Zeitschrift: L'Enseignement Mathématique

Herausgeber: Commission Internationale de l'Enseignement Mathématique

Band: 18 (1972)

Heft: 1: L'ENSEIGNEMENT MATHÉMATIQUE

Artikel: THÉORIE ADDITIVE DES NOMBRES PROBLÈME DE WARING ET THÉORÈME DE HILBERT

Autor: Dress, François

DOI: https://doi.org/10.5169/seals-45368

Nutzungsbedingungen

Die ETH-Bibliothek ist die Anbieterin der digitalisierten Zeitschriften auf E-Periodica. Sie besitzt keine Urheberrechte an den Zeitschriften und ist nicht verantwortlich für deren Inhalte. Die Rechte liegen in der Regel bei den Herausgebern beziehungsweise den externen Rechteinhabern. Das Veröffentlichen von Bildern in Print- und Online-Publikationen sowie auf Social Media-Kanälen oder Webseiten ist nur mit vorheriger Genehmigung der Rechteinhaber erlaubt. Mehr erfahren

Conditions d'utilisation

L'ETH Library est le fournisseur des revues numérisées. Elle ne détient aucun droit d'auteur sur les revues et n'est pas responsable de leur contenu. En règle générale, les droits sont détenus par les éditeurs ou les détenteurs de droits externes. La reproduction d'images dans des publications imprimées ou en ligne ainsi que sur des canaux de médias sociaux ou des sites web n'est autorisée qu'avec l'accord préalable des détenteurs des droits. En savoir plus

Terms of use

The ETH Library is the provider of the digitised journals. It does not own any copyrights to the journals and is not responsible for their content. The rights usually lie with the publishers or the external rights holders. Publishing images in print and online publications, as well as on social media channels or websites, is only permitted with the prior consent of the rights holders. Find out more

Download PDF: 15.07.2025

ETH-Bibliothek Zürich, E-Periodica, https://www.e-periodica.ch

THÉORIE ADDITIVE DES NOMBRES PROBLÈME DE WARING ET THÉORÈME DE HILBERT [1]

par François DRESS

1. INTRODUCTION HISTORIQUE

Depuis Pythagore, de très nombreux mathématiciens se sont intéréssés aux propriétés « magiques » des nombres entiers, en particulier des nombres triangulaires, des nombres carrés, des nombres m-gonaux (les triangulaires sont les entiers de la forme $n(n+1)/2$: nombre des éléments d'un triangle de points de base n, par exemple, les nombres carrés sont les carrés: nombre d'éléments d'un carré de points, etc...).

C'est Bachet qui semble avoir remarqué le premier, en 1621, que tout entier positif pouvait s'écrire comme somme de 4 carrés (certains étant éventuellement nuls), encore qu'il y ait quelques raisons de penser que Diophante connaissait déjà ce résultat empirique. Ensuite Fermat a noté, en 1636, que tout entier positif pouvait s'écrire comme somme de 3 nombres triangulaires, de 4 carrés, de m nombres m-gonaux (on s'apercevra plus tard que ce résultat général n'est pas le meilleur possible).

Ainsi, pour les nombres triangulaires:

$$\begin{aligned}
&\cdots\cdots\cdots\cdots\\
12 &= 10+1+1 = 6+6+0 = 6+3+3\\
13 &= 10+3+0 = 6+6+1\\
14 &= 10+3+1\\
15 &= 15+0+0 = 6+6+3\\
16 &= 15+1+0 = 10+6+0 = 10+3+3\\
&\cdots\cdots\cdots\cdots
\end{aligned}$$

et pour les carrés:

$$\begin{aligned}
&\cdots\cdots\cdots\cdots\\
12 &= 9+1+1+1 = 4+4+4+0\\
13 &= 9+4+0+0 = 4+4+4+1
\end{aligned}$$

[1]) Cet article reprend et complète une partie de la conférence qu'a prononcée F. DRESS aux Journées Arithmétiques de 1971 (24 au 29 mai 1971, à Marseille).

$$14 = 9 + 4 + 1 + 0$$
$$15 = 9 + 4 + 1 + 1$$
$$16 = 16 + 0 + 0 + 0 = 4 + 4 + 4 + 4$$
$$\cdots\cdots\cdots\cdots$$

Fermat avait ajouté, dans la lettre à Mersenne où il indiquait cette découverte, qu'il n'avait pas la place d'en donner la démonstration, mais qu'il y consacrerait un livre entier. Le livre ne fut jamais publié... On peut d'ailleurs douter que Fermat ait été en possession d'une démonstration correcte car des efforts infructueux furent déployés pendant plus d'un siècle, par Euler en particulier, pour tenter de résoudre ce problème. C'est finalement Lagrange, en 1770, qui en donna la première démonstration.

