Agrégation externe de mathématiques, session 2017
Épreuve de modélisation, option informatique

Résumé : Nous nous intéressons à l'évaluation du signe d'expressions polynomiales en arithmétique flottante, et à la certification du résultat. Les questions sont illustrées sur des prédicats géométriques.

Mots clefs : précision numérique, représentation approchée des réels, géométrie

---

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, vous sont proposées en fin de texte. Vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury appréciera que la discussion soit accompagnée d'exemples traités sur ordinateur.*

1. Motivation : les propriétés usuelles de la géométrie euclidienne

Plaçons nous dans le plan euclidien $\mathbf{R}^2$ muni d'un repère orthonormé direct. Les coordonnées d'un point p sont notées (x_p, y_p). La droite passant par deux points p, q est notée pq.

Un prédicat géométrique est une fonction prenant un n-uplet de points en argument et dont la valeur est un signe : $+, -$, ou 0.

Soient a, b, c, d quatre points, ordonnés en abscisses : $x_a < x_b < x_c < x_d$. Si c est au dessus de ab et d est au dessus de bc, alors d est évidemment au dessus de ab. En machine, si les nombres sont représentés en virgule flottante (par exemple en utilisant le type `double` ou `float`), il est possible que le calcul donne le contraire ! En effet, les nombres utilisés ne représentent pas fidèlement le corps des réels $\mathbf{R}$, et les évaluations d'expressions arithmétiques souffrent d'erreurs d'arrondi.

Théorème 1. *Soient quatre points p, q, r, v dans $\mathbf{R}^2$. Si les triangles pqv, qrv et rpv sont orientés positivement, alors pqr est lui aussi orienté positivement.*

Le calcul de l'orientation d'un triangle pqr est un exemple de prédicat géométrique. Il s'exprime comme le signe de l'expression polynomiale de degré total deux en les coordonnées des points :

$$P_{orient}(p,q,r) = \begin{vmatrix} 1 & 1 & 1 \\ x_p & x_q & x_r \\ y_p & y_q & y_r \end{vmatrix}.$$

Théorème 2. *Soient B_1 et B_2 deux disques ouverts dont les bords se coupent en deux points a et b. Soit p un point appartenant à B_1 mais pas à B_2, et B le disque ouvert circonscrit au triangle pab. Alors $B \subset B_1 \cup B_2$.*

() Option informatique

Le prédicat d'appartenance d'un point s au disque circonscrit à trois autres points p, q, r non alignés s'écrit comme le signe d'une expression polynomiale de degré total quatre :

$$P_{dans\_disque}(p,q,r,s) = \begin{vmatrix} 1 & 1 & 1 & 1 \\ x_p & x_q & x_r & x_s \\ y_p & y_q & y_r & y_s \\ x_p^2+y_p^2 & x_q^2+y_q^2 & x_r^2+y_r^2 & x_s^2+y_s^2 \end{vmatrix}.$$

Les algorithmes géométriques prennent leur décisions en fonction de la valeur de prédicats. En raison des erreurs d'arrondi, l'évaluation des prédicats n'est pas toujours correcte, et les théorèmes géométriques usuels ne sont pas toujours satisfaits. Ceci entraîne des incohérences dans l'exécution des algorithmes, qui utilisent bien entendu ces théorèmes géométriques. Ceci aboutit en pratique à l'échec des programmes.

2. L'arithmétique flottante

2.1. *Représentation des nombres*

Un nombre de type `double` est représenté en machine par une séquence de 64 chiffes binaires (c'est-à-dire valant 0 ou 1) :

$$(-1)^s \times (1/2 + m \times 2^{-53}) \times 2^{e-1024}$$

où [1] s est un chiffre binaire indiquant le signe ; e est un entier, l'exposant, codé sur 11 chiffres binaires ; m est un entier, la mantisse, codée sur 52 chiffres binaires. Ces nombres décrivent non pas **R**, mais un ensemble discret et borné.

Pour simplifier les calculs, on considère également un modèle jouet, dans lequel les nombres sont représentés par deux chiffres décimaux (et non plus binaires ici) significatifs : par exemple, 13 et 4,5 et 0,35 et 370 sont représentés exactement dans ce modèle, mais 2,47 ne l'est pas.

