TELECOM Nancy *(ESIAL)*
Maths Numériques

devoir sur l'arithmétique flottante

Question de cours

Qu'est-ce qu'un "underflow" ?

Exercice

Dans cet exercice, on suppose que les calculs effectués par la machine ne provoquent ni overflow ni underflow.

On cherche à calculer précisément la fonction $f(x) = \sqrt{1+x} - \sqrt{1-x}$ lorsque $|x|$ est petit ($|x| < 1$).

1. On considère d'abord "l'algorithme" immédiat :

$$y1 := sqrt(1 \oplus x) \ominus sqrt(1 \ominus x)$$

où $sqrt(u)$ désigne la racine calculée en machine (pour cette opération la norme IEEE impose aussi que $sqrt(u) = fl(\sqrt{u})$, on a donc $sqrt(u) = \sqrt{u}(1+\epsilon)$ avec $|\epsilon| \leq \mathbf{u}$ pour tout flottant $u \geq 0$ qui n'est pas un nombre spécial (car avec la racine carrée $[\sqrt{\mu}, \sqrt{M}] \subset [m, M]$)).

   (a) Sans faire d'analyse d'erreur expliquer pourquoi cet algorithme risque d'être peu précis lorsque $|x| << 1$.

   (b) Analyser l'erreur obtenue par cet algorithme et trouver une "quasi" borne supérieure pour l'erreur relative (qui sera fonction de x). Aide : (i) On pourra considérer que la dernière soustraction est exacte (soustraction de deux nombres proches), (ii) vous allez rencontrer des termes de la forme $\sqrt{1+\epsilon_1}(1+\epsilon_2)$ avec $|\epsilon_i| < \mathbf{u}$ que vous pourrez écrire sous la forme $1+\delta_1$. Comme $\sqrt{1+x} = 1 + x/2 + O(x^2)$ on peut voir que δ_1 est "quasi-borné" par $\frac{3}{2}\mathbf{u}$. (iii) On pourra simplifier les calculs en utilisant $\sqrt{1+x} \simeq 1 + x/2$ puisqu'on s'intéresse à la précision lorsque $|x|$ est petit. Rmq : la quasi-borne de l'erreur relative est $3\mathbf{u}/|x|$.

2. Écrire f différemment pour trouver le bon algorithme (lorsque $|x|$ est petit). Aide : si a et b sont deux réels positifs $a - b = (\sqrt{a} - \sqrt{b})(\sqrt{a} + \sqrt{b})$. Une analyse d'erreur montre que cet algorithme est stable (erreur relative quasi-bornée par $\frac{7}{2}\mathbf{u}$), ne pas la faire.

3. Illustrer cet exercice par un code python comme pour l'exercice 3 du TP, c'est à dire en considérant la deuxième formule comme quasi-exacte. Expliquer et commenter vos résultats !