En même temps d'autres théorèmes empiriques, de nature additive également, étaient énoncés. Les deux plus célèbres sont le problème de Goldbach, formulé en 1742 dans une lettre à Euler: tout entier pair est-il somme de 2 nombres premiers ? et le problème de Waring, formulé en 1770 dans un livre (avec addition du « etc... » dans l'édition de 1782!): tout entier positif est-il somme de 9 cubes, de 19 bicarrés, etc... ?

On verra que ces questions sont, malgré la simplicité de leur énoncé, extrêmement ardues et que l'intervalle qui sépare l'énoncé empirique de sa démonstration se compte en dizaines d'années ou plus souvent en siècles (phénomène assez courant en arithmétique!).

2. Notions fondamentales en théorie additive des nombres

En donnant le premier résultat partiel intéressant dans le problème de Goldbach, Schnirelman esquissa, en 1930, un cadre général pour tous les problèmes additifs relatifs à des suites d'entiers.

Etant donné 2 suites croissantes d'entiers strictement positifs $A = \{a_1 < a_2 < \ldots\}$ et $B = \{b_1 < b_2 \ldots\}$, on appelle somme de A et B et on note $A + B$ la suite croissante obtenue en réordonnant l'ensemble $A \cup B \cup \{a_i + b_j \mid a_i \in A, b_j \in B\}$ (on convient parfois que $a_0 = 0 \in A$, $b_0 = 0 \in B$, auquel cas on considère simplement l'ensemble $\{a_i + b_j\}$).

On peut en particulier effectuer les sommes $A + A = 2A$, $A + A + A = 3A, \ldots, A + \ldots + A = hA, \ldots$ On dit alors que la suite A est une base (d'ordre $\leqslant h$) des entiers s'il existe h tel que $hA = \mathbf{N}$ (A est exactement d'ordre h si $(h-1)\,A \subsetneqq \mathbf{N}$). On peut également définir la notion de base relativement à une sous-suite de $\mathbf{N}$ (les entiers pairs dans le problème de Goldbach, par exemple).

Etant donné une suite $A = \{a_k\}$, on définit la fonction $A(n) = \sum_{1 \leq a_k \leq n} 1$ = nombre des a_k compris entre 1 et n. On définit ensuite la densité de Schnirelman de la suite A par:

$$d(A) = \inf_n \frac{A(n)}{n}.$$

Cette définition appelle deux remarques. Primo, on a $1 \in A$ dès que $d(A) > 0$. Secundo, la notion de densité de Schnirelman est très différente de celle, classique, de densité asymptoptique, définie par $\liminf \frac{A(n)}{n}$. En particulier, $d(A) = 1$ équivaut à $A = \mathbf{N}$, tandis que d. asympt. $(A) = 1$ équivaut à « presque tous les entiers appartiennent à A ». Indiquons enfin une notion intermédiaire, « tous les entiers assez grands appartiennent à A », fréquemment utilisée en théorie additive (théorème de Vinogradov sur les entiers impairs sommes de 3 nombres premiers, constantes $G(k)$ du problème de Waring).

3. Théorèmes de Schnirelman et de Mann

La densité de Schnirelman est un outil remarquable pour prouver que certaines suites sont des bases. Il ne faut pas cependant en attendre plus que des résultats d'existence, avec au mieux une majoration délirante de l'ordre (2.10^{10} pour les nombres premiers, nettement pire dans le problème de Waring par la méthode de Linnik et Khintchine, par exemple), en raison de l'influence très pathologique des premiers termes de la suite (et comme « premiers » n'est jamais que le contraire de « à l'infini », cela peut entraîner fort loin... !).

Les principaux théorèmes en la matière sont les suivants:

Théorème (Schnirelman, 1930). $d(A+B) \geqslant d(A) + d(B) - d(A)\,d(B)$.

La démonstration, que nous ne donnerons pas ici, est fort simple et s'appuie, modulo la minoration $\forall n\ [A(n) \geqslant n \,.\, d(A)]$, sur un dénombrement très banal.

Lemme. $d(A) + d(B) \geqslant 1 \Rightarrow A + B = \mathbf{N}$.

Une simple affaire de tiroirs, tout aussi banale.

COROLLAIRE. *Toute suite de densité de Schnirelman strictement positive est une base.*

En effet, l'inégalité de Schnirelman peut s'écrire

$$\big(1-d(A+B)\big) \leqslant \big(1-d(A)\big)\big(1-d(B)\big).$$

Il s'ensuit, en particulier, que $(1 - d(hA)) \leqslant (1 - d(A))^h \leqslant \frac{1}{2}$ si $d(A) > 0$ et $h \geqslant h_0$. Et le lemme précédent entraîne immédiatement que $2h_0A = \mathbf{N}$.