2.2. *Règles d'arrondi*

Les `double` représentant seulement un ensemble discret de nombres, le résultat d'une opération arithmétique $(+, -, \times, \div)$ entre deux `double` n'est pas toujours un nombre représentable. Ce résultat est donc approché, et divers modes d'arrondi sont possibles : au plus proche, par défaut, par excès, ou vers 0.

On note $\oplus, \ominus, \otimes$ les opérations arrondies au plus proche, correspondant à $+, -, \times$ respectivement. En cas d'égalité entre les distances aux deux nombres les plus proches, on choisit le nombre le plus proche dont le dernier chiffre binaire de la mantisse est pair.

Dans notre modèle jouet, si on prend l'arrondi au plus proche, alors $36 \oplus 0,5 = 36$; $37 \oplus 0,5 = 38$; $(35 \oplus 3,3) \oplus 0,4 = 38$, mais $35 \oplus (3,3 \oplus 0,4) = 39$. En particulier, l'ensemble des `double` muni des opérations arithmétiques $\oplus$ et $\otimes$ ne forme pas un anneau.

1. En réalité, la définition du type `double` est plus complète, mais par souci de simplicité, nous nous limiterons à cette définition.

() Option informatique

Le signe d'un déterminant 2×2 dont les entrées sont des `double` exacts est évalué sans inversion de signe (le résultat peut être 0 pour un signe correct positif, mais pas négatif). Mais le signe d'un déterminant 3×3 (ou plus généralement $n \times n, n \geq 3$) peut être éventuellement inversé, même si les entrées sont exactes.

3. Incohérences géométriques

Revenons à l'évaluation du prédicat d'orientation de trois points du plan. Écrivons le polynome P_{orient} défini plus haut sous la forme

$$P_{orient}(p,q,r) = \begin{vmatrix} x_q - x_p & x_r - x_p \\ y_q - y_p & y_r - y_p \end{vmatrix}.$$

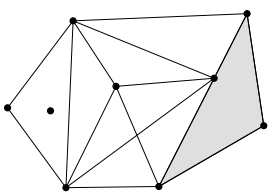
L'évaluation du signe $P_{orient}(p,q,r)$ dans notre modèle jouet, pour les points $p(-94,0), q(400,180), r(92,68)$, ne donne pas le résultat correct. Si on considère ensuite le point $v(-5,34)$, on obtient un contre-exemple au Théorème 1.

Les conséquences de telles erreurs d'évaluation de prédicats sont lourdes pour les algorithmes géométriques. Intéressons nous par exemple au cas des triangulations. Dans tout ce qui suit, $\mathscr{P}$ est un ensemble fini de points du plan $\mathbf{R}^2$.

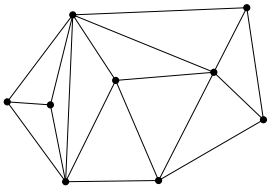
Définition 1 (enveloppe convexe)**.** *L'enveloppe convexe $\mathscr{E}(\mathscr{P})$ de $\mathscr{P}$ est le plus petit convexe de $\mathbf{R}^2$ contenant $\mathscr{P}$.*

Définition 2 (triangulation)**.** *Une triangulation de $\mathscr{P}$ est une partition de $\mathscr{E}(\mathscr{P})$ en triangles dont tous les sommets sont des points de $\mathscr{P}$ et telle que*

- *tout point de $\mathscr{P}$ est sommet d'au moins un triangle,*
- *l'intersection de deux triangles, lorsqu'elle n'est pas vide, est soit un sommet commun, soit une arête commune.*



pas une triangulation



triangulation

L'algorithme simple suivant permet, "sur le papier", de parcourir tous les triangles ayant le sommet v, en tournant autour de v dans le sens positif. t représente un triangle de sommet v. t_0 est le triangle sur lequel le parcours commence.

$t := t_0$
`faire`
    `soient` v,p,q `les sommets de` t`, pris dans le sens positif :` $P_{orient}(v,p,q) > 0$
    t`:= triangle voisin du triangle` vpq`, partageant l'arête` vq `avec` vpq
`tant que` $t \neq t_0$

En pratique, le programme peut ne pas s'exécuter correctement si le prédicat d'orientation n'est pas évalué correctement.

