Calcul sur des nombres à virgule flottante symboliques

Antoine PLET

LIP (AriC) - ENS de Lyon

RAIM - 9 avril 2015

Travail réalisé avec C.-P. Jeannerod, N. Louvet et J.-M. Muller

Nombres à virgule flottante

Définition

En base β et précision p :

$$(-1)^s \cdot m \cdot \beta^e$$

avec $s \in \{0, 1\}$, $m \in \mathbb{N}$ (mantisse) tel que $m < \beta^p$ et $e \in \mathbb{Z}$ (exposant).
On note $\mathbb{F}_p$ leur ensemble.

Nombres à virgule flottante

Définition

En base β et précision p :

$$(-1)^s \cdot m \cdot \beta^e$$

avec $s \in \{0, 1\}$, $m \in \mathbb{N}$ (mantisse) tel que $m < \beta^p$ et $e \in \mathbb{Z}$ (exposant).
On note $\mathbb{F}_p$ leur ensemble.

Formats standards

La norme IEEE 754-2008 définit deux formats standards en base 2 :

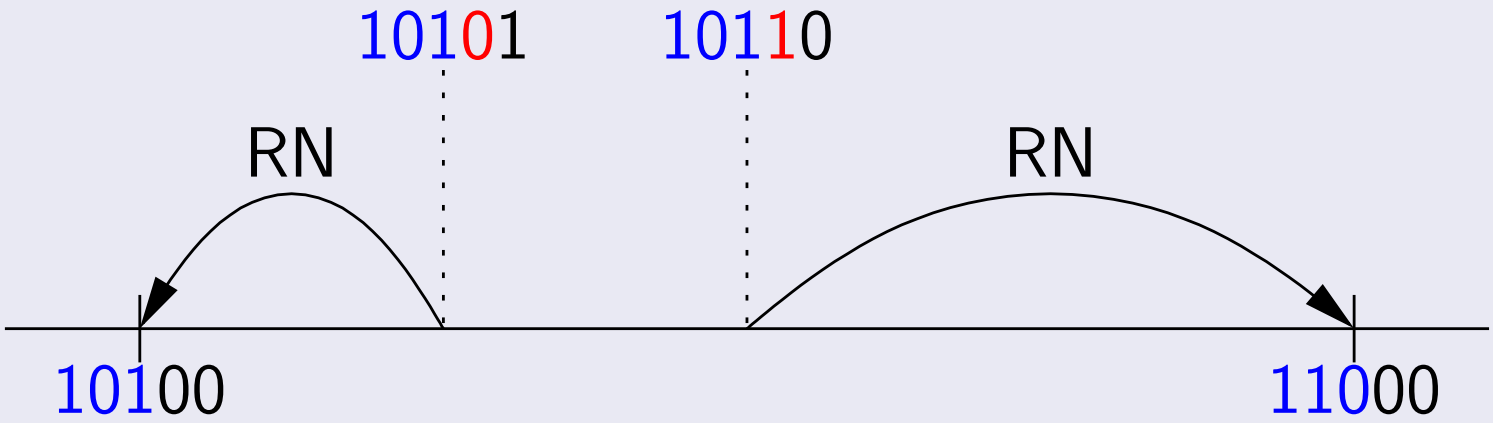
- binary32 (type *float* en C), avec $p = 24$,
- binary64 (type *double* en C), avec $p = 53$.

Approximation des réels pour les calculs, avec une règle (arrondi correct) définie par la norme pour les opérations de base.

Fonction d'arrondi

Arrondi au plus proche vers la mantisse paire (RN)

