Analyse Numérique
Corrigés des Travaux Pratiques 2024 – 2025
Séance 2

Attention : dans toutes les manipulations les termes $\mathcal{O}(u^2)$ (ou le produit de $\mathcal{O}(u^2)$ avec d'autres termes) ont été négligés. Par ailleurs, les références au Chapitre 1 correspondent aux transparents de ce chapitre.

1. Soit

$$x \;=\; \pm\, \overline{0.d_1 d_2 \cdots d_t d_{t+1} \cdots} \cdot \beta^e$$

avec $d_1 \neq 0$ et $e \in [e_{\min}, e_{\max}]$. La relation (2) du cours montre que

$$|\mathrm{fl}(x) - x| \;\leq\; u\beta^{e-1}$$

et, comme $d_1 \neq 0$,

$$|\mathrm{fl}(x)| \;\geq\; \overline{0.d_1 d_2 \cdots\cdots d_t 0 \cdots}\beta^e \;\geq\; \overline{0.1}\,\beta^e \;=\; \beta^{e-1}\,.$$

Combinant les deux inégalités on a

$$\frac{|\mathrm{fl}(x) - x|}{|\mathrm{fl}(x)|} \;\leq\; u\,.$$

Utilisant le raisonnement pour passer de la formule (3) à la formule (4) du Chapitre 1 (mais permutant les rôles de x et $\mathrm{fl}(x)$) on a

$$x \;=\; \mathrm{fl}(x)(1+\epsilon)\,, \quad |\epsilon| \;\leq\; u\,,$$

et le résultat voulu en découle.

2. Pour cet exercice, e n'est que la représentation machine de la célèbre constante. Analyse (tous les ϵ_i satisfont $|\epsilon_i| \leq u$) :

$$\begin{aligned}
(e \odot (1 \oslash 7)) \odot 7 \;&=\; (e \odot (1/7)(1+\epsilon_1)) \odot 7 \\
&=\; e \cdot (1/7)(1+\epsilon_1)(1+\epsilon_2) \odot 7 \\
&=\; (e/7)(1+2\epsilon_3) \odot 7 \\
&=\; (e/7)(1+2\epsilon_3) \cdot 7(1+\epsilon_4) \\
&=\; e(1+3\epsilon_5)
\end{aligned}$$

et donc l'erreur relative par rapport à e vaut au plus $3u \approx 3.3\ 10^{-16}$. Avec `res=(e*(1/7))*7` l'instruction `abs(res-e)/e` donne `1.6337e-16`.

Il reste à noter que comme `res`, `e`$\in \mathbb{F}$, la valeur de `abs(res-e)/e` est déterminée avec ... une erreur **relative** d'au plus $2u \;=\; 2.2 \cdot 10^{-16}$ (pourquoi ?).

3. Avec $|\epsilon_i| \leq u$ on a pour la multiplication

$$\begin{aligned}
\widetilde{x} \odot \widetilde{y} \;&=\; (1+\epsilon_1)x \odot (1+\epsilon_2)y \\
&=\; ((1+\epsilon_1)x \cdot (1+\epsilon_2)y)(1+\epsilon_3) \\
&=\; xy(1+3\epsilon_4)\,;
\end{aligned}$$

l'erreur relative sur le résultat est donc au plus $3u$. Le résultat pour la division est similaire. Quant à l'addition, on a

$$\begin{aligned}\widetilde{x} \oplus \widetilde{y} &= (1+\epsilon_1)x \oplus (1+\epsilon_2)y \\ &= ((1+\epsilon_1)x + (1+\epsilon_2)y)(1+\epsilon_3) \\ &= (x+y)(1+\frac{\epsilon_1 x + \epsilon_2 y}{x+y})(1+\epsilon_3)\,.\end{aligned}$$

Comme x, y sont positifs (pourquoi ?)

$$\frac{|\epsilon_1 x + \epsilon_2 y|}{|x+y|} \leq \frac{|x|+|y|}{|x+y|}u = u\,, \tag{1}$$

et donc

$$\begin{aligned}\widetilde{x} \oplus \widetilde{y} &= (x+y)(1+\epsilon_4)(1+\epsilon_3) \\ &= (x+y)(1+2\epsilon_5)\,.\end{aligned}$$

L'erreur relative sur le résultat est donc au plus $2u$.

Notez que l'erreur relative maximale sur les trois opérations est semblable au cas où $\widetilde{x}$, $\widetilde{y}$ contiennent déjà des erreurs d'arrondi et au cas où ils n'en contiennent pas (Chapitre 1, p.9). En particulier, on peut effectuer plusieurs opérations d'addition de nombres positifs, de multiplication et de division sans craindre une soudaine perte de précision. Cela contraste avec l'opération de soustraction de deux nombres positifs, surtout si ces deux nombres sont proches (Exemple 2 du cours).