() Option informatique

4. Évaluation certifiée

Certaines bibliothèques logicielles fournissent une arithmétique exacte, c'est-à-dire que le résultat des opérations est bien le résultat mathématiquement correct. L'arithmétique exacte est en général trop lente pour qu'il soit raisonnable de bâtir des logiciels reposant entièrement sur ces bibliothèques.

Les prédicats dont il a été question plus haut peuvent fournir une réponse correcte et garantie, sans que le recours à des calculs arithmétiques exacts soit toujours nécessaire. On parle alors de d'évaluation certifiée des prédicats.

Si le point r est très loin de la droite pq, la valeur de $P_{orient}(p,q,r)$ est grande, et le calcul de son signe en `double` donnera le résultat correct. Mais si r est presque aligné avec pq, alors la valeur de $P_{orient}(p,q,r)$ est très proche de 0, et son signe peut ne pas être conservé lors de l'arrondi. Dans une telle configuration, difficile, on aura recours à du calcul arithmétique exact.

Si les points sont assez bien distribués dans le plan, on peut penser que le recours au calcul arithmétique exact sera rare, et donc que le l'évaluation certifiée sera très rapide. On combine ainsi la rapidité du calcul en `double` avec les garanties du calcul arithmétique exact.

Seul le calcul arithmétique exact peut garantir qu'une expression polynomiale est exactement nulle.

On réutilise les notations $\oplus, \ominus, \otimes$ introduites à la section 2.2 pour les opérations sur les `double`, qui fournissent un résultat arrondi. $+, -, \times$ représentent toujours les opérations (exactes) sur les réels. Pour deux nombres x et y, $x \oplus y$ est une approximation de $x + y$:

$$x \oplus y = x + y + \varepsilon_{x+y},$$

où ε_{x+y} est l'erreur d'arrondi commise. La mantisse de $x + y$ a été arrondie au plus proche à 53 chiffres binaires :

$$|\varepsilon_{x+y}| \leq |x + y| \cdot 2^{-53}.$$

Le signe de $x + y$ est alors évalué de manière certifiée de la façon suivante :

— Si

$$|x \oplus y| > |\varepsilon_{x+y}|$$

alors le signe de $x \oplus y$ est le même que le signe de $x + y$.

— Sinon, on évalue $x + y$ grâce à du calcul arithmétique exact.

De même,

$$\begin{aligned} x \ominus y &= x - y + \varepsilon_{x-y}, \quad \text{avec } |\varepsilon_{x-y}| \leq |x - y| \cdot 2^{-53}, \text{ et} \\ x \otimes y &= x \times y + \varepsilon_{x\times y}, \quad \text{avec } |\varepsilon_{x\times y}| \leq |x \times y| \cdot 2^{-53}. \end{aligned}$$

L'erreur commise lors de l'évaluation d'une expression polynomiale peut ensuite être majorée par propagation des erreurs. Le principe s'étend ainsi à l'évaluation de signes de déterminants, et permet donc une évaluation certifiée des prédicats géométriques évoqués plus haut.

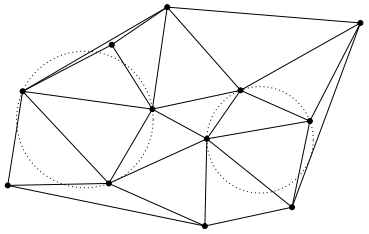
5. Cas dégénérés

L'etude des sections précédentes permet d'admettre ci qu'il est possible d'évaluer de manière garantie les prédicats géométriques évoqués dans la section 1. Le cas où un prédicat

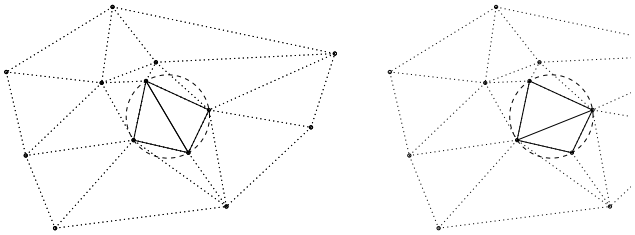
est nul correspond à une configuration dégénérée : trois points aignés ou quatre points cocycliques.