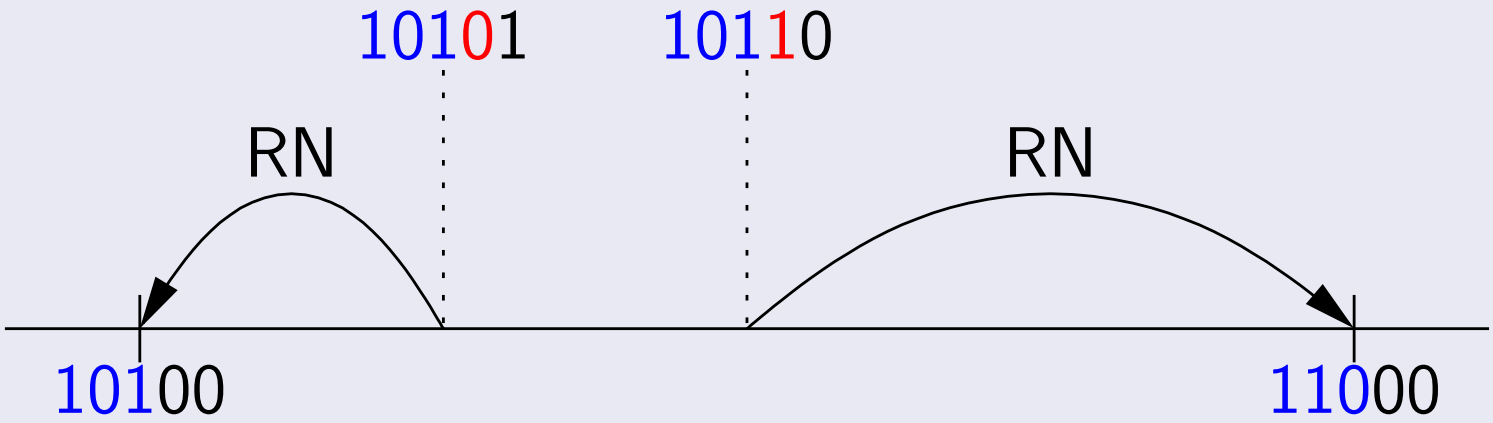
En base 2 et précision $p = 3$:



Fonction d'arrondi

Arrondi au plus proche vers la mantisse paire (RN)

En base 2 et précision $p = 3$:



- Opérations de base : $+$, $-$, $\times$, $\div$, $\sqrt{}$ et FMA (Fused Multiply-Add) qui calcule $a \times b + c$. Calcul de l'arrondi des résultats exacts.
- Erreur relative $\frac{|\text{valeur calculée} - \text{valeur exacte}|}{|\text{valeur exacte}|}$ bornée par $u = \frac{1}{2}\beta^{1-p}$.

Calcul de $r = ab + cd$

algorithm naïf(a, b, c, d)
  $\widehat{p}_1 := \mathrm{RN}(ab)$;
  $\widehat{p}_2 := \mathrm{RN}(cd)$;
  $\widehat{r} := \mathrm{RN}(\widehat{p}_1 + \widehat{p}_2)$;
  return $\widehat{r}$

Calcul de $r = ab + cd$

algorithm naïf(a, b, c, d)
  $\widehat{p}_1 := \mathrm{RN}(ab)$;
  $\widehat{p}_2 := \mathrm{RN}(cd)$;
  $\widehat{r} := \mathrm{RN}(\widehat{p}_1 + \widehat{p}_2)$;
  return $\widehat{r}$

Borne d'erreur [Higham, '02] :

$$\frac{|r - \widehat{r}|}{|r|} \leqslant (2u + u^2)\frac{|ab| + |cd|}{|ab + cd|}.$$

Calcul de $r = ab + cd$

algorithm naïf(a, b, c, d)
$\widehat{p}_1 := \mathrm{RN}(ab)$;
$\widehat{p}_2 := \mathrm{RN}(cd)$;
$\widehat{r} := \mathrm{RN}(\widehat{p}_1 + \widehat{p}_2)$;
return $\widehat{r}$

Borne d'erreur [Higham, '02] :

$$\frac{|r - \widehat{r}|}{|r|} \leqslant (2u + u^2)\frac{|ab| + |cd|}{|ab + cd|}.$$

Exemple de mauvais cas

En entrée :

- $a = 2^{p-1} + 2^{p-2} - 1$
- $b = -a$
- $c = a + 1$
- $d = a - 1$

En sortie :

- $r = -1$
- $\widehat{r} = -2^p$

$\Rightarrow$ Erreur relative :
$2^p - 1 = \frac{1}{u} - 1 \gg 1.$

Algorithme de Cornea, Harrison et Tang pour $r = ab + cd$

algorithm CHT(a, b, c, d)
$\widehat{p}_1 := \mathrm{RN}(ab)$;
$e_1 := \mathrm{RN}(ab - \widehat{p}_1)$;
$\widehat{p}_2 := \mathrm{RN}(cd)$;
$e_2 := \mathrm{RN}(cd - \widehat{p}_2)$;
$\widehat{r} := \mathrm{RN}(\widehat{p}_1 + \widehat{p}_2)$;
$\widehat{e} := \mathrm{RN}(e_1 + e_2)$;
$\widehat{s} := \mathrm{RN}(\widehat{r} + \widehat{e})$;
return $\widehat{s}$

