Les expansions flottantes (Tâche 2)

- Introduction
  - Le calcul du déterminant
  - Définition des expansions
- Propriétés
  - Propriétés sur les flottants
  - Brique de base des expansions
  - Propriétés sur les expansions
  - Lemme
- Algorithmes
  - Adaptativité
  - Pseudo-expansions
  - Opérations élémentaires sur les pseudo-expansions
- Conclusion

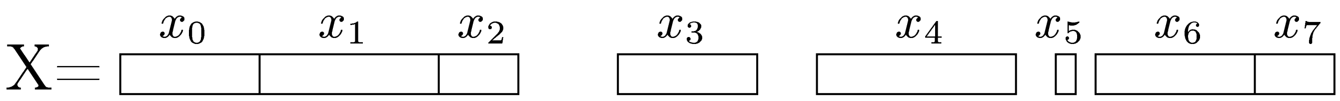
Calcul de déterminant

- Méthodes spécifiques :
  - Géométrie algorithmique avertie
  - Calculs adaptatifs
- Méthodes générales :
  - L'arithmétique stochastique (LIP6)
  - La multiprécision :
    - Maple, Mathematica...
    - GMP
    - MPFun
    - Arithmétique modulaire
    - Expansions, pseudo-expansions

Les expansions

Une **expansion** est la représentation d'un grand nombre flottant **x** par un n-uplet de nombres flottants représentables en machine $(x_n, ..x_1, x_0)$ tels que :

- **x** soit égal à la somme exacte des x_i ;
- les chiffres significatifs des x_i ne se superposent pas ;
- les x_i soient triés selon leur valeur absolue.



Fonctions α, ω et ulp

53 bits

$$x = \boxed{1.0110100111010010000000000 \ldots 0}\ .2^{18}$$

$\alpha(x)$ $\omega(x)$ $ulp(x)$

$\alpha(x) = 2^{18}$ $\omega(x) = 2^{3}$ $ulp(x) = 2^{-34}$

$$\begin{aligned}
\alpha(x) &= \inf\left\{w \in \mathbb{N}^* \,|\, w = 2^k; k \in \mathbb{Z}; 2w > |x|\right\} \\
\omega(x) &= \inf\left\{w \in \mathbb{N}^* \,|\, w = 2^k; k \in \mathbb{Z}; wx \in \mathbb{Z}\right\} \\
ulp(x) &= \alpha(x) \cdot ulp(1)
\end{aligned}$$

Propriétés (1)

Soient a et b deux nombres flottants. On a

$$\alpha(a) \leq |a| < 2\alpha(a) \quad (1)$$

$$\alpha(a)\alpha(b) \leq \alpha(ab) \leq 2\alpha(a)\alpha(b) \quad (2)$$

$$|b| \cdot \alpha(a) < 2|a| \cdot \alpha(b) \quad (3)$$

$$
\begin{array}{rl}
 & a = \boxed{1.011010011101001000000000000000} \\
+ & \\
 & b = \qquad \boxed{1.100110010100100011011101001100} \\
\hline
\end{array}
$$

a' b'

Brique de base

Brique de base : *Soient a et b deux nombres flottants et a' et b' deux nombres flottants tels que $a' = \circ(a+b)$ et $a' + b' = a + b$. On a*

$$\begin{aligned} ulp(a') &\leq 2\max(ulp(a), ulp(b)) \quad (4) \\ \omega(b') &\geq \min\left(\omega(a), \omega(b)\right) \quad (5) \\ \alpha(b') &\leq \min\left(\frac{ulp(a')}{2}, \frac{\omega(a')}{4}\right) \quad (6) \end{aligned}$$

Définition *Soit* $\mathbf{A}$ *un ensemble de nombres flottants et leur somme $A = \sum_{a \in \mathbf{A}} a$. On dit que $\widehat{A}$ est une* **approximation précise** *de A si on a :*

- *$\widehat{A}$ est l'arrondi au plus près de A ou*
- *$\widehat{A}$ est l'arrondi au plus près de A_1 avec $|A - A_1| \leq ulp(ulp(\widehat{A}))$*

Propriétés (2)