Schnirelman a ainsi prouvé que la suite $P = \{1\} \cup \{\text{nombres premiers}\}$ était une base en démontrant par une méthode de crible que $d(2P) > 0$.

THÉORÈME (Mann, 1942). $d(A+B) \geqslant \min(d(A) + d(B), 1)$.

La démonstration de ce théorème est assez ardue et il faut ajouter qu'il représente en un sens le meilleur résultat possible. Si l'on a par exemple (k étant un entier $\geqslant 2$)

$$A = B = (1, k+1, 2k+1, \ldots),$$

alors

$$A + B = (1, 2, k+1, k+2, 2k+1, 2k+2, \ldots),$$

cependant que l'on a $d(A) = d(B) = \dfrac{1}{k}$ et $d(A+B) = \dfrac{2}{k} = d(A) + d(B)$.

4. GÉNÉRALITÉS SUR LE PROBLÈME DE WARING

Nous en avons déjà vu l'énoncé: pour $k = 2, 3, 4, \ldots$, la suite des puissances k-ièmes est-elle une base ? La réponse est affirmative, comme nous le verrons, et l'on désigne traditionnellement par $g(k)$ l'ordre de cette base. Pour les premières valeurs de k, l'évidence empirique conduit à conjecturer les valeurs suivantes:

$$g(2) = 4, \quad g(3) = 9, \quad g(4) = 19.$$

En exceptant le théorème des 4 carrés de Lagrange, la première démonstration d'existence, dans le cas particulier des bicarrés, est due à Liouville, en 1859, avec la majoration $g(4) \leqslant 53$. Sa démonstration,

que nous donnerons au paragraphe 6, utilise conjointement le théorème des 4 carrés et une identité algébrique.

Puis on s'apercevra assez vite que, moyennant une identité algébrique « convenable » (que l'on découvrira effectivement pour les petites valeurs de k), l'existence de $g(k)$ entraîne celle de $g(2k)$, avec une majoration du style $g(2k) \leqslant a_k\, g(k) + b_k$.

Par contre, les valeurs impaires de k posent des problèmes délicats, car les identités ont alors une tendance fâcheuse à fournir des sommes de puissances k-ièmes dont certaines sont négatives ! Enfin Maillet réussit, en 1895, à démontrer l'existence dans le cas des cubes, et donne la majoration $g(3) \leqslant 21$.

De nouveaux cas sont résolus, les majorations sont petit à petit améliorées, puis Hilbert, en 1909, démontre le théorème général d'existence de $g(k)$. Le théorème de Hilbert ne donnant aucune majoration explicite, la course aux majorations continue donc, et continue toujours... mais surtout pour une autre constante dont nous allons parler maintenant.

En 1909 également, Wieferich prouve que $g(3) = 9$, cependant que Landau montre qu'il existe N_0 tel que tout entier supérieur à N_0 soit somme d'au plus 8 cubes. En d'autres termes, la valeur de $g(3)$ dépend d'un nombre fini d'entiers — en fait 23 et 239 — qui sont seuls à exiger 9 cubes, et ne réflète nullement les propriétés « à l'infini ». Plus généralement on est amené à définir $G(k)$ le minimum de p, tel que tout entier suffisamment grand soit somme d'au plus p puissances k-ièmes. On trivialement $G(k) \leqslant g(k)$ et il est en outre bien clair que l'existence de l'une quelconque des deux constantes entraîne celle de l'autre. La disproportion entre les deux constantes est énorme: on sait actuellement que $g(k)$ est équivalent à 2^k, tandis que l'on dispose pour $G(k)$ de majorations de l'ordre de $k \operatorname{Log} k$.

Nous nous occuperons dans la suite de cet article de méthodes élémentaires tournant autour des identités algébriques. Il importe néanmoins de signaler que Hardy et Littlewood en 1920, puis Vinogradov en 1924, ont introduit des méthodes analytiques extrêmement puissantes qui permettent d'obtenir des résultats numériques remarquables. Et notons enfin que Linnik a donné, en 1943, une démonstration du théorème de Hilbert par la méthode de Schnirelman; mais les majorations explicites qu'il obtient sont catastrophiques.

5. Sommes de carrés

Avant d'exposer, dans le cas des bicarrés et dans celui des cubes, le principe des méthodes élémentaires, puis de donner une démonstration simple du théorème de Hilbert, nous tenons à rappeler très brièvement quelques résultats essentiels sur les sommes de carrés.

Le résultat essentiel sur les sommes de 2 carrés, énoncé en 1625 par Girard puis un peu plus tard par Fermat, démontré en 1749 par Euler, est le suivant: N est somme de 2 carrés si et seulement si les nombres premiers de la forme $4n + 3$ qui figurent dans sa décomposition y figurent avec un exposant pair — en autres termes, si N est le produit d'un carré par un entier composé uniquement avec des facteurs premiers 2 ou de la forme $4n + 1$. On notera que les sommes de 2 carrés forment une suite de densité asymptotique nulle (la densité jusqu'à x est équivalente à $1/\mathrm{Log}\, x$).

