MÉMOIRES DE LA S. M. F.

FRANCIS SERGERAERT

Utilisation des flottants du hardware pour le calcul rationnel exact

Mémoires de la S. M. F., tome 49-50 (1977), p. 187-194

<http://www.numdam.org/item?id=MSMF_1977__49-50__187_0>

Calculateurs en Math. (1975 - LIMOGES)
Bull. Soc. math. France
Mémoire 49-50, 1977, p. 187-194

UTILISATION DES FLOTTANTS DU HARDWARE POUR LE CALCUL RATIONNEL EXACT

(Francis SERGERAERT - POITIERS)

1 - Enoncé du problème -

Soit $\mathcal{O}_n$ l'anneau des germes de fonctions holomorphes définies au voisinage de 0 dans $\mathbb{C}^n$. On note $\mathcal{M}$ l'idéal maximal unique de $\mathcal{O}_n$. Si $f \in \mathcal{O}_n$, on note Jf l'idéal de $\mathcal{O}_n$ engendré par les dérivées partielles du premier ordre de f ; c'est l'idéal jacobien de f. On dit que f a une singularité en 0 si $Jf \subset \mathcal{M}$, autrement dit si toutes les dérivées partielles de f s'annulent en 0, ou encore si $df(0) = 0$. Cette singularité est isolée si 0 est racine isolée de $df = 0$; le Nullstellensatz implique alors qu'il existe un entier k tel que $Jf \supset \mathcal{M}^k$.

On note alors :

$$m(f) = \inf \{m : f^m \in Jf\}$$

En 1973 il existait deux conjectures au sujet de la valeur de l'entier $m(f)$ dans le cas où $n=2$, c'est-à-dire dans le cas où f est un germe de deux variables complexes. La première, attribuée à J. Mather par Wall [W1], affirmait que $m(f) \leq 2$.

La seconde suppose en plus que f est un germe irréductible. On peut alors définir (voir par exemple [L1]) l'entier $p(f)$, nombre des paires de Puiseux de f. A titre d'exemple :

$$p(y^2 - x^3) = 1$$

$$p(y^4 - 2x^3y^2 - 4x^5y + x^6 - x^7) = 2$$

$$p(y^8 - 4x^3y^6 - 8x^5y^5 + 6x^6y^4 - 26x^7y^4 + 16x^8y^3$$
$$-4x^9y^2 - 24x^9y^3 + 36x^{10}y^2 - 8x^{11}y - 20x^{11}y^2 + x^{12}$$
$$+16x^{12}y + 6x^{13} - 8x^{13}y + 21x^{14} - x^{15}) = 3$$

Remarquons que ces trois polynômes sont polynôme minimal de $x^{3/2}$, $x^{3/2} + x^{7/4}$, $x^{3/2} + x^{7/4} + x^{15/8}$ respectivement.

On note $P(x,y)$ le troisième de ces polynômes.

La seconde conjecture, énoncée par A'Campo, affirmait que si f est un germe irréductible de deux variables, alors $m(f) \geq p(f)$.

On voit que les deux conjectures sont contradictoires pour $f=P$. Il était donc tentant de déterminer, pour le cas particulier de P , qui avait raison, de Mather ou d'A'Campo. C'est ce calcul que nous voulons maintenant expliquer.

2 - *Description de l'algorithme* -

Il faut déterminer s'il est possible de trouver A , A' tels que :

$$P^2 = AP'_x + A'P'_y$$

Notons :

$$P'_x = \Sigma\, X_{i,j}\, x^i y^j \quad , \quad P'_y = \Sigma\, Y_{i,j}\, x^i y^j$$

$$A = \Sigma\, a_{i,j}\, x^i y^j \quad , \quad A' = \Sigma\, a'_{i,j}\, x^i y^j$$

$$P^2 = \Sigma\, b_{i,j}\, x^i y^j$$

On utilise la méthode des coefficients indéterminés qui donne les équations :

$$\sum_{\substack{0\leq i'\leq i \\ 0\leq j'\leq j}} a_{i',j'}\, X_{i-i',j-j'} + a'_{i',j'}\, Y_{i-i',j-j'} = b_{i,j} \qquad (Eq_{i,j})$$