Soit $\mathbf{A}$ *un ensemble de nombres flottants et leur somme* $A = \sum_{a \in \mathbf{A}} a$ *telle que* $\widehat{A}$ *est une approximation précise de* A*. Les propriétés suivantes sont satisfaites.*

$$\left|A - \widehat{A}\right| \leq ulp\left(\frac{1}{2} + ulp\right)\alpha(\widehat{A}) \tag{7}$$

$$\frac{\left|\widehat{A}\right|}{2} \leq |A| \leq 2\left|\widehat{A}\right| \tag{8}$$

$$\frac{1}{2}\alpha(\widehat{A}) \leq \alpha(A) \tag{9}$$

$$\alpha(A) \leq \alpha(\widehat{A}) \tag{10}$$

Lemme

Lemme 1 *Soit* $\mathbf{A}$ *un ensemble de nombres flottants et leur somme* $A = \sum_{a \in \mathbf{A}} a$ *telle que* $\widehat{A}$ *est une approximation précise de* A*. La propriété suivante est satisfaite pour tout nombre flottant* b *et pour tout ensemble* $\mathbf{C}$ *de nombres flottants tel que* $C = b.A$ *—avec* $C = \sum_{c \in \mathbf{C}} c$ *et* $\widehat{C}$ *est une approximation précise de* C*.*

$$\alpha(\widehat{C}) \leq 2\alpha(b) \cdot \alpha(\widehat{A})$$

Adaptativité des calculs

Avec les expansions :

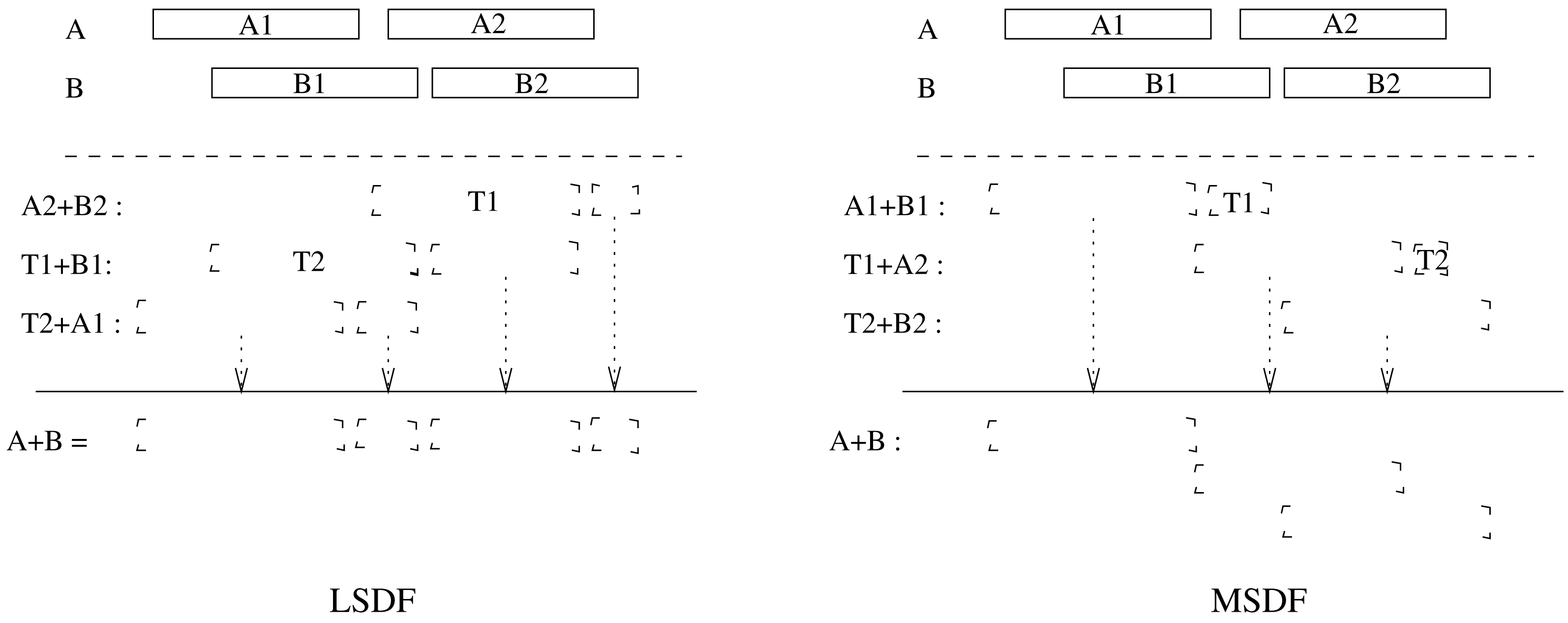
Addition, multiplication : bits de poids faibles en tête (LSDF)

Division : bits de poids forts en tête (MSDF)

$\Rightarrow$ Passage aux opérateurs bits de poids forts en tête.

