Outils pour un calcul numérique fiable
22-24 septembre 1999
Paris (Jussieu)

CONTRIBUTION À MPFR

CALCUL RAPIDE ET EXACT EN PRÉCISION ARBITRAIRE DE LA MOYENNE ARITHMÉTICO-GÉOMÉTRIQUE

Sylvie Boldo,
élève à l'ENS Lyon,
travail effectué dans l'équipe PolKA du LORIA,
dirigée par Mr Paul Zimmermann.

PLAN

1 Les problèmes du calcul exact

2 La bibliothèque MPFR

3 Calcul exact de la moyenne arithmético-géométrique

4 Ma contribution à MPFR

5 Complexité

6 Conclusions

1 Les problèmes du calcul exact

Il n'existe qu'un nombre fini de
"nombres machine".
Les autres réels sont **arrondis**.

1.1 Arrondis

La norme IEEE-754 définit les 4 modes d'arrondi suivants :

- arrondi vers $-\infty$: $\nabla(x)$;
- arrondi vers $+\infty$: $\Delta(x)$;
- arrondi vers 0 : $\mathcal{Z}(x)$;
- arrondi au plus proche : $\mathcal{N}(x)$.

1.2 Un résultat exact ?

Le résultat EXACT d'une suite de calculs est celui donné par l'algorithme idéal :

- effectuer **chaque calcul** avec une PRÉCISION INFINIE ;
- arrondir *ensuite* le résultat obtenu selon le mode d'arrondi demandé.

1.3 Dilemme du fabricant de tables

Soit

$$r = +1.\underbrace{b\ldots\ldots\ldots 0}_{q \text{ bits}}\ \underbrace{11\ldots\ldots 111}_{p \text{ bits}}$$
$$= x0\ 1^p$$

Supposons

- qu'il faille arrondir r vers zéro à la précision q ;
- que l'erreur soit de 2^{-p-q},

alors :

$$x01^{p-1}0 \leq r_{exact} \leq x10^p.$$

Or $\mathcal{Z}(x01^{p-1}0) = x0$ et $\mathcal{Z}(x10^p) = x1$.

Il est **impossible de déterminer** $\mathcal{Z}(r)$ à la précision q. Il aurait fallu calculer avec plus de précision.

2 La bibliothèque MPFR

2.1 Qu'est-ce que MPFR ?

- **bibliothèque de fonctions** ;
- calcul en **précision quelconque** ;
- donnant le **résultat exact** ;
- en langage C ;
- développée au LORIA à Nancy ;
- gratuite et accessible à tous.

2.2 L'utilisation de MPFR

Pour obtenir le résultat EXACT d'un calcul, il suffit de fournir :

- les **opérandes** de précisions arbitraires,
- la **précision voulue** pour le résultat,
- le **mode d'arrondi** souhaité.

3 Calcul exact de la moyenne arithmético-géométrique

3.1 Définition

AG(a, b) est la limite commune de :

$$u_0 = \min(a, b) \qquad \text{et} \qquad v_0 = \max(a, b)$$

$$u_{n+1} = \sqrt{u_n \, v_n} \qquad \text{et} \qquad v_{n+1} = \frac{u_n + v_n}{2}$$

3.2 Calcul approché

Chaque calcul a une précision p.

$$\tilde{u}_0 = \mathcal{N}(u_0) \qquad \text{et} \qquad \tilde{v}_0 = \mathcal{N}(v_0)$$

$$\tilde{u}_{n+1} = \mathcal{N}\left(\sqrt{\mathcal{N}(\tilde{u}_n \, \tilde{v}_n)}\,\right)$$

$$\tilde{v}_{n+1} = \frac{\mathcal{N}(\tilde{u}_n + \tilde{v}_n)}{2}$$

Si $\delta = 2^{-p}$, il est alors facile de montrer que :

$$u_n\,(1-\delta)^{\frac{3}{2}n+1} \leq \tilde{u}_n \leq u_n\,(1+\delta)^{\frac{3}{2}n+1}$$

$$v_n\,(1-\delta)^{\frac{3}{2}n+1} \leq \tilde{v}_n \leq v_n\,(1+\delta)^{\frac{3}{2}n+1}$$