C'est un sytème linéaire en les $a_{i,j}$ et $a'_{i,j}$. On classe les équations $Eq_{i,j}$ d'abord par rapport à $i+j$, puis par rapport à j . La première équation non triviale donne une relation :

$$a_{i_1,j_1} = C_{i_1,j_1} + \sum_{(i,j)\neq(i_1,j_1)} \alpha_{i,j}\, a_{i,j} + \alpha'_{i,j}\, a'_{i,j}$$

On reporte ce résultat dans la deuxième équation non triviale, on obtient (par exemple) :

$$a'_{i_2,j_2} = C'_{i_2,j_2} + \Sigma \ldots$$

et on continue.

Après substitution dans une $Eq_{i,j}$ non triviale des résultats précédemment obtenus, trois cas peuvent se produire :

1°) $Eq_{i,j}$ devient triviale $(0=0)$. C'est que $Eq_{i,j}$ est conséquence des $Eq_{i',j'}$ pour $(i',j') < (i,j)$.

2°) $Eq_{i,j}$ devient une relation "impossible" $(0=b\neq 0)$: les coefficients de toutes les inconnues sont annulés, mais le "second membre" n'est pas nul. Le système linéaire n'admet donc pas de solution ; c'est que $P^2 \notin JP$, ce qui donne raison à A' Campo.

3°) $Eq_{i,j}$ devient une relation non triviale "possible" ; on obtient alors une nouvelle relation :

$$a_{i_p,j_p} = \ldots \quad \text{ou} \quad a'_{i_p,j_p} = \ldots$$

et on poursuit.

Dans le 1°) on déduit que

$$x^i y^j \notin JP \quad \text{(regarder !)}.$$

En particulier $\mathfrak{m}^{i+j} \notin JP$ et ceci donne un renseignement sur le plus petit k tel que $JP \supset \mathfrak{m}^k$; on a $k > i+j$.

Dans le 3°) on peut affirmer que

$$x^i y^j \in JP + \sum_{\substack{(i',j')>(i,j) \\ i'+j' = i+j}} \mathbb{R}\, x^{i'} y^{j'} + \mathfrak{m}^{i+j+1}$$

En particulier, s'il existe un entier k tel que

a) Le cas 2°) ne se produit pas pour $i+j \leq k$

b) Le cas 3°) se produit pour tout (i,j) tel que $i+j=k$,

on voit que :

a) $\Rightarrow P^2 \in JP + \mathfrak{m}^{k+1}$

b) $\Rightarrow \mathfrak{m}^k \subset JP + \mathfrak{m}^{k+1}$

Mais le lemme de Nakayama implique alors que $\mathfrak{m}^k \subset JP$, a fortiori $\mathfrak{m}^{k+1} \subset JP$, et donc $P^2 \in JP$; c'est que Mather a raison.

Comme on sait d'avance qu'il existe un entier k tel que $JP \supset \mathcal{M}^k$, on voit qu'on a bien un algorithme pour déterminer qui, de Mather ou de A'Campo, a raison au sujet de P .

Cependant le calcul est trop compliqué pour être effectué à la main ; il est nécessaire de recourir à l'ordinateur. L'utilisation de l'ordinateur soulève pourtant de nouvelles difficultés qu'on va maintenant exposer.

3 - Résoudre un système linéaire rationnel à l'aide d'un ordinateur -

Supposons qu'on veuille résoudre à l'aide d'un ordinateur le système linéaire à coefficients et seconds membres rationnels :

$$\sum_{i=1}^{m} a_{ij} x_i = b_j \qquad (1 \leq j \leq n).$$

C'est un problème très classique. La difficulté en ce qui nous concerne est double. D'abord, la taille du système, qu'on ne connaît d'ailleurs pas au départ de l'algorithme, est grande : a posteriori il se trouve que $m = 240$ et $n = 225$! La seconde difficulté est plus intéressante. Il faut décider si oui ou non le système proposé admet une solution. Il n'est pas question pour ce faire d'utiliser les flottants fournis par les constructeurs. Supposons en effet qu'on veuille déterminer la nature du système

$$3x_1 + x_2 = 0$$

$$3x_1 + x_2 = 1$$

Un calcul en flottants déduira de la première équation que $x_1 = -0.33x_2$, résultat qui, reporté dans la seconde donne $-0.99x_2 + x_2 = 0.01x_2 = 1$. On obtient ainsi la réponse : le système admet une solution ! Ce résultat inexact provient des erreurs commises par les opérateurs binaires lors d'opération sur des flottants.