$\Rightarrow$ Adaptativité. Possibilité de faire du calcul en ligne.

$\Rightarrow$ Introduction des pseudo-expansions.

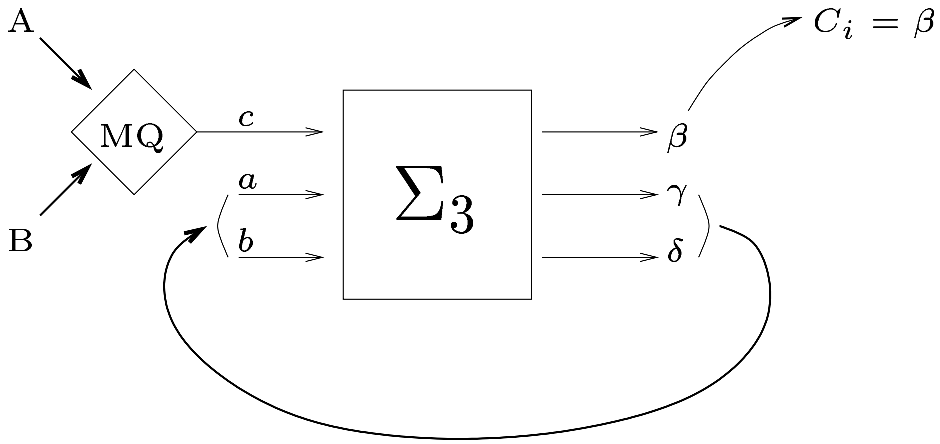

A
A1
A2
B
B1
B2
A2+B2 :
T1
T1+B1:
T2
T2+A1 :
A+B =
LSDF
A
A1
A2
B
B1
B2
A1+B1 :
T1
T1+A2 :
T2
T2+B2 :
A+B :
MSDF

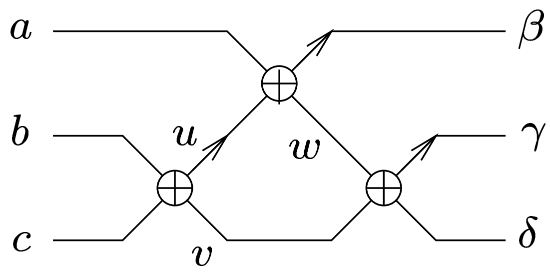
Les pseudo-expansions

Une **pseudo-expansion** est la représentation d'un grand nombre flottant **x** par un n-uplet de nombres flottants représentables en machine $(x_1, x_2, ..x_n)$ tels que :

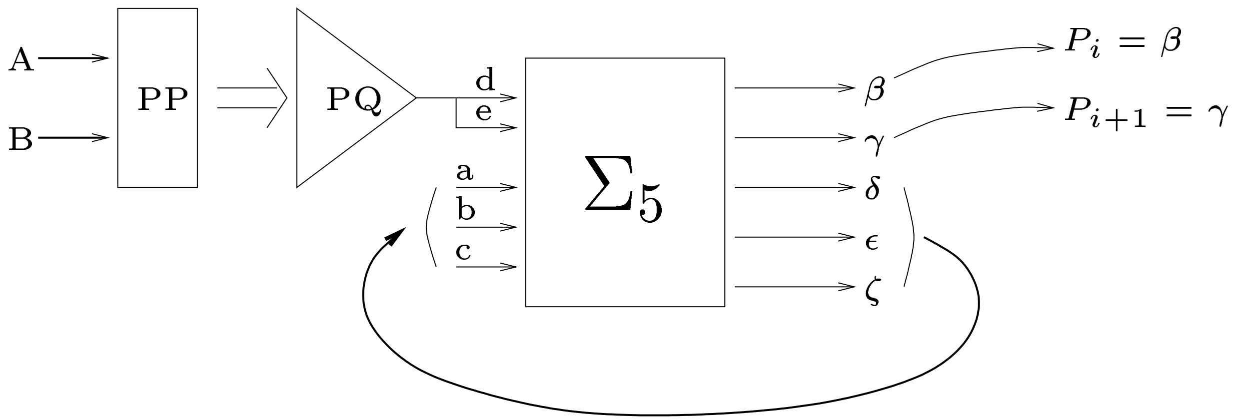
- **x** soit égal à la somme exacte des x_i ;
- les x_i soient triés selon leur valeur absolue ;
- deux composantes, séparées par au moins une autre composante, ne se superposent pas.

