquement à x et y (autrement dit, que $x = y \Rightarrow \rho(x) = \rho(y)$, $x < y \Rightarrow \rho(x) < \rho(y)$, $x > y \Rightarrow \rho(x) > \rho(y)$).
3. Cela reste-t'il vrai si l'on considère également des nombres négatifs ? Si non, que faut-il prendre en compte ?
4. Déduire de ce qui précède que la comparaison de deux nombres flottants (représentant des nombres finis) peut se faire essentiellement « sans calcul »
5. En déduire également qu'il est possible de trier un tableau de tels nombres en temps linéaire en la longueur du tableau (on pourra se contenter de détailler le cas de nombres non négatifs).
6. Cela reste-t'il possible si certains nombres flottants représentent des nombres infinis ?
7. Cela reste-t'il possible si certains nombres flottants ne représentent pas des nombres ?

Exercice 3. *Un exercice de Jean-Baptiste Bianquis*

À votre grand bonheur, vous avez reçu pour Noël une balance d'excellente qualité : elle offre trois chiffres (décimaux) de précision, et ce autant pour des masses de l'ordre du gramme que de l'ordre de la tonne. Vous décidez d'utiliser cette balance pour mesurer la masse m_h de votre hamster h. On suppose pour simplifier que m_h est de l'ordre de 100 grammes (un gros hamster, d'après Wikipédia).

1. Si h accepte de monter docilement sur la balance et d'y rester le temps qu'elle fasse sa mesure, avec quelle précision obtiendrez-vous m_h ?
2. h (qui est d'une intelligence assez rare pour un rongeur) vous soupçonne, à tort ou à raison, de vouloir utiliser cette pesée pour justifier une mise au régime. Il descend donc immédiatement de la balance à chaque fois que vous l'y posez, et ce avant que la mesure n'ait été faite. Vous décidez alors de le peser indirectement : vous vous pesez une première fois avec h dans la main, puis une deuxième fois sans h, et vous faites la différence. Avec quelle précision obtenez-vous m_h ?