Algorithme de Cornea, Harrison et Tang pour $r = ab + cd$

algorithm CHT(a, b, c, d)
$\widehat{p}_1 := \mathrm{RN}(ab)$;
$e_1 := \mathrm{RN}(ab - \widehat{p}_1)$;
$\widehat{p}_2 := \mathrm{RN}(cd)$;
$e_2 := \mathrm{RN}(cd - \widehat{p}_2)$;
$\widehat{r} := \mathrm{RN}(\widehat{p}_1 + \widehat{p}_2)$;
$\widehat{e} := \mathrm{RN}(e_1 + e_2)$;
$\widehat{s} := \mathrm{RN}(\widehat{r} + \widehat{e})$;
return $\widehat{s}$

- Erreur relative $\leqslant \mathcal{O}(u)$
  [Cornea, Harrison et Tang, '02],
- Erreur relative $\leqslant 2u$
  [Jeannerod, '14].

Algorithme de Cornea, Harrison et Tang pour $r = ab + cd$

algorithm CHT(a, b, c, d)
$\widehat{p}_1 := \mathrm{RN}(ab)$;
$e_1 := \mathrm{RN}(ab - \widehat{p}_1)$;
$\widehat{p}_2 := \mathrm{RN}(cd)$;
$e_2 := \mathrm{RN}(cd - \widehat{p}_2)$;
$\widehat{r} := \mathrm{RN}(\widehat{p}_1 + \widehat{p}_2)$;
$\widehat{e} := \mathrm{RN}(e_1 + e_2)$;
$\widehat{s} := \mathrm{RN}(\widehat{r} + \widehat{e})$;
return $\widehat{s}$

- Erreur relative $\leqslant \mathcal{O}(u)$
  [Cornea, Harrison et Tang, '02],
- Erreur relative $\leqslant 2u$
  [Jeannerod, '14].

Certificat d'optimalité en base 2
[Muller, '13] :

- $a = c = 2^p - 1$,
- $b = 2^{p-3} + 2^{-1}$,
- $d = 2^{p-3} + 2^{-2}$.

$$r = 2^{2p-2} + 2^{p-1} - 3 \cdot 2^{-2}$$
$$\widehat{s} = 2^{2p-2}$$

Algorithme de Cornea, Harrison et Tang pour $r = ab + cd$

algorithm CHT(a, b, c, d)
$\widehat{p}_1 := \mathrm{RN}(ab)$;
$e_1 := \mathrm{RN}(ab - \widehat{p}_1)$;
$\widehat{p}_2 := \mathrm{RN}(cd)$;
$e_2 := \mathrm{RN}(cd - \widehat{p}_2)$;
$\widehat{r} := \mathrm{RN}(\widehat{p}_1 + \widehat{p}_2)$;
$\widehat{e} := \mathrm{RN}(e_1 + e_2)$;
$\widehat{s} := \mathrm{RN}(\widehat{r} + \widehat{e})$;
return $\widehat{s}$

- Erreur relative $\leqslant \mathcal{O}(u)$ [Cornea, Harrison et Tang, '02],
- Erreur relative $\leqslant 2u$ [Jeannerod, '14].

Certificat d'optimalité en base 2 [Muller, '13] :

- $a = c = 2^p - 1$,
- $b = 2^{p-3} + 2^{-1}$,
- $d = 2^{p-3} + 2^{-2}$.

$$r = 2^{2p-2} + 2^{p-1} - 3 \cdot 2^{-2}$$
$$\widehat{s} = 2^{2p-2}$$

$$\Rightarrow \text{Erreur relative} \geqslant 2u - 7u^2$$

La borne d'erreur est asymptotiquement optimale.

Analyse d'algorithme en arithmétique flottante

Deux étapes :

- Calcul d'une borne sur l'erreur relative : analyse d'erreur a priori pour p « suffisamment grand ».
- Étude de la qualité de cette borne : recherche de mauvais cas à la main.

Inconvénients de la vérification de certificats à la main

- Fort risque d'erreur.
- Demande beaucoup de temps.
- Tout est à reprendre en cas de changement de base.