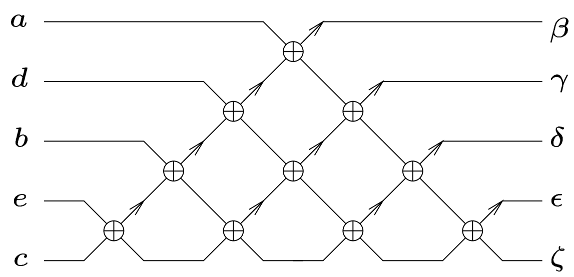
Addition


A
B
MQ
c
a
b
Σ_3
β
γ
δ
$C_i = \beta$

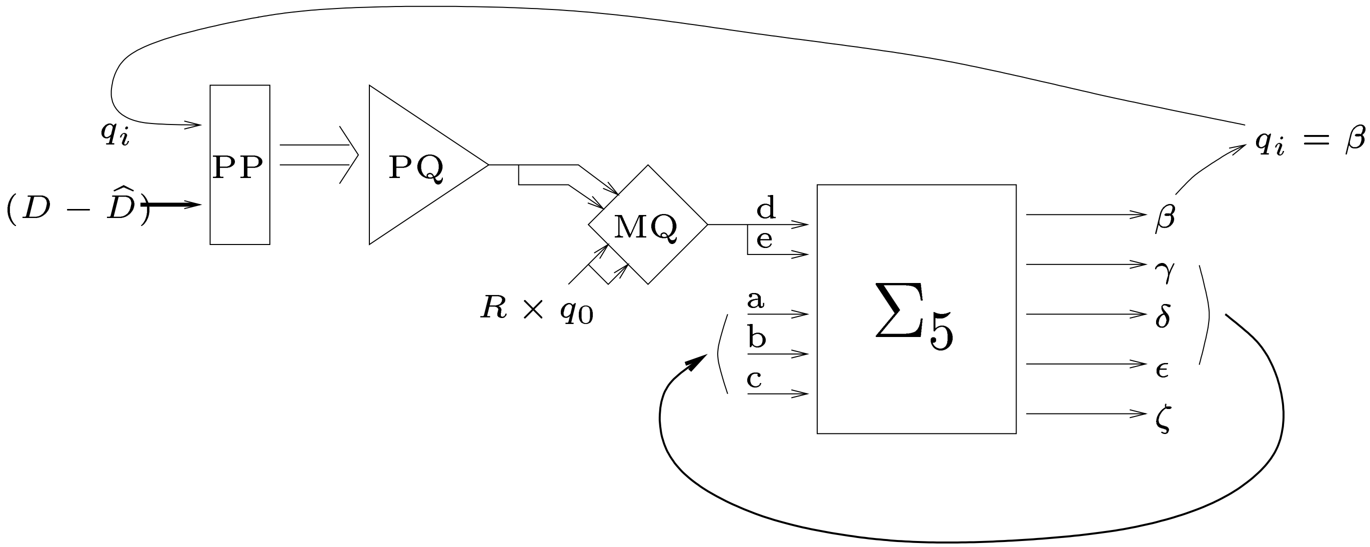

a
b
c
u
v
w
β
γ
δ

Multiplication





Division


q_i
$(D - \widehat{D})$
PP
PQ
MQ
$R \times q_0$
d
e
a
b
c
Σ_5
β
γ
δ
ϵ
ζ
$q_i = \beta$

Conclusion

On a une bibliothèque de calcul sur les pseudo-expansions

$\Rightarrow$ On sait calculer le déterminant de manière exacte et efficace grâce à cette bibliothèque

$\Rightarrow$ On peut faire du calcul adaptatif.

$\Rightarrow$ On est prêt à faire du calcul en multi-précision où la précision est fixée (continuation avec la bibliothèque de quad-double)