Par souci d'ordre, énonçons le théorème de Lagrange: tout N (positif ou nul) est somme de 4 carrés (positifs ou nuls). On en trouvera un peu partout des démonstrations très nombreuses et plus ou moins variées...

Le problème de déterminer les sommes de 3 carrés est beaucoup plus difficile que les précédents. Il est en liaison très étroite avec la théorie des formes quadratiques binaires. Legendre, en 1798, puis Gauss, en 1801, ont établi le résultat suivant: N est somme de 3 carrés si et seulement si il n'est pas de la forme $4^k(8n+7)$.

On remarquera pour terminer que le résultat sur les nombres triangulaires est un corollaire immédiat du théorème des 3 carrés. En effet

$$m = \frac{a(a+1)}{2} + \frac{b(b+1)}{2} + \frac{c(c+1)}{2}$$

$$\Leftrightarrow \# \; 8m + 3 = (2a+1)^2 + (2b+1)^2 + (2c+1)^2$$

et les entiers de la forme $8m + 3$ sont bien sommes de 3 carrés (impairs, pour de banales raisons de congruences modulo 8).

6. Sommes de bicarrés

Nous allons rapporter la méthode de Liouville, puis nous indiquerons ensuite très sommairement comment la majoration obtenue peut être améliorée en conservant les méthodes élémentaires.

On considère l'identité

$$6(a_1{}^2+a_2{}^2+a_3{}^2+a_4{}^2)^2 = \sum_{i<j} [(a_i+a_j)^4 + (a_i-a_j)^4] = B_{12},$$

en désignant par B_q un entier qui est somme de q bicarrés (en fait, l'identité que nous donnons ici est due à Lucas, mais celle qu'utilisait Liouville lui est équivalente). Comme tout entier est somme de 4 carrés, on obtient ainsi

$$6a^2 = B_{12}, \text{ pout tout } a,$$

puis

$$6m = 6(a^2+b^2+c^2+d^2) = B_{48}, \text{ pour tout } m,$$

et enfin, comme tout entier n est de la forme $6m + h \,.\, 1^4$ (avec $h = 0, 1, ..., 5$),

$$n = B_{53}, \text{ pour tout } n, \text{ i.e. } g(4) \leqslant 53,$$

ce qui est très exactement le résultat donné par Liouville.

Pour améliorer cette majoration, on a utilisé essentiellement, outre diverses petites astuces, deux remarques:

— certains entiers a peuvent s'écrire $a = a_1{}^2 + a_2{}^2 + 2a_3{}^2$, de sorte qu'alors $6a^2 = B_{11}$;

— certains entiers m peuvent s'écrire comme sommes de 3 carrés.

Mais on peut faire un peu mieux. L'idée que nous exposerons pour les sommes de cubes et surtout pour la démonstration générale du théorème de Hilbert nous a permis de « partir » de sommes de 2 carrés au lieu de sommes de 3, et nous avons ainsi obtenu $g(4) \leqslant 30$. C'est la meilleure majoration actuellement connue, mais il est vraisemblable que ce n'est pas pour bien longtemps, des travaux sont en cours où les méthodes analytiques reprendraient leurs droits...

Signalons pour terminer ce paragraphe la majoration $g(4) \geqslant 19$. Des tables numériques extrêmement étendues ont été calculées qui laissent à penser qu'il n'y a que 7 entiers qui nécessitent 19 bicarrés: 79, 159, 239, 319, 399, 479 et 559.

7. Sommes de cubes

L'identité « historique » est

$$6x(x^2+a_1{}^2+a_2{}^2+a_3{}^2) = \sum_i [(x+a_i)^3 + (x-a_i)^3] = C_6,$$

en désignant par C_q un entier qui est somme de q cubes. Cette identité permet de montrer qu'un nombre de la forme $6x\,(x^2+m)$ est C_6 sous deux conditions. La première, mineure, est que m soit une somme de 3 carrés; la seconde, beaucoup plus gênante mais essentielle pour que les cubes soient positifs, est que m ne soit pas trop grand (on devra imposer a priori $0 \leqslant m \leqslant x^2$). Toute la difficulté est alors de « raccrocher » un entier quelconque à un nombre $6x\,(x^2+m)$ convenable.