4.a Comme $f = x^2$ on a donc $f' = 2x$ et (en arithmétique exacte) $f'_h = \frac{(x+h)^2 - x^2}{h} = 2x + h$; par conséquent

$$f'_h(x) - f'(x) = h\,,$$

qui est satisfait pour tout x, et en particulier pour $x = 1$.

4.b Le risque d'annulation existe car $f(x) \approx f(x+h)$ pour h «petit», et donc on soustrait deux réels proches.

4.c La fonction qui évalue la valeur absolue de l'erreur numérique

```
function err = aprxerr(h)
fh = ((1+h)^2-1)/h; % approximation de f'(1)
fp = 2;             % valeur de f'(1)
err = abs(fh-fp);
```

et le résultat d'évaluation est comme suit

```
aprxerr(1e-6)  % donne 9.9992e-07
aprxerr(1e-8)  % donne 1.2155e-08
aprxerr(1e-10) % donne 1.6548e-07
aprxerr(1e-12) % donne 1.7780e-04
```

La dépendance de l'erreur en fonction de h est aussi représentée sur la Figure 1 pour une plage plus large de valeurs de h. On constate que l'erreur cesse de diminuer pour $h \to 0$ aux alentours de $h = 10^{-8}$, une situation qui ne peut être justifiée que par des erreurs d'arrondi.

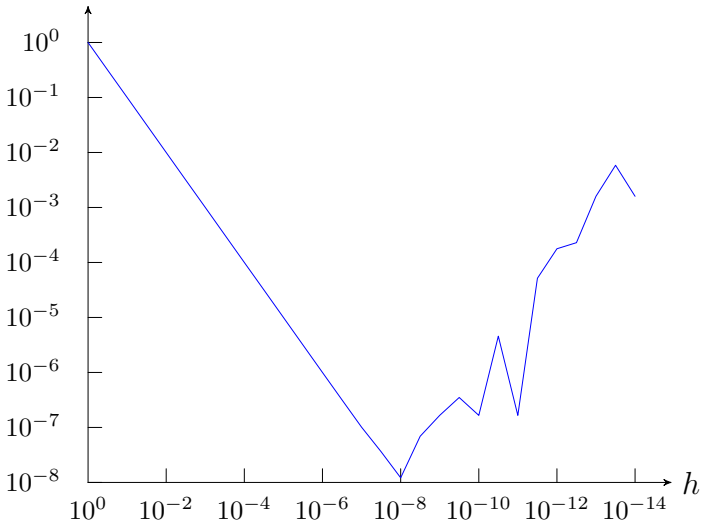

FIGURE 1 – L'erreur $|f'_h(x) - f'(x)|$ en arithmétique en virgule flottante pour différentes valeurs de h.

4.d En arithmétique en virgule flottante $f'(1) = 2$, alors que (en utilisant $|\epsilon_i| \leq u$) on a aussi

$$\begin{aligned}
f'_h(1) &= ((1 \oplus h) \odot (1 \oplus h) \ominus 1) \oslash h \\
&= \left((1+h)^2 \underbrace{(1+\epsilon_1)^2(1+\epsilon_2)}_{1+3\epsilon_3} \ominus 1 \right) \oslash h \\
&= \underbrace{(1+\epsilon_4)(1+\epsilon_5)}_{1+2\epsilon_6} \left((1+h)^2(1+3\epsilon_3) - 1 \right) /h \\
&= (1+2\epsilon_6) \left(2h + h^2 + 3\epsilon_3 + \mathcal{O}(hu) \right) /h \\
&= (1+2\epsilon_6) \left(2 + h + 3\frac{\epsilon_3}{h} + \mathcal{O}(u) \right) \\
&= 2 + h + 3\frac{\epsilon_3}{h} + \mathcal{O}(u) + \mathcal{O}(u^2/h)
\end{aligned}$$

et donc

$$f'_h(1) - \widetilde{f}(1) \approx h + 3\frac{\epsilon_3}{h}.$$

Notez que d'après ce résultat l'erreur minimale est commise quand les deux termes sont égaux ; en précision double ($u = 1.1\,10^{-16}$) cela correspond à $h \approx 10^{-8}$.

Annexe

A titre d'information, voici le code qui a permis de générer la Figure 1 ; il ne fait formellement pas partie de la matière.

```
% exposants unif. espacés
e = 0:0.5:14;
h = 10.^(-e);
for i=1:length(h)
    err(i) = aprxerr(h(i)); % erreur d'approximation
end
loglog(1./h,err);
box off
```