Annexes
TD8 - Codage des flottants

1 Quelques repères importants

1.1 Codage des flottants en Python

En Python, les nombres réels sont codés selon la norme IEEE 754 double précision, c'est à dire sur 64 bits répartis comme suit :

— 1 bit pour le signe, 0 pour les réels positifs, 1 pour les réels négatifs,
— 11 bits pour l'exposant e (avec décalage de 1023), donc $e \in [-1023, 1024]$,
— 52 bits pour la mantisse m, $1 \leq m < 2$.

En particulier en double précision on a :

Cas	**Représentation**
plus grand nombre positif	$1.7976931348623157 \times 10^{308}$
plus petit nombre positif	$2.2250738585072014 \times 10^{-308}$
0+	s = 0 ; e = - 1023 ; m = 0
0-	s = 1 ; e = - 1023 ; m = 0
$+\infty$	s = 0 ; e = 1024 ; m = 0
$-\infty$	s = 1 ; e = 1024 ; m = 0
NaN (Not a Number)	e = 1024 ; m$\neq$ 0

1.2 Précision machine

La précision correspond au nombre de chiffres significatifs qui peuvent être stockés. Pour les flottants en double précision, elle correspond aux nombres de bits utilisés pour coder la mantisse, plus le bit implicite.

Comme $2^{-52} \simeq 2.2 \times 10^{-16}$, on peut stocker environ 15 à 16 décimales pour la mantisse.

1.3 Répartition des flottants parmi les réels

Si un flottant N s'écrit :

$$N = m \times 2^e,$$

alors les mantisses de ses 2 plus proches voisins N^- et N^+ s'obtiennent en décrémentant ou incrémentant la mantisse de N de 2^{-52} :

$$N^- = (m - 2^{-52}) \times 2^e$$
$$N^+ = (m + 2^{-52}) \times 2^e$$

Ainsi on a :

$$N^- = (m - 2^{-52}) \times 2^e = m \times 2^e - 2^{-52} \times 2^e = N - 2^{e-52}$$
$$N^+ = (m + 2^{-52}) \times 2^e = m \times 2^e + 2^{-52} \times 2^e = N + 2^{e-52}$$

Donc deux nombres flottants consécutifs sont définis par un incrément ou décrément constant de leur mantisse, mais cela correspond à un écart entre ces deux flottants qui dépend de leur exposant e. L'écart entre deux flottants consécutifs est donc de :

$$2^{e-52}$$

Il en résulte que deux flottants consécutifs avec un grand exposant sont plus éloignés sur l'axe des réels que deux flottants consécutifs avec un petit exposant. Autrement dit, l'échantillonnage de l'axe des réels n'est pas uniforme, la densité des flottants diminue au fur et à mesure que l'on s'éloigne de zéro pour aller vers les infinis (cf Figure 1).

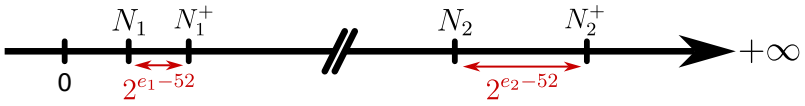

FIGURE 1 – La densité des flottants varie en fonction de la valeur du flottant. N_1 et N_1^+ (respectivement N_2 et N_2^+) sont des flottants consécutifs d'exposant e_1 (respectivement e_2). Comme $e_1 < e_2$, $N_1 - N_1^+ < N_2 - N_2^+$

1.4 Opérations d'arrondi

Comme les flottants sont codés sur un nombre fini de bits, il y en a un nombre fini. Pour coder les réels, qui sont en nombre infini, on fait donc des arrondis.

Lorsque un réel ne coïncide pas directement avec un flottant, mais qu'il n'est pas hors limite, il est arrondi au flottant le plus proche, *i.e.* le flottant le moins éloigné de lui sur la droite des réels. En cas d'égalité de distance avec 2 flottants, on prend le flottant avec le dernier bit de la mantisse nul, d'où l'expression "arrondi au chiffre pair le plus proche".

Ainsi, en pratique, tout flottant représente tout un intervalle de réels.

Si le nombre réel est supérieur en valeur absolue au plus grand nombre flottant (normalisé) positif, alors il sera arrondi vers un infini, on parle alors de dépassement (overflow). Par exemple, le nombre -1.8×10^{308} sera arrondi vers $-\infty$.

Si le nombre réel est inférieur en valeur absolu au plus petit nombre flottant (normalisé) positif, alors il sera arrondi vers 0^+ ou 0^-. Dans ce cas, on parle de soupassement (underflow). Par exemple, le nombre 10^{-324} est arrondi vers 0^+.

1.5 Erreur d'arrondi maximale pour un flottant donné

Soit un flottant N, son successeur N^+ et son prédécesseur N^-.