Nous allons exposer une manière de le faire, qui utilise des « cubes arbitraires » et une seconde identité algébrique, dans l'esprit de notre démonstration du théorème de Hilbert. Mais nous devrons tout d'abord modifier la première identité en lui ajoutant 2 carrés:

$$10x^3 + 6x\,(a_1{}^2+a_2{}^2+a_3{}^2+a_4{}^2+a_5{}^2) = \sum_i \left[(x+a_i)^3 + (x-a_i)^3\right],$$

soit, de manière abrégée,

$$10x^3 + 6xm = C_{10}$$

Moyennant certaines conditions de congruences modulo 8 et certaines majorations sur m (que nous verrons plus tard), les 2 carrés $a_4{}^2$ et $a_5{}^2$ peuvent être choisis arbitrairement, ce qui revient encore à dire que les deux cubes $(x-a_4)^3$ et $(x-a_5)^3$ peuvent être choisis arbitrairement. Nous les prendrons alors égaux tous deux à t^3 (nous verrons plus tard également comment choisir t) et nous utiliserons l'identité

$$2t^3 = (t+1)^3 + (t-1)^3 - 6t\,.$$

Et, en négligeant les diverses conditions qui devront être satisfaites, on obtient le résultat brut que non seulement les nombres $10x^3 + 6xm$, mais aussi ceux $10x^3 + 6xm + 6t$ sont C_{10}. Pour simplifier la présentation, nous commencerons par réécrire l'identité initiale en tenant compte de la seconde:

$$\begin{aligned} &10x^3 + 6xm + 6t \\ &\quad = 10x^3 + 6x\left(a_1{}^2+a_2{}^2+a_3{}^2+2\,(x-t)^2\right) + 6t = C_{10}\,. \end{aligned}$$

Sous les deux conditions

$$\begin{cases} m \equiv 3 \text{ ou } 5 \,(\text{mod. } 8) \\ 2x^2 \leqslant m \leqslant \dfrac{17}{8}\,x^2 \end{cases}$$

nous pourrons choisir t quelconque dans l'intervalle $[1, \frac{1}{4}x]$. Cette limitation (qui ne peut guère être relâchée) ne permet pas de « couvrir la plage » comprise entre deux valeurs acceptables consécutives de m et nous devrons introduire un cube supplémentaire, qui permettra par la même occasion de régler les questions de congruences modulo 6.

En posant $m = 2x^2 + n$, on écrira un entier N comme somme de 11 cubes en suivant le processus suivant:

— on prend pour x le plus grand entier tel que

$$22x^3 + 6(5x) + 125 \leqslant N$$

(le terme $5x$ provient des conditions de congruences sur m: dans certains cas, on ne pourra pas choisir n inférieur à 5; le terme 125 provient de la condition de congruence modulo 6 — qu'on verra un peu plus loin — sur le 11e cube: dans certains cas, on ne pourra pas le choisir inférieur à $5^3 = 125$);

— on prend pour n le plus grand entier acceptable modulo 8 tel que

$$22x^3 + 6xn + 125 \leqslant N,$$

et on a alors un reste défini par

$$N = 22x^3 + 6xn + r;$$

— on choisit enfin h le plus grand entier congru à r modulo 6 et tel que

$$h^3 \leqslant r.$$

Comme $h^3 \equiv h$ (mod. 6), on a donc $r = h^3 + 6t$, avec les majorations

$$r \leqslant 6(6x) + 125$$

$$6t \leqslant 3r^{2/3}$$

et l'on constate que l'on obtient une valeur admissible pour t (i.e. vérifiant $t \leqslant \frac{1}{4}x$) dès que $x \geqslant 10\,375$, ce qui sera le cas dès que $N \geqslant 2{,}4569.10^{13} > 22\,(10\,375)^3$. On remarquera qu'à cette valeur, il y a belle lurette que les intervalles $[22x^3, 22\frac{3}{4}x^3]$ se recouvrent (ces intervalles correspondent à la condition d'encadrement donnée plus haut pour m).

Tout entier à partir de $2{,}4569.10^{13}$ étant donc somme de 11 cubes (positifs) il reste, pour finir de prouver la majoration

$$g(3) \leqslant 11\,,$$

à montrer que tous les entiers inférieurs à cette limite sont également C_{11}. La vérification numérique se fait par une méthode de descente très simple, en ôtant de chaque entier le plus grand cube inférieur ou égal (avec une légère modification pour les deux dernières étapes): il suffit que tout entier inférieur à $2{,}5355.10^9$ soit C_{10}, que tout entier inférieur à $5{,}578.10^6$ soit C_9, que tout entier compris entre 240 et 94 758 soit C_8, et enfin que tout entier compris entre 455 et 6 665 soit C_7. Cette dernière condition résulte des tables connues (jusqu'à 40 000, 239 est le plus grand nombre qui nécessite 9 cubes, 454 le plus grand qui en nécessite 8, tous ceux au-delà étant C_7).

8. Intermède: le problème facile de Waring

Alias « the easier problem of Waring ».