Exemple de l'évaluation de $ab + cd$ par l'algorithme naïf

Entrées :

$$a = 2^{p-1} + 2^{p-2} - 1, \quad b = -a, \quad c = a + 1, \quad d = a - 1.$$

- $a^2 = 2^{2p-2} + 2^{2p-4} + 1 + 2 \cdot 2^{2p-3} - 2 \cdot 2^{p-1} - 2 \cdot 2^{p-2}$
  $= 2^{2p-1} + 2^{2p-4} - 2^p - 2^{p-1} + 1$
- $ab = -a^2 = -\left(2^{2p-1} + 2^{2p-4} - 2^p - 2^{p-1} + 1\right)$
- $\widehat{p_1} = \mathrm{RN}(ab) = -\left(2^{2p-1} + 2^{2p-4} - 2^p\right)$
- $cd = a^2 - 1 = 2^{2p-1} + 2^{2p-4} - 2^p - 2^{p-1}$
- $\widehat{p_2} = \mathrm{RN}(cd) = 2^{2p-1} + 2^{2p-4} - 2^{p+1}$
- $\mathrm{RN}(ab) + \mathrm{RN}(cd) = -2^{p+1} + 2^p = -2^p$
- $\widehat{r} = \mathrm{RN}(\mathrm{RN}(ab) + \mathrm{RN}(cd)) = -2^p$

Objectif

Proposer des méthodes automatiques pour le calcul flottant symbolique pour

- accélérer la recherche de mauvais cas,
- vérifier les certificats d'optimalité existants.

- Première approche symbolique, inspirée d'une méthode de calcul numérique.

- Seconde approche numérique, fondée sur l'évaluation-interpolation.

Régime asymptotique

Les résultats symboliques sont valables pour p « suffisamment grand ».

Illustration

- $\alpha = 2^{2p-2} + 2^{p-1} - 3$.
- $\mathrm{RN}(\alpha) = 2^{2p-2} + 2^{p-1}$ pour $p \geqslant 4$.
- Si $p = 3$:
  - $\rightarrow$ $\alpha = 17 \Rightarrow \mathrm{RN}(\alpha) = 16$,
  - $\rightarrow$ mais $2^{2p-2} + 2^{p-1} = 20$.

$\Rightarrow$ Résultat symbolique valable pour $p \geqslant p_0$, avec $p_0 = 4$.

Conséquence

Il faut maîtriser le régime asymptotique et trouver une valeur correcte pour p_0.

Plan de l'exposé

Automatiser le calcul flottant symbolique

Hypothèses :

- On travaille en base β paire fixée ($\beta = 2$ ou $\beta = 10$).
- La précision p est symbolique.
- Les calculs sont faits avec l'arrondi au plus proche vers la mantisse paire.

Automatiser le calcul flottant symbolique

Hypothèses :

- On travaille en base β paire fixée ($\beta = 2$ ou $\beta = 10$).
- La précision p est symbolique.
- Les calculs sont faits avec l'arrondi au plus proche vers la mantisse paire.

Choix de conception :

- Choix d'une structure de données pour la représentation des nombres.
- Implantation des opérations élémentaires exactes ($+$ et $\times$).
- Implantation de la fonction d'arrondi.

Choix de représentation

Flottants « creux » $\Rightarrow$ représentation inspirée des polynômes creux :
un flottant est représenté par une liste de couples (coefficient, exposant), possiblement vide :

$$[(c_1, e_1), (c_2, e_2), \dots (c_n, e_n)] \mapsto \sum_{i=1}^{n} c_i \beta^{e_i}.$$

Contraintes sur la structure de données :

- coefficients c_i entiers dans $[1-\beta, \beta-1]$, $c_1 \neq 0$,
- exposants : fonctions affines en p,
- il existe p_0 tel que pour tout $p \geqslant p_0$ et $1 \leqslant i < n$,

$$e_i > e_{i+1},$$
$$\text{si } c_i c_{i+1} < 0, \text{ alors } e_i > e_{i+1} + 1.$$

$\hookrightarrow 2^p - 2^{p-1} - 2^{p-2} = 2^{p-2}$

Exemple de calcul numérique d'arrondi

$\beta = 2$, $p = 6$ et $\alpha = 812 = 1100101100_2$. Que vaut RN($\alpha$) ?

Exemple de calcul numérique d'arrondi

$\beta = 2$, $p = 6$ et $\alpha = 812 = 1100101100_2$. Que vaut RN($\alpha$) ?

- Bit de poids fort : 1100101100_2.

Exemple de calcul numérique d'arrondi

$\beta = 2$, $p = 6$ et $\alpha = 812 = 1100101100_2$. Que vaut RN($\alpha$) ?

- Bit de poids fort : 1100101100_2.
- Bit d'arrondi : 1100101100_2.

Exemple de calcul numérique d'arrondi

$\beta = 2$, $p = 6$ et $\alpha = 812 = 1100101100_2$. Que vaut RN($\alpha$) ?

