Présentation

AFAE

Tayyib Patel

January 4, 2016

Stupid is as Stupid Does: Taking the Square Root of the Square of a Floating-Point Number

Sylvie Boldo

Plan

Comment est représenté un nombre x en virgule flottante ?

- Base considérée β
- Signe s_x
- Mantisse m_x
- Exposant e_x entre e_{min} et e_{max}

$$x = (\text{-}1)^{s_x} * m_x * 2^{e_x}$$
$$\text{avec } m_x = x_0.x_1x_2..x_n,\ x_i \text{ app. } \{0,..,\beta - 1\}$$

Avant 1985, représentations différentes selon le langage, l'architecture...

1985: Norme IEEE754

$\beta = 2$

Deux types de représentation:

- format simple précision : 32 bits, 1 bit pour s_x, 24 bits de mantisse (1 implicite + 23 de fraction), 8 bits pour e_x
- format double précision : 64 bits, 1 bit pour s_x, 53 bits de mantisse (1 implicite + 52 de fraction), 11 bits pour e_x

Comment faire lorsque x non représentable exactement ?
$\rightarrow$ Utilisation d'arrondis

Types d'arrondis proposés par la norme IEEE754:

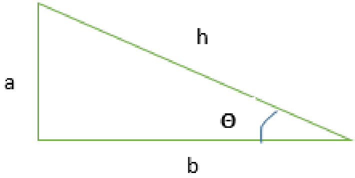
- arrondi vers $-\infty$ noté $\Delta(x)$: plus grand nombre machine $\leq x$
- arrondi vers $+\infty$ noté $\nabla(x)$: plus petit nombre machine $\geq x$
- arrondi autour de 0 noté Z(x) : $\Delta(x)$ pour $x < 0$ et $\nabla(x)$ pour $x > 0$
- arrondi au plus près noté $\circ(x)$: nombre machine le plus proche de x
  Si x au milieu de deux nombres machines
  Arrondi pair: Vers le nombre dont la mantisse se termine par 0

Autres règles d'arbitrage pour l'arrondi au plus près

- Arrondi "autour de 0": Z(x)
- Arrondi "vers $-\infty$": $\Delta(x)$
- Arrondi "vers $+\infty$": $\nabla(x)$

Représentation des nombres sur ordinateur: Danger !

- Ariane 5 -> conversion d’une valeur flottante 64 bits vers un entier signé 16 bits
- Missile Patriot -> erreur d’arrondi: 0.1 non représentable en virgule flottante (en binaire)



$\Theta = arcsin(a/\sqrt{a^2 + b^2})$

$\frac{a}{\sqrt{a^2+b^2}}$ non exactement représenté en machine -> arrondi $\circ$ utilisé dans tout l'article

$x = \circ(\frac{a}{\circ(\sqrt{\circ(\circ(a^2)+\circ(b^2))})}) \Rightarrow$ 5 arrondis nécessaires

Problème: Comment s'assurer que $-1 \leq x \leq 1$?

Analyse du pire cas ($|x| = 1$), atteint pour b=0

$$x = \circ\left(\frac{a}{\circ(\sqrt{\circ(a^2)})}\right)$$

Il suffit alors pour s'assurer que $-1 \leq x \leq 1$ de prouver que

$$\circ(\sqrt{\circ(a^2)}) = |a|$$

Hypothèses:

- Base considérée égale à 2
- Exposant non borné: $e_{min} = -\infty$ et $e_{max} = +\infty$
- Précision p > 1

Format utilisé correspondant à ces hypothèses: FLX

Outils:

- Coq Proof Assistant (vérification de la preuve)
- Librairie Flocq

Plusieurs lemmes nécessaires à la preuve de la propriété

Notation: ulp: valeur du dernier bit de la mantisse d’un nombre à virgule flottante ou d’un nombre réel

Lemme 1

Soit v un réel et u un nombre à virgule flottante positif.
Si $v < u + \frac{ulp(u)}{2}$, alors $\circ(v) \leq u$.