Il faut donc chercher à travailler dans Q : un "nombre" sera un couple (p,q) d'entiers du hardware ; (p,q) représente le rationnel p/q . Un jeu de sous-programmes permet d'effectuer les opérations sur de tels nombres selon les règles bien connues.

On rencontre alors une nouvelle difficulté : les rationnels résultats des calculs demandent très rapidement pour être représentés des entiers p,q dépassant la capacité maximum prévue par le constructeur, même si on réduit systématiquement p/q à sa forme irréductible.

là encore la solution est classique : utiliser des sous-programmes de calcul en multi-précision. Cependant, cette méthode est très lourde : le temps de calcul devient très long, et on se heurte à de redoutables problèmes d'encombrement mémoire. Il n'est pas sûr qu'une telle méthode réussise à traiter notre problème ; de toute façon, elle coûterait cher !

Nous avons utilisé une autre méthode, d'esprit complètement différent, et qui s'est révélée très efficace.

Soit Z l'ensemble des entiers positifs représentables par un entier du hardware, et F l'ensemble des réels positifs représentables par un flottant du hardware. Décidons qu'un "nombre" est un quadruplet (f,s,α,β) où f est un bit, indicateur de format, qui vaut 0 ou 1, s est un signe qui vaut $-1,0$ ou 1 ; si $f=1$, alors $\alpha,\beta \in Z$ et le nombre représenté est le rationnel $s\alpha/\beta$; si $f=0$, alors $\alpha,\beta \in F$, $0 < \alpha < \beta$, et le nombre représenté est un réel (inconnu) élément de $[s\alpha,s\beta]$; dans ce cas, on ne connaît qu'une approximation par défaut et une approximation par excès du nombre représenté ; par contre, dans le premier cas, on connaît le nombre exactement.

On va expliquer maintenant comment on effectue les opérations d'addition et de multiplication de ces nombres. Supposons d'abord qu'on veuille effectuer :

$$(0,s,\alpha,\beta) \times (0,s',\alpha',\beta') = ?$$

On détermine des flottants $\alpha'',\beta'' \in F$ tels que $\alpha'' \leq \alpha\alpha' < \beta\beta' \leq \beta''$. Le résultat est alors $(0,ss',\alpha'',\beta'')$; naturellement on cherchera $\beta''-\alpha''$ aussi petit que possible. Supposons maintenant qu'on veuille effectuer :

$$(1,s,p,q) \times (0,s',\alpha',\beta') = ?$$

On cherche alors $\alpha'',\beta'' \in F$ tels que $\alpha'' \leq p\alpha'/q < p\beta'/q \leq \beta''$. Le résultat est encore $(0,ss',\alpha'',\beta'')$.

Si on veut calculer :

$$(1,s,p,q) \times (1,s',p',q') = ?$$

Soit $p"/q" = pp'/qq'$ la représentation irréductible du produit des modules. Si $p",q" \in Z$, le résultat sera $(1,ss',p",q")$. Si par contre $p"$ ou $q"$ dépasse la capacité prévue par le conducteur, on détermine deux flottants $\alpha",\beta" \in F$ tels que $\alpha" \leq p"/q" \leq \beta"$; on décidera alors que le résultat de la multiplication est $(0,ss',\alpha",\beta")$. L'esprit de la méthode est très simple : si possible on calcule dans Q ; sinon on calcule sur des intervalles de sécurité de $\mathbb{R}$.

On procède de la même façon pour l'addition. Il se présente cependant une difficulté. Soit à calculer :

$$(0,+,\alpha,\beta) + (0,-,\alpha',\beta') = ?$$

Supposons d'abord que $\alpha > \beta'$; on cherche alors $\alpha",\beta"$ dans F tels que $0 < \alpha" \leq \alpha-\beta' < \beta-\alpha' \leq \beta"$, et le résultat est $(0,+,\alpha",\beta")$. De même si $\beta < \alpha'$. Cependant, si $\alpha \leq \beta'$ et $\beta \geq \alpha'$, le résultat est un nombre indéterminé de l'intervalle $[\alpha-\beta', \beta-\alpha']$ qui contient 0 et <u>dont on ne peut déterminer le signe</u>. On considère qu'une telle addition est une <u>erreur</u> dans le même sens qu'on dit qu'une multiplication provoquant un dépassement de capacité est une erreur : le résultat n'est pas représentable.