Ce problème nous sera utile non pas pour son énoncé et ses résultats mais pour les identités qui interviennent dans sa résolution. Il s'agit d'écrire tout entier sous la forme $N = \pm\, y_1{}^k \pm y_2{}^k \pm \ldots \pm y_s{}^k$ (les y_j étant des entiers positifs, mais cela n'a guère d'importance) et d'établir l'existence d'une constante $v(k)$ telle que l'on puisse toujours prendre $s \leqslant v(k)$.

On utilise des identités valables pour les entiers dans certaines progressions arithmétiques. Ainsi pour les cubes:

$$6n = (n+1)^3 + (n-1)^3 - 2n^3$$

$$6n + 3 = (2n-5)^3 + n^3 - (2n-4)^3 - (n-4)^3$$

et pour les bicarrés:

$$4\,080n = (2n-1)^4 + (n+8)^4 - (2n+1)^4 - (n-8)^4$$

L'existence de $v(k)$ dans le cas général résulte de l'identité

$$n^k - C_{k-1}^1 (n-1)^k + C_{k-1}^2 (n-2)^k - \ldots + (-1)^{k-1} (n-k+1)^k = k\,!\,n + \beta$$

(β entier indépendant de n)

(la démonstration est immédiate: calcul de la $(k-1)$-ième différence finie du polynôme x^k).

Nous retiendrons cette identité sous la forme suivante:

LEMME. *Pour tout entier positif k il existe des entiers positifs $R = R(k)$, $S = S(k)$, α, $a_1, \ldots, a_R$, $c_1, \ldots, c_S$, et des entiers quelconques β, $b_1, \ldots, b_R$, $d_1, \ldots, d_S$, tels que l'on ait l'identité*

$$\sum_{i=1}^{R} (a_i n + b_i)^{2k} = \sum_{j=1}^{S} (c_j n + d_j)^{2k} - (\alpha n + \beta) .$$

9. THÉORÈME DE HILBERT. L'IDENTITÉ FONDAMENTALE

La méthode de Hilbert pour démontrer l'existence de $g(n)$ est fondée sur la donnée, pour tout k, d'une identité de même forme que celle que nous avons vue pour les sommes de bicarrés:

$$M(x_1^2 + x_2^2 + \ldots + x_5^2)^2 = \sum_{i=1}^{N} m_i (a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{i5}x_5)^{2k} .$$

Mais cette identité permet uniquement de démontrer l'existence de $g(2k)$ en supposant établie celle de $g(k)$. Hilbert a donc dû, pour montrer l'existence de $g(n)$ pour les valeurs impaires de n, imaginer un raisonnement par récurrence que nous trouvons personnellement assez compliqué. Après Hilbert, de nombreux mathématiciens se sont efforcés de simplifier sa démonstration mais les améliorations ont pratiquement toutes porté sur l'établissement de l'identité fondamentale.

Nous nous proposons ici de supprimer la seconde partie de la démonstration de Hilbert et de prouver, sans aucune récurrence, que $g(n)$ existe pour tout n pair (d'où il s'ensuit trivialement que $g(n)$ existe aussi pour tout n impair). Outre l'utilisation déjà annoncée des identités du problème facile de Waring, nous aurons besoin au préalable de préciser quelque peu l'identité fondamentale de Hilbert.

LEMME. *Pour tout entier positif k il existe des entiers positifs M, $N = (2k+1) \ldots (2k+5)/24$, $m_1, \ldots, m_{N-1}$, m_N, avec M et m_N strictement positifs, et des entiers $a_{11}, \ldots, a_{15}$, $a_{21}, \ldots, a_{N5}$, tels que l'on ait l'identité*

$$M(x_1^2 + \ldots + x_5^2)^k = \sum_{i=1}^{N-1} m_i (a_{i1}x_1 + \ldots + a_{i5}x_5)^{2k} + m_N x_5^{2k} .$$

(L'innovation par rapport à l'identité de Hilbert est: m_N est strictement positif).

COROLLAIRE. *Pour tout entier positif k il existe des entiers positifs $M = M(k)$ et $Q = Q(k)$ tels que, pour tout entier l et tout entier $x \leqslant \sqrt{l}$ on ait*

$$Ml^k = x^{2k} + \sum_1^Q u_h{}^{2k}$$

$$(\textit{avec les } u_h \in \mathbf{Z})$$

Ce corollaire est la généralisation du résultat que nous avions utilisé pour disposer d'un « cube arbitraire ».