Preuve:

$\circ$ monotone : si $x < y$, alors $\circ_1(x) \leq \circ_2(y)$ quelques soient les 2 règles d'arbitrage.
En prenant la règle d'arbitrage paire pour v, et la règle d'arbitrage autour de 0 pour $u + \frac{ulp(u)}{2}$, comme $v < u + \frac{ulp(u)}{2}$, on a $\circ(v) \leq u$, puisque $u + \frac{ulp(u)}{2}$ est situé à mi-distance entre u et son successeur.

Lemme 2

Soit v un réel et u un nombre à virgule flottante positif.
Supposons que u n'est pas une puissance de 2, et que $u - \frac{ulp(u)}{2} < v$.
Alors $u \leq \circ(v)$.

Preuve:

u pas une puissance de la base, et l'exposant n'est pas borné $\Rightarrow$
Prédecesseur de u a même exposant et même ulp, et est positif.
Reste de la preuve similaire au précédent cas.

Lemme 3

Soit u un réel positif. Alors
$ulp(u^2) + \frac{(ulp(u))^2}{2} < 2 * u * ulp(u)$.

Preuve:

Soit e t.q $ulp(u) = 2^{e-p}$, et i t.q $ulp(u^2) = 2^{i-p}$. Alors soit
$i = 2 * e - 1$, soit $i = 2 * e$.
Dans chaque cas, l'inéquation est vérifiée.

Lemme 4

Soit u un nombre à virgule flottante positif. Alors $\circ(\sqrt{\circ(u^2)}) = u$.

Preuve:

- Si u est puissance de la base, u^2 exactement représenté (car exposant non borné) et pas d'erreur d'arrondi.
- Sinon, soit $y = \circ(u^2)$.
  On prouve d'abord $\circ(\sqrt{\circ(u^2)}) \geq u$.
  Pour cela, on prouve que $u - \frac{ulp(u)}{2} < \sqrt{y}$ afin de pouvoir utiliser le Lemme 2. La preuve de cette inégalité fait appel au Lemme 3.
  Pour prouver $\circ(\sqrt{\circ(u^2)}) \leq u$, on prouve que $\sqrt{y} < u + \frac{ulp(u)}{2}$ afin de pouvoir utiliser le Lemme 1. Pour cela, on utilise encore une fois le Lemme 3, ce qui nous amène au résultat.

Théorème 5

Soit u un nombre à virgule flottante. Alors $\circ(\sqrt{\circ(u^2)}) = |u|$.

Preuve:

Si u négatif, on utilise le théorème précédent sur -u,et si u est zéro, le résultat est vrai (puisque $\circ(0) = 0$).

Théorème 6

Soit a un nombre à virgule flottante et b un réel. Alors $-1 \leq \circ(\frac{a}{\circ(\sqrt{\circ(\circ(a^2)+\circ(b^2))})}) \leq 1$.

Preuve:

Le théorème 5 nous donne $|\circ(\frac{a}{\circ(\sqrt{\circ(a^2)})})| = |\circ(\frac{a}{|a|})| = 1$.

Comme $|\circ(\frac{a}{\circ(\sqrt{\circ(\circ(a^2)+\circ(b^2))})})| \leq |\circ(\frac{a}{\circ(\sqrt{\circ(a^2)})})|$, le résultat est prouvé.

Conclusion

- Sous toutes les hypothèses considérées, test sur x avec x la représentation en virgule flottante de $\frac{a}{\sqrt{a^2+b^2}}$ non nécessaire
- Résultat valide pour |a| entre 2^{-511} et 2^{511} en double précision.
- Résultat également valide pour règle d'arbitrage au plus près vers $+\infty$.
- Apparemment, si mantisse petite, résultat vrai dans n'importe quelle base

Limites

- Résultat non valide pour arrondis dirigés.
- Résultat non valide dans le cas général pour les autres bases

Remarque: Même dans le cas où le résultat n’est pas valide, dans tous les cas testés, $|x|$ n’est jamais supérieur à 1.

Conclusion générale

Ne pas répéter les erreurs du passé:

- Ne pas négliger les tests, surtout en représentation machine
- Ne pas non plus faire de tests inutiles (coûteux en temps/mémoire)

Merci de votre attention!