Parmi les triangulations définies plus haut, on admettra qu'il en en existe une, définie comme suit et qui est unique si $\mathscr{P}$ ne contient aucun sous-ensemble de quatre points cocycliques.

Définition 3 (triangulation parfaite)**.** *Une triangulation parfaite $\mathscr{D}(\mathscr{P})$ de $\mathscr{P}$ est une triangulation possédant la propriété suivante : le disque circonscrit à chaque triangle de $\mathscr{D}(\mathscr{P})$ ne contient aucun point de $\mathscr{P}$ en son intérieur.*



Triangulation parfaite



cas dégénéré : $\mathscr{D}(\mathscr{P})$ pas unique

Pour calculer $\mathscr{D}(\mathscr{P})$, il est nécessaire de savoir, pour tout triplet de points p,q,r définissant un triangle (c'est-à-dire pour tout triplet de points p,q,r non alignés) et pour tout point s différent de p,q,r, si s est à l'intérieur ou à l'extérieur du disque circonscrit à pqr (les algorithmes efficaces éviteront bien sûr de tester tous les quadruplets de points). Si $\mathscr{P}$ ne contient aucun sous-ensemble de quatre points cocycliques, l'ensemble des triangles dont le disque circonscrit est vide constitue la triangulation parfaite, unique.

En revanche, si $\mathscr{P}$ contient un sous-ensemble de $k \geq 4$ points cocycliques, alors n'importe quelle triangulation du polygone convexe formé par ces k points est parfaite. Par exemple, pour quatre points p,q,r,s cocycliques, c'est-à-dire pour lesquels $P_{dans\_disque}(p,q,r,s)$ est nul, le quadrilatère qu'ils forment peut être triangulé au choix dans $\mathscr{D}(\mathscr{P})$ par l'une ou l'autre de ses deux diagonales. $\mathscr{D}(\mathscr{P})$ n'est pas définie uniquement par la définition ci-dessus. On pourrait prendre l'une des deux diagonales aléatoirement, le choix de $\mathscr{D}(\mathscr{P})$ ne serait alors pas déterministe.

Pour définir $\mathscr{D}(\mathscr{P})$ uniquement, une possibilité est d'utiliser une perturbation symbolique. Pour p,q,r non alignés, et s différent de p,q,r, le déterminant P_{dans_disque} est perturbé de la façon suivante :

$$P^{(\epsilon)}_{dans\_disque}(p,q,r,s) = \begin{vmatrix} 1 & 1 & 1 & 1 \\ x_p & x_q & x_r & x_s \\ y_p & y_q & y_r & y_s \\ x_p^2+y_p^2+\epsilon^{i(p)} & x_q^2+y_q^2+\epsilon^{i(q)} & x_r^2+y_r^2+\epsilon^{i(r)} & x_s^2+y_s^2+\epsilon^{i(s)} \end{vmatrix},$$

où i est une application injective de $\mathscr{P}$ dans $\mathbf{N}$. La perturbation est dite symbolique car ϵ est une variable, abstraite, et non pas une valeur donnée, aussi petite soit elle. Le signe de $P_{dans\_disque}(p,q,r,s)$ est la limite du signe de $P^{(\epsilon)}_{dans\_disque}(p,q,r,s)$ lorsque ϵ tend vers zéro en gardant des valeurs strictement positives.

() Option informatique

Pour p,q,r,s donnés satisfaisant les hypothèses ci-dessus, $P^{(\epsilon)}_{dans\_disque}(p,q,r,s)$ est un polynôme dans $\mathbf{R}[\epsilon]$, dont le terme constant est nul si et seulement si p,q,r,s sont cocycliques. Les coefficients des autres termes sont les expressions polynomiales P_{orient} sur les triplets de points construits à partir de $\{p,q,r,s\}$. Un raisonnement géométrique simple sur les quatre points montre que $P^{(\epsilon)}_{dans\_disque}(p,q,r,s)$ ne peut pas être le polynôme nul de $\mathbf{R}[\epsilon]$. Son signe est alors le signe du premier coefficient non nul, si on les examine dans l'ordre croissant des exposants de ϵ. On décide d'attribuer à $P_{dans\_disque}(p,q,r,s)$ le signe de $P^{(\epsilon)}_{dans\_disque}(p,q,r,s)$.