- Bit de poids fort : 1100101100_2.
- Bit d'arrondi : 1100101100_2.
- Encadrement : $800 = 1100100000_2 \leqslant \alpha \leqslant 1100110000_2 = 816$.

Exemple de calcul numérique d'arrondi

$\beta = 2$, $p = 6$ et $\alpha = 812 = 1100101100_2$. Que vaut RN($\alpha$) ?

- Bit de poids fort : 1100101100_2.
- Bit d'arrondi : 1100101100_2.
- Encadrement : $800 = 1100100000_2 \leqslant \alpha \leqslant 1100110000_2 = 816$.
- RN(α) $= 816$.

Calcul symbolique d'arrondi

On cherche RN(α) où

$$\alpha = \underbrace{2^{2p-2}}_{\text{exposant maximal}} + \underbrace{2^{2p-3} + 2^{2p-6} + 2^{p+1} + 2^{p-1} + 2^2}_{\text{suite de l'expression}}$$

Calcul symbolique d'arrondi

On cherche RN(α) où

$$\alpha = \underbrace{2^{2p-2}}_{\text{exposant maximal}} + \underbrace{2^{2p-3} + 2^{2p-6} + 2^{p+1} + 2^{p-1} + 2^2}_{\text{suite de l'expression}}$$

- $2^{2p-2} \leqslant \alpha < 2^{2p-1}$

Calcul symbolique d'arrondi

On cherche RN(α) où

$$\alpha = \underbrace{2^{2p-2}}_{\text{exposant maximal}} + \underbrace{2^{2p-3} + 2^{2p-6} + 2^{p+1} + 2^{p-1} + 2^2}_{\text{suite de l'expression}}$$

- $2^{2p-2} \leqslant \alpha < 2^{2p-1}$
- On peut localiser le bit d'arrondi associé à l'exposant $p-2$:

$$\alpha = \underbrace{2^{2p-2} + 2^{2p-3} + 2^{2p-6} + 2^{p+1} + \underbrace{2^{p-1}}_{\text{parité de la mantisse}}}_{\text{partie supérieure}} + \underbrace{\underbrace{0 \cdot 2^{p-2}}_{\text{bit d'arrondi}} + 2^2}_{\text{partie inférieure}}$$

Calcul symbolique d'arrondi

On cherche RN(α) où

$$\alpha = \underbrace{2^{2p-2}}_{\text{exposant maximal}} + \underbrace{2^{2p-3} + 2^{2p-6} + 2^{p+1} + 2^{p-1} + 2^2}_{\text{suite de l'expression}}$$

- $2^{2p-2} \leqslant \alpha < 2^{2p-1}$
- On peut localiser le bit d'arrondi associé à l'exposant $p-2$:

$$\alpha = \underbrace{2^{2p-2} + 2^{2p-3} + 2^{2p-6} + 2^{p+1} + \underbrace{2^{p-1}}_{\text{parité de la mantisse}}}_{\text{partie supérieure}} + \underbrace{\underbrace{0 \cdot 2^{p-2}}_{\text{bit d'arrondi}} + 2^2}_{\text{partie inférieure}}$$

- partie supérieure $\leqslant \alpha \leqslant$ partie supérieure $+ 2^{p-1}$

Calcul symbolique d'arrondi

On cherche RN(α) où

$$\alpha = \underbrace{2^{2p-2}}_{\text{exposant maximal}} + \underbrace{2^{2p-3} + 2^{2p-6} + 2^{p+1} + 2^{p-1} + 2^2}_{\text{suite de l'expression}}$$

- $2^{2p-2} \leqslant \alpha < 2^{2p-1}$
- On peut localiser le bit d'arrondi associé à l'exposant $p-2$:

$$\alpha = \underbrace{2^{2p-2} + 2^{2p-3} + 2^{2p-6} + 2^{p+1} + \underbrace{2^{p-1}}_{\text{parité de la mantisse}}}_{\text{partie supérieure}} + \underbrace{\underbrace{0 \cdot 2^{p-2}}_{\text{bit d'arrondi}} + 2^2}_{\text{partie inférieure}}$$