Pour la démonstration de notre identité, nous utiliserons la méthode de Schmidt, reprise par Ellison, qui s'appuie sur les propriétés des ensembles convexes (dans un espace vectoriel réel). Nous rappelons tout d'abord — sans démonstration — les définitions et résultats dont nous aurons besoin:

— étant donné un ensemble $S \subset \mathbf{R}^N$, on appelle enveloppe convexe (ou clôture convexe) de S et on note $h(S)$ le plus petit ensemble convexe qui contient S (i.e. l'intersection de tous les ensembles convexes qui contiennent S);

— étant donné un ensemble $S \subset \mathbf{R}^N$, tout vecteur $V \in h(S)$ peut s'écrire sous la forme $V = \sum_{i=1}^N m_i s_i$, avec $s_i \in S$, $m_i \in \mathbf{R}$, $m_i \geqslant 0$ et $\sum_{i=1}^N m_i = 1$. De plus, si tous les vecteurs de S sont à coordonnées rationnelles, et si V est également à coordonnées rationnelles, les m_i peuvent être choisis rationnels;

— le barycentre d'une masse continûment répartie dans un ensemble S borné se trouve toujours à l'intérieur de $h(S)$ (cet intérieur étant « pris » dans la plus petite variété affine support de $h(S)$ — munie de la topologie ordinaire).

Nous allons maintenant donner la démonstration de l'identité de Hilbert telle qu'elle est exposée par Ellison. Il nous suffira ensuite d'un petit complément pour obtenir la précision supplémentaire: m_N est strictement positif.

L'ensemble des formes homogènes de degré $2k$ en 5 variables et à coefficients réels constitue un espace vectoriel sur $\mathbf{R}$ de dimension $N = (2k+1) \dots (2k+5)/24$ (N est le nombre de termes de la forme générale de degré $2k$ en $x_1, \dots, x_5$). On considère alors dans cet espace $\mathbf{R}^N$ l'ensemble S de toutes les formes $(a_1x_1 + \dots + a_5x_5)^{2k}$, les a_i appartenant

à $\mathbf{Q}$, ainsi que son enveloppe convexe $h(S)$. Puis on considère les sous-ensembles T et T' constitués par les formes $(a_1x_1 + \ldots + a_5x_5)^{2k}$ vérifiant $a_1{}^2 + \ldots + a_5{}^2 \leqslant 1$, les a_i appartenant respectivement à $\mathbf{Q}$ et à $\mathbf{R}$. On a l'inclusion $h(T) \subset h(S)$, cependant que $h(T)$ et $h(T')$ ont même intérieur.

On étudie ensuite l'intégrale

$$\int_{\mathscr{S}} (a_1x_1 + \ldots + a_5x_5)^{2k}\, da_1 \ldots da_5 \;/\; \int_{\mathscr{S}} da_1 \ldots da_5$$

$$(\mathscr{S} \text{ étant l'hypersphère } a_1{}^2 + \ldots + a_5{}^2 \leqslant 1)$$

et on établit, à l'aide d'un banal changement de variables, qu'elle est égale à $c\,(x_1{}^2 + \ldots + x_5{}^2)^k$, avec

$$c = \int_{\mathscr{S}} t_1{}^{2k}\, dt_1 \ldots dt_5 \;/\; \int_{\mathscr{S}} dt_1 \ldots dt_5) > 0\,.$$

La forme $f = c\,(x_1{}^2 + \ldots + x_5{}^2)^k$ se trouve par conséquent à l'intérieur de $h(T')$, donc de $h(T)$, donc de $h(S)$, ainsi d'ailleurs que toutes les formes λf (λ réel $\in [0, 1]$). Et on peut conclure en choisissant λ tel que $\lambda c \in \mathbf{Q}$.

Mais nous pouvons faire un peu mieux : en effet la forme f est à l'intérieur de $h(S)$ tandis que la forme $g = x_5{}^2$ est dans S donc dans la variété affine support de S; ce qui permet d'en déduire qu'il existe μ_0 réel > 0 tel que, pour tout μ réel $\in [0, \mu_0]$, la forme $f - \mu g$ se trouve dans $h(S)$, ainsi d'ailleurs que toutes les formes $\lambda f - \lambda\mu g$ (λ réel $\in [0, 1]$). Nous choisissons alors λ et μ tels que λc et $\lambda\mu$ soient rationnels et, en utilisant les résultats rappelés sur les ensembles convexes et les vecteurs à coordonnées rationnelles, nous en déduisons l'identité cherchée.

10. Théorème de Hilbert. Fin de la démonstration

Théorème. *Pour tout entier positif k il existe des entiers positifs $A = A(k)$ et $T = T(k)$ tels que tout intervalle $[m - A, m]$ contienne un nombre qui soit somme de T puissances $2k$-ièmes.*

Corollaire (théorème de Hilbert). *Pour tout entier positif n, $g(n)$ est fini.*

On pourra remarquer que la recherche d'une majoration explicite de $g(2k)$ en utilisant notre démonstration dépend essentiellement des constantes M et m_i qui interviennent dans l'identité fondamentale, les autres constantes (celles que l'on trouve dans l'identité relative au problème facile de Waring comme celles qui interviendront dans la suite de la démonstration) étant aisément estimables ou majorables.