Grâce à cette méthode de perturbation, aucun quadruplet de points n'est jamais considéré comme cocyclique : l'un des quatre points est toujours considéré comme étant situé à l'intérieur du disque circonscrit aux trois autres.

La triangulation $\mathscr{D}(\mathscr{P})$ est ainsi définie de façon unique, même en présence de configurations dégénérées.

Exercice de programmation :

► *Il vous est demandé de rédiger un programme conforme aux spécifications ci-dessous dans l'un des langages C, Caml ou Java à votre choix. Ce programme devra être accompagné d'un exemple d'exécution permettant d'en vérifier le bon fonctionnement. La clarté et la concision du programme seront des éléments importants d'appréciation pour le jury.*

Cet exercice consiste à observer le comportement des nombres en virgule flottante de la machine. Toutes les variables utilisées seront de type `double` ou `float` selon le langage choisi.

On écrit un programme qui évalue

$$S = \sum_{i=0}^{\infty} \left(1 - \frac{1}{\rho}\right)^i$$

où ρ est une puissance de 2 suffisamment grande, par exemple $\rho = 2^{20}$.

Si calcule naïvement $S_n = \sum_{i=0}^{n} \left(1 - \frac{1}{\rho}\right)^i$ en ajoutant les termes un par un dans l'ordre des puissances croissantes, on remarque quand n est suffisamment grand, $S_n = S_n \oplus \left(1 - \frac{1}{\rho}\right)^{n+1}$ lorsqu'on calcule en `double` (ou `float`).

En évaluant S sous la forme

$$S = \sum_{j=0}^{\infty} \sum_{k=0}^{K-1} \left(1 - \frac{1}{\rho}\right)^{jK+k}$$

ce phénomène se produit pour des rangs plus grands que précedemment, l'approximation de S calculée est donc meilleure. On pourra essayer plusieurs valeurs de K (1,10,100,1000,10000) et comparer. $K = 1$ correspond au calcul naïf de S ci-dessus.

Optionnellement, le programme pourra également calculer une borne sur l'erreur commise sur S.

() Option informatique

Suggestions pour le développement

► *Soulignons qu'il s'agit d'un menu à la carte et que vous pouvez choisir d'étudier certains points, pas tous, pas nécessairement dans l'ordre, et de façon plus ou moins fouillée. Vous pouvez aussi vous poser d'autres questions que celles indiquées plus bas. Il est très vivement souhaité que vos investigations comportent une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats.*

— Préciser l'ensemble des `double` et étudier ses propriétés : plus petit nombre représentable, plus grand nombre représentable, distance entre deux nombres, commutativité, associativité, évaluation du signe d'un déterminant...
— Quelles sont les propriétés des opérations sur les entiers inférieurs à 2^{53}, si on les représente dans des `double` ?
— Montrer la non-validité du théorème 1 dans le modèle jouet sur les points proposés au début de la section 3.
— Expliciter le problème que peut rencontrer l'algorithme présenté à la fin de la section 3 pour tourner autour d'un sommet dans une triangulation.
— En supposant que les valeurs absolues des coordonnées des points sont toutes plus petites qu'une borne B donnée (c'est-à-dire que les points sont contenus dans un carré $[-B,B]\times[-B,B]$), calculer un majorant de l'erreur commise lors du calcul en `double` de l'expression polynomiale $P_{orient}(p,q,r)$ en fonction de B.
— Démontrer les propriétés du polynôme $P^{(\epsilon)}_{dans\_disque}(p,q,r,s)$ de $\mathbf{R}[\epsilon]$ énoncées dans la section 5, pour p,q,r,s donnés.
— Démontrer qu'en utilisant la perturbation symbolique définie à la section 5, la triangulation parfaite est bien définie par la définition 3, et de façon unique, pour un ensemble de k points cocycliques.