En ce qui nous concerne, un nombre indéterminé d'un intervalle contenant 0 ne peut être considéré : il risquerait d'empêcher de conclure sur la nature du système linéaire étudié. On voit que notre méthode ne permet de déterminer certains nombres qu'avec une certaine approximation ; mais, tant que l'"erreur" qu'on vient d'étudier ne se produit pas, on connaît de façon certaine si un "nombre" est nul ou non. Il reste donc à espérer qu'aucune "erreur" ne se produise ; cela revient à espérer que si, au cours du calcul le résultat d'une soustraction n_1-n_2 est nul, les nombres n_1 et n_2 ne résultent pas d'une trop grande quantité de calculs, et sont encore représentés sous forme rationnelle ; on peut alors garantir le résultat nul.

Dans notre cas, aucune "erreur" ne s'est produite : on trouve ainsi en 20 secondes d'IBM/370-168 que la conjecture d'A'Campo est fausse.

La méthode est très efficace, car un "nombre" a un encombrement mémoire fixe et réduit ; par ailleurs les sous-programmes de calcul ne contiennent aucune séquence itérative : on gagne sur tous les tableaux, mémoire et temps de calcul.

Pour résoudre le système linéaire, le choix du pivot est crucial ; nous avons choisi la méthode suivante : chercher si possible un pivot rationnel, et, si c'est le cas un pivot $(1,s,p,q)$ tel que p^2+q^2 soit minimum (de façon à "compliquer" le moins possible les calculs à suivre). La méthode naïve consistant à choisir le plus grand pivot possible échoue sur une "erreur" !

Dans nos sous-programmes, les résultats rationnels étaient systématiquement réduits. On peut imaginer la variante suivante : un nombre est un quintuplet (f,r,s,p,q) où f,s,p,q sont comme avant, et où r est un bit indiquant, dans le cas où $f=1$, si p/q est réduit. On peut ainsi n'effectuer les réductions, qui prennent beaucoup de temps, qu'à bon escient, quand elles présentent un intérêt. On gagne ainsi 20% du temps de calcul. Cependant le choix du pivot est moins rigoureux : on pourrait préférer le pivot $(1,1,1,11,4)$ au pivot $(1,0,1,564764,564764)$; ce serait un mauvais choix.

Dans le même temps Briançon [B1] démontrait indépendamment la conjecture de Mather. Depuis il a montré en collaboration avec Skoda un résultat beaucoup plus général [B2] impliquant que pour un germe de n variables f, on a $m(f) \leq n$. Nous espérons que la méthode de calcul que nous venons d'exposer pourra néanmoins rendre d'autres services en calcul rationnel.

Les programmes ont été rédigés en PL/I Optimizer d'IBM. L'utilisation de ce langage très puissant rend ici de grands services: contrôle absolu sur la représentation interne, maniement de structures, aides puissantes à la mise au point ont permis de réaliser sans aucune difficulté l'implémentation de notre méthode.

BIBLIOGRAPHIE

[B1] BRIANÇON Joël - A propos d'une question de J. Mather ; preprint ; Nice 1973.

[B2] BRIANÇON, Joël ; SKODA, Henri - Sur la clôture intégrale d'un idéal de germes de fonctions holomorphes en un point de $\mathbb{C}^n$. Comptes-Rendus de l'Académie des Sciences de Paris. T.278, Série A, pp. 949-951, 1974.

[L1] LÊ Dũng Tráng - Sur les noeuds algébriques. Compositio math., 1972, n° 25, pp. 281-321.

[W1] WALL, C.-T.-C. - Sur le théorème de préparation, in Proceedings of Liverpool Singularities. Springer Lecture Notes 192.

Francis SERGERAERT
Département de Mathématiques
Faculté des Sciences
86022 POITIERS