- partie supérieure $\leqslant \alpha \leqslant$ partie supérieure $+ 2^{p-1}$
- Finalement, bit d'arrondi $= 0$ donc

$$\text{RN}(\alpha) = 2^{2p-2} + 2^{2p-3} + 2^{2p-6} + 2^{p+1} + 2^{p-1}.$$

Résultats

/home/aplet/Documents/these/trunk/code1/FP.mw* - [Server 1] - Maple 17

File Edit View Insert Format Table Drawing Plot Spreadsheet Tools Window Help

RAIM 2015

```
> FP_Naif := proc(a, b, c, d, p, beta)
  local A, B, C, D, p1, p1_hat, p2, p2_hat, r_tmp, r_hat, p0, p_tmp:

  p0 := 2;

  #Convert input into struct
  A, p_tmp := FP_convert2struct(beta, p, a);
  p0 := max(p0, p_tmp);
  B, p_tmp := FP_convert2struct(beta, p, b);
  p0 := max(p0, p_tmp);
  C, p_tmp := FP_convert2struct(beta, p, c);
  p0 := max(p0, p_tmp);
  D, p_tmp := FP_convert2struct(beta, p, d);
  p0 := max(p0, p_tmp);

  #Algorithm
  p1, p0 := FP_mult(beta, p, p0, A, B);
  p1_hat, p0 := FP_round(beta, p, p0, p1, p);
  p2, p0 := FP_mult(beta, p, p0, C, D);
  p2_hat, p0 := FP_round(beta, p, p0, p2, p);
  r_tmp, p0 := FP_add(beta, p, p0, p1_hat, p2_hat);
  r_hat, p0 := FP_round(beta, p, p0, r_tmp, p);

  #Convert to readable
  return FP_convert2read(beta, r_hat), p0;
  end:
> FP_Naif_num := proc(a, b, c, d, p, beta)
```

Résultats

/home/aplet/Documents/these/trunk/code1/FP.mw* - [Server 1] - Maple 17

File Edit View Insert Format Table Drawing Plot Spreadsheet Tools Window Help

```
> FP_Naif_num := proc(a, b, c, d, p, beta)
  return roundFP(beta, p, roundFP(beta, p, a·b) + roundFP(beta, p, c·d));
  end:
> beta := 2;
  a := 2^(p-1) + 2^(p-2) - 1;
  b := -a;
  c := a + 1;
  d := a - 1;
  r := expand(a·b + c·d);
  r_hat, p0 := FP_Naif(a, b, c, d, p, beta);
  abs_err := r - r_hat;
  rel_err := abs_err / abs(r);
```

$$\beta := 2$$

$$a := 2^{p-1} + 2^{p-2} - 1$$

$$b := -2^{p-1} - 2^{p-2} + 1$$

$$c := 2^{p-1} + 2^{p-2}$$

$$d := 2^{p-1} + 2^{p-2} - 2$$

$$r := -1$$

$$r\_hat, p0 := -2^{p}, 6$$

$$abs\_err := -1 + 2^{p}$$

$$rel\_err := -1 + 2^{p} \qquad (4.6.1)$$

Résultats

```
> for p0_opt from p0 to 2 by -1 do
    print(p0_opt, FP_Naif_num(op(subs(p = p0_opt, [a, b, c, d]))), p0_opt, 2) = subs(p = p0_opt, r_hat));
  od;
```

$$6, -64 = -64$$
$$5, -32 = -32$$
$$4, 0 = -16$$
$$3, 0 = -8$$
$$2, -1 = -4 \qquad (4.6.2)$$

Résultats

```
> for p0_opt from p0 to 2 by -1 do
    print(p0_opt, FP_Naif_num(op(subs(p = p0_opt, [a, b, c, d]))), p0_opt, 2) = subs(p = p0_opt, r_hat));
  od;
```

$$
\begin{aligned}
6,\ -64 &= -64\\
5,\ -32 &= -32\\
4,\ 0 &= -16\\
3,\ 0 &= -8\\
2,\ -1 &= -4
\end{aligned}
\qquad (4.6.2)
$$

Vérification des certificats d'optimalité présentés dans :

- *Error Bounds on Complex Floating-Point Multiplication*, [Brent, Percival et Zimmermann, '07] (2 exemples) ;
- *Further analysis of Kahan's algorithm for the accurate computation of* 2×2 *determinants*, [Jeannerod, Louvet et Muller, '13] (8-12 exemples) ;
- *On the componentwise accuracy of complex floating-point division with an FMA*, [Jeannerod, Louvet et Muller, '13] (8-11 exemples) ;
- *On the error of computing* $ab + cd$ *using Cornea, Harrison and Tang's method*, [Muller, '14] (1 exemple).

Plan de l'exposé

Lien avec les polynômes

Structure de données $\leftrightarrow$ polynômes. On rassemble les exposants ayant le même coefficient linéaire en p pour créer un monôme.