Les réels compris entre $\frac{N^-+N}{2}$ et $\frac{N+N^+}{2}$ seront arrondis vers N (voir Figure 2). Donc si un réel, R est arrondi vers N, l'erreur d'arrondi en valeur absolue est au plus de 2^{e-53} (voir Eqs. (1-2)).

$$|N - (\frac{N+N^-}{2})| = |\frac{N-N^-}{2}| = |\frac{N-(N-2^{e-52})}{2}| = |2^{e-53}| \quad (1)$$

$$|(\frac{N+N^+}{2}) - N| = |\frac{N^+-N}{2}| = |\frac{(N+2^{e-52})-N}{2}| = |2^{e-53}| \quad (2)$$

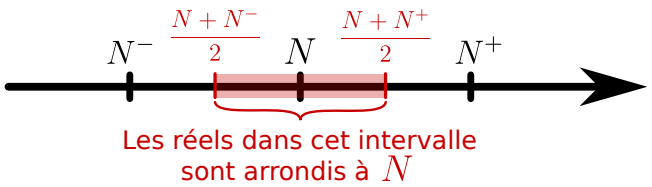

FIGURE 2 – Les réels ne peuvent pas être tous représentés par des flottants. N^-, N et N^+ sont des flottants consécutifs. Tous les réels entre ces valeurs seront arrondis à un de ces flottants.

2 Arithmétique flottante

2.1 Déroulement d'une opération élémentaire

Pour une opération élémentaire (+, -, ×, /, ...), la procédure de calcul est la suivante :

— Convertir les opérandes en flottants (2 arrondis)
— Effectuer les opérations sur les opérandes arrondis
— Convertir le résultat en flottant (1 arrondi)
— Stocker le résultat

Donc pour une simple opération, on aura 3 opérations d'arrondis.

2.2 Combinaison de plusieurs calculs

Les règles générales qui s'appliquent sont les mêmes qu'en arithmétique classique :
— On traite d'abord les calculs entre parenthèses
— Puis les multiplications et les divisions
— Et enfin les additions et les soustractions
En cas de même ordre de priorité, les calculs sont effectués de gauche à droite.

2.3 Propriétés de l'arithmétique flottante

Soit $\mathcal{F}$ l'ensemble des flottants.
— **Commutativité**
$\forall N_1, N_2 \in \mathcal{F}$ on a : $N_1 + N_2 = N_2 + N_1$

— **Non associativité**
$\exists N_1, N_2, N_3 \in \mathcal{F}$ tels que : $N_1 + (N_2 + N_3) \neq (N_1 + N_2) + N_3$
À cause des erreurs d'arrondis qui ne seront pas forcément les mêmes.

— **Non distributivité de la multiplication par rapport à l'addition**
$\exists N_1, N_2, N_3 \in \mathcal{F}$ tels que : $N_1 \times (N_2 + N_3) \neq N_1 \times N_2 + N_1 \times N_3$
Encore à cause des erreurs d'arrondis.

2.4 Quelques phénomènes innatendus

Les phénomènes suivants sont à connaître car il peuvent être source de nombreuses erreurs en calcul numérique.

2.4.1 Absorption

Le phénomène d'absorption se produit quand on additionne ou on soustrait un nombre relativement petit à un nombre relativement grand.

Dans ce cas, l'un des opérandes va absorber l'autre, c'est à dire que tout se passe comme si l'opération n'avait pas eu lieu (voir Fig 3).

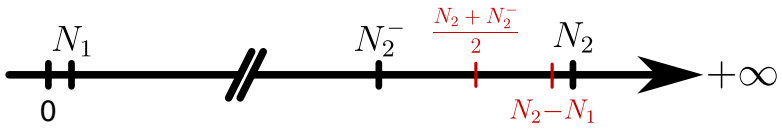

FIGURE 3 – Le réel $N_2 - N_1$ va être arrondis au flottant N_2. Tout se passe comme si la soustraction n'avait pas eu lieu.

2.4.2 Annulation catastrophique

L'annulation catastrophique est le phénomène selon lequel soustraire deux très bonnes approximations de deux nombres proches peut donner une très mauvaise approximation de la valeur de leur différence.

Exemple :
Soit deux personnes mesurant respectivement $t_1 = 175.6$cm et $t_2 = 174.4$cm.
On les mesure avec un mètre précis au centimètre près et on obtient les mresures suivantes : $\hat{t}_1 = 176$cm et $\hat{t}_2 = 174$cm.
L'erreur relative d'approximation de leur taille est donc :

$$t_1^{\text{err}} = \frac{|t_1 - \hat{t}_1|}{t_1} \simeq 0.2\% \qquad t_2^{\text{err}} = \frac{|t_2 - \hat{t}_2|}{t_2} \simeq 0.2\%$$

Si on s'intéresse maintenant à la différence de taille de ces personnes. La différence réelle de leur taille est $d = t_1 - t_2 = 1.2$cm.
Mais la différence approximée est $\hat{d} = \hat{t}_1 - \hat{t}_2 = 2$cm. On a donc une erreur sur la différence de :

$$d^{\text{err}} = \frac{|d - \hat{d}|}{d} \simeq 66\%$$

On a donc perdu de la précision lors de la soustraction de ces deux nombres.

L'annulation catastrophique n'est pas spécifique de l'arithmétique flottante, mais inhérente à l'opération de soustraction.

En arithmétique flottante, la différence entre deux nombres flottants très proche est calculée de manière exacte (pour aller plus loin voir le lemme de Sterbenz). Néanmoins, comme nous l'avons vu précedemment, la majorité des réels ne peuvent pas être stockés de manière exacte en mémoire, et sont donc arrondis au flottant le plus proche. Ainsi la soustractions de réels proches est en fait une soustraction de leur version approximée et donc est soumise au phénomène d'annulation catastrophique.