3.3 Erreur commise

$$\boxed{\frac{|\tilde{v}_n - \mathrm{AG}(u_0, v_0)|}{8\,v_0} \leq n \cdot 2^{\frac{3}{2}n-p-1} + 4\,\frac{-\,u_0 \cdot 2^n}{v_0 - u_0}}$$

Le nombre d'itérations optimal est environ :

$$n_{opt} = \left\lceil \log_2\left(\frac{v_0 - u_0}{2\,u_0}\right) + \log_2 p \right\rceil$$

L'erreur commise en fonction de la précision est donc :

$$\frac{|\tilde{v}_{n_{opt}} - AG(u_0, v_0)|}{v_0} \leq$$

$$8 \cdot 2^{-p} \cdot \log_2 \left(\frac{p\ (v_0 - u_0)}{2\ u_0} \right) \cdot \left(\frac{p\ (v_0 - u_0)}{2\ u_0} \right)^{\frac{3}{2}}$$

3.4 Précision nécessaire

Pour obtenir une précision q sur le résultat, il faut calculer avec une précision p telle que :

$$p \geq q + 3 + \frac{3}{2}\ \log_2 \left(\frac{p\ (v_0 - u_0)}{2\ u_0} \right) + \log_2 \left(\log_2 \left(\frac{p\ (v_0 - u_0)}{2\ u_0} \right) \right)$$

Algorithme

agm(a, b, rnd_mode, q)

si a ou b est négatif, renvoyer NaN

si a ou b est nul, renvoyer 0

si $a = b$, renvoyer a

$p = p_{min}(q) + 2$

$u = a$

$v = b$

répéter

calculer $err(p)$

tant que $|u - v|$ varie

$tmp = \mathcal{N}\left(\sqrt{\mathcal{N}\left(u \cdot v\right)}\right)$

$v = \dfrac{\mathcal{N}(u + v)}{2}$

$u = tmp$

s'il est possible d'arrondir v
à la précision q,
suivant l'arrondi rnd_mode
et connaissant $err(p)$

renvoyer cet arrondi

sinon

$p = p + 5$

4 Ma contribution à MPFR

- implantation de la moyenne arithmético-géométrique

- contribution au débogage de la bibliothèque

- implantation du logarithme

Pour obtenir p bits du logarithme de x :

$$\ln x = \frac{\pi}{2\ \mathrm{AG}(4/s, 1)} - m \ln 2 + \circ(\ln(x) \cdot 2^{-p})$$

où $s = x \cdot 2^m > 2^{p/2}$

Les erreurs sont alors dues :

– à la dérive du calcul ;

– à la perte de bits significatifs lors de la soustraction : $\left\lceil \log_2 \left(\frac{p \ln 2}{2\ |x-1|} \right) \right\rceil$ bits perdus.

5 Complexité

5.1 Complexité théorique

$\mathsf{M}(n)$: complexité de la **multiplication** de deux nombres de taille n.

$$\mathsf{AGM}(n) = 7\ \mathsf{M}(n)\ \left(\log_2\left(\frac{v_0 - u_0}{2\ u_0}\right) + \log_2 n\right)$$

$$\mathsf{LN}(n)\ =\ \frac{8}{7}\ (\mathsf{PI}(n) + \mathsf{LN2}(n)) + 16\ \mathsf{M}(n)\ \log_2 n$$

5.2 Dans la réalité

- bien plus **rapide** que MAPLE ;
- équivalent à PARI jusqu'à 2000 bits de précision, nettement plus **rapide** au-delà ;

- 85% du temps d'agm dans les **racines carrées** ;
- temps équivalents pour agm, π et ln 2 ;
- 80% du temps dans les fonctions de GMP ;
- impossible d'arrondir le résultat obtenu dans **6% des cas**.

6 Conclusions

6.1 Les réalisations

- un algorithme
  - prouvé,
  - implanté,
  - validé.
- un programme
  - débogué,
  - testé,
  - rapide.

6.2 Les perspectives de MPFR :

- **s'enrichir** d'autres fonctions mathématiques usuelles ;
- être bêta-testée.