Illustration

Sur l'algorithme de Cornea, Harrison et Tang, avec $x = 2^p$, on a en entrée :

- $a = 2^p - 1 = x - 1$,
- $b = 2^{p-3} + 2^{-1} = 2^{-3}x + 2^{-1}$,
- $c = 2^p - 1 = x - 1$,
- $d = 2^{p-3} + 2^{-2} = 2^{-3}x + 2^{-2}$;

et en sortie :

- $\widehat{s} = 2^{2p-2} = 2^{-2}x^2$.

Formalisation

Notation : $\mathbb{F} = \bigcup_{p \geqslant 2} \mathbb{F}_p$ ensemble des réels à support fini en base β.

Théorème

Soient $a, b \in \mathbb{Q}$ avec $a > 0$ et $x = \beta^{ap+b}$. Pour tout polynôme $Q \in \mathbb{F}[X]$, il existe un polynôme $R \in \mathbb{F}[X]$ et un entier p_0 tels que

$$\mathrm{RN}_p\left(Q(x)\right) = R(x) \quad \text{pour tout } p \geqslant p_0.$$

Formalisation

Notation : $\mathbb{F} = \bigcup_{p \geqslant 2} \mathbb{F}_p$ ensemble des réels à support fini en base β.

Théorème

Soient $a, b \in \mathbb{Q}$ avec $a > 0$ et $x = \beta^{ap+b}$. Pour tout polynôme $Q \in \mathbb{F}[X]$, il existe un polynôme $R \in \mathbb{F}[X]$ et un entier p_0 tels que

$$\mathrm{RN}_p\left(Q(x)\right) = R(x) \quad \text{pour tout } p \geqslant p_0.$$

Remarques :

- On peut calculer p_0 à partir de Q.
- $\deg(R) = \deg(Q)$.

Évaluation-interpolation

$P \in \mathbb{Q}[X]$, $d = \deg(P)$. Il suffit de connaître la valeur de P en $d+1$ points distincts pour retrouver ses coefficients.

Évaluation-interpolation

$P \in \mathbb{Q}[X]$, $d = \deg(P)$. Il suffit de connaître la valeur de P en $d+1$ points distincts pour retrouver ses coefficients.

Conséquence

On peut calculer sur des flottants symboliques avec une méthode numérique : l'évaluation-interpolation.
Calcul d'une opération suivie d'un arrondi :

- choix de $x = \beta^{ap+b}$ et calcul de Q,
- déduction de p_0 et du degré d de R,
- évaluation de $\mathrm{RN}(Q(x))$ pour $x_i = \beta^{ap_i+b}$, $i = 0, \ldots, d$,
- interpolation pour trouver R,
- évaluation de R en x.

Exemple : calcul de l'arrondi d'une multiplication

En base 2, on veut calculer RN(cd) sur les entrées $c = 2^p - 1$ et $d = 2^{p-3} + \frac{1}{2}$.

- $x = 2^{p-3} \Rightarrow c = 8x - 1$ et $d = x + \frac{1}{2}$;

Exemple : calcul de l'arrondi d'une multiplication

En base 2, on veut calculer RN(cd) sur les entrées $c = 2^p - 1$ et $d = 2^{p-3} + \frac{1}{2}$.

- $x = 2^{p-3} \Rightarrow c = 8x - 1$ et $d = x + \frac{1}{2}$;
- $Q(x) = cd = 8x^2 + 3x - \frac{1}{2}$;

Exemple : calcul de l'arrondi d'une multiplication

En base 2, on veut calculer RN(cd) sur les entrées $c = 2^p - 1$ et $d = 2^{p-3} + \frac{1}{2}$.

- $x = 2^{p-3} \Rightarrow c = 8x - 1$ et $d = x + \frac{1}{2}$;
- $Q(x) = cd = 8x^2 + 3x - \frac{1}{2}$;
- $p_0 = 5$ et $\deg(R) = 2$;
- Points d'évaluation : $x_0 = 2^2$ ($p = 5$), $x_1 = 2^3$ ($p = 6$), $x_2 = 2^4$ ($p = 7$) ;

Exemple : calcul de l'arrondi d'une multiplication

En base 2, on veut calculer $\mathrm{RN}(cd)$ sur les entrées $c = 2^p - 1$ et $d = 2^{p-3} + \frac{1}{2}$.

- $x = 2^{p-3} \Rightarrow c = 8x - 1$ et $d = x + \frac{1}{2}$;
- $Q(x) = cd = 8x^2 + 3x - \frac{1}{2}$;
- $p_0 = 5$ et $\deg(R) = 2$;
- Points d'évaluation : $x_0 = 2^2$ $(p = 5)$, $x_1 = 2^3$ $(p = 6)$, $x_2 = 2^4$ $(p = 7)$;
- Évaluation :

$$\begin{aligned} \mathrm{RN}_5(Q(2^2)) &= \mathrm{RN}_5(139, 5) = 136 \\ \mathrm{RN}_6(Q(2^3)) &= \mathrm{RN}_6(535, 5) = 528 \\ \mathrm{RN}_7(Q(2^4)) &= \mathrm{RN}_7(2095, 5) = 2080 \end{aligned}$$

Exemple : calcul de l'arrondi d'une multiplication

En base 2, on veut calculer $\mathrm{RN}(cd)$ sur les entrées $c = 2^p - 1$ et $d = 2^{p-3} + \frac{1}{2}$.

- $x = 2^{p-3} \Rightarrow c = 8x - 1$ et $d = x + \frac{1}{2}$;
- $Q(x) = cd = 8x^2 + 3x - \frac{1}{2}$;
- $p_0 = 5$ et $\deg(R) = 2$;
- Points d'évaluation : $x_0 = 2^2$ $(p = 5)$, $x_1 = 2^3$ $(p = 6)$, $x_2 = 2^4$ $(p = 7)$;
- Évaluation :

$$\begin{aligned} \mathrm{RN}_5(Q(2^2)) &= \mathrm{RN}_5(139, 5) = 136 \\ \mathrm{RN}_6(Q(2^3)) &= \mathrm{RN}_6(535, 5) = 528 \\ \mathrm{RN}_7(Q(2^4)) &= \mathrm{RN}_7(2095, 5) = 2080 \end{aligned}$$

- Interpolation :

$$\begin{aligned} R(X) &= 136\frac{(X-x_1)(X-x_2)}{(x_0-x_1)(x_0-x_2)} + 528\frac{(X-x_0)(X-x_2)}{(x_1-x_0)(x_1-x_2)} + 2080\frac{(X-x_0)(X-x_1)}{(x_2-x_0)(x_2-x_1)} \\ &= 8X^2 + 2X \end{aligned}$$

Exemple : calcul de l'arrondi d'une multiplication

En base 2, on veut calculer $\mathrm{RN}(cd)$ sur les entrées $c = 2^p - 1$ et $d = 2^{p-3} + \frac{1}{2}$.

- $x = 2^{p-3} \Rightarrow c = 8x - 1$ et $d = x + \frac{1}{2}$;
- $Q(x) = cd = 8x^2 + 3x - \frac{1}{2}$;
- $p_0 = 5$ et $\deg(R) = 2$;
- Points d'évaluation : $x_0 = 2^2$ $(p = 5)$, $x_1 = 2^3$ $(p = 6)$, $x_2 = 2^4$ $(p = 7)$;
- Évaluation :

$$\begin{aligned} \mathrm{RN}_5(Q(2^2)) &= \mathrm{RN}_5(139, 5) = 136 \\ \mathrm{RN}_6(Q(2^3)) &= \mathrm{RN}_6(535, 5) = 528 \\ \mathrm{RN}_7(Q(2^4)) &= \mathrm{RN}_7(2095, 5) = 2080 \end{aligned}$$

- Interpolation :

$$\begin{aligned} R(X) &= 136\frac{(X-x_1)(X-x_2)}{(x_0-x_1)(x_0-x_2)} + 528\frac{(X-x_0)(X-x_2)}{(x_1-x_0)(x_1-x_2)} + 2080\frac{(X-x_0)(X-x_1)}{(x_2-x_0)(x_2-x_1)} \\ &= 8X^2 + 2X \end{aligned}$$

- $\mathrm{RN}(cd) = R(x) = 2^{2p-3} + 2^{p-2}$.

Conclusion

- Implantation des opérations +, -, x et FMA en Maple.
- Vérification rapide de nombreux certificats d'optimalité.
- Méthode de calcul alternative : l'évaluation-interpolation.

Conclusion

- Implantation des opérations +, -, x et FMA en Maple.
- Vérification rapide de nombreux certificats d'optimalité.
- Méthode de calcul alternative : l'évaluation-interpolation.

Perspectives

- Comparaison des deux approches.
- Implantation des autres modes d'arrondi.
- Ajout de la division.
- Vers une formalisation en Coq.
- Gestion d'une base symbolique.