Nous allons donc montrer que, pour tout entier m, il existe $r < A$ tel que $m - r$ soit somme de T puissances $2k$-ièmes (rappelons que les diverses constantes que nous allons rencontrer: R, M, ... ont déjà été définies, soit au paragraphe 8 soit au paragraphe 9, et qu'elles dépendent toutes de k, et de k seulement). Si l^k est la plus grande puissance k-ième inférieure ou égale à $\dfrac{m}{RM}$, nous pouvons tout d'abord écrire

$$m = R.Ml^k + r_1 , \qquad \text{avec} \qquad \frac{1}{2}\left(\frac{m}{RM}\right)^{1/k} \leqslant l \leqslant \left(\frac{m}{RM}\right)^{1/k}$$

$$\text{et} \qquad 0 \leqslant r_1 \leqslant kRM\left(\frac{m}{RM}\right)^{(k-1)/k}$$

(la constante $\dfrac{1}{2}$ n'est pas essentielle; en toute rigueur, l'inégalité où elle figure n'est vérifiée que pour $m \geqslant m_0\,(k)$, mais il nous semble inutile d'alourdir notre démonstration avec de tels détails qui ne peuvent avoir d'importance que dans la recherche éventuelle d'une majoration explicite de $g\,(2k)$).

Nous pouvons maintenant, en utilisant le corollaire du lemme qui énonce l'identité fondamentale, écrire

$$m = \sum_{i=1}^{R} x_i^{2k} + \sum_{h=1}^{QR} u_h^{2k} + r_1 ,$$

les x_i étant des entiers arbitraires inférieurs ou égaux à $\dfrac{1}{\sqrt{2}}\left(\dfrac{m}{RM}\right)^{1/2k}$.

Il nous faut alors employer le lemme sur l'identité relative au problème facile de Waring. Nous supposerons que l'on a $|\beta| < |\alpha|$, ce qui est toujours possible par une translation sur n. Définissons

$$a = \max_i a_i , \quad b = \max_i |b_i/a_i|$$

de sorte que pour tout $n \leqslant \dfrac{1}{a}\left\{\dfrac{1}{\sqrt{2}}\left(\dfrac{m}{RM}\right)^{1/2k}\right\} - b$, nous pouvons toujours poser, pour tout i, $x_i = a_i n + b_i$. Ce qui permet d'écrire

$$m = \sum_{j=1}^{S} y_j^{2k} + \sum_{h=1}^{QR} u_h^{2k} + r_1 - (\alpha n + \beta) ,$$

la condition de majoration sur n pouvant s'écrire (ici encore pour m assez grand):

$$n \leqslant Cm^{1/2k}.$$

Il est clair qu'en général r_1 est trop grand pour pouvoir être « presque annulé » par le terme $-(\alpha n+\beta)$, mais nous pouvons résoudre cette difficulté de la même manière que nous l'avons fait pour les sommes de cubes. On extraira donc la plus grande puissance $2k$-ième inférieure ou égale à r_1, puis on répétera ce processus:

$$r_1 = z_1^{2k} + r_2$$

$$\text{avec} \quad 0 \leqslant r_2 \leqslant k\left\{kRM\left(\frac{m}{RM}\right)^{(k-1)/k}\right\}^{(k-1)/k} = c_2 m^{\gamma^2}\left(\gamma = \frac{k-1}{k}\right)$$

$$r_2 = z_2^{2k} + r_3 \quad \text{avec} \quad 0 \leqslant r_3 \leqslant c_3 m^{\gamma^3}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$r_{t-1} = z_{t-1}^{2k} + r_t \quad \text{avec} \quad 0 \leqslant r_t \leqslant c_t m^{\gamma^t}.$$

En prenant t tel que γ^t soit supérieur à $1/2k$, il sera alors possible (toujours pour m assez grand) de choisir n de telle façon que le reste final $r = r_t - (\alpha n+\beta)$ vérifie

$$r < \alpha,$$

et nous avons ainsi obtenu le résultat cherché: pour $A = \alpha$, il existe toujours $r < A$ tel que $m - r$ soit somme de $T = S + QR + t - 1$ puissances $2k$-ièmes.

Appendice

Tableau des valeurs ou des meilleurs encadrements de $G(k)$ et de $g(k)$ actuellement connus pour les petites valeurs de k:

k	2	3	4	5	6	7	8	9	10
$G(k)$	4	4-7	16	6-23	9-36	8-52	32-73	13-99	12-122
$g(k)$	4	9	19-30	37	73	143	279	548	1079

BIBLIOGRAPHIE DIDACTIQUE

ELLISON. Waring's problem. *Amer. Math. Monthly*, vol. 78, 1971, pp. 10-36.
HALBERSTAM and ROTH. *Sequences*, vol. 1. Oxford, at the Clarendon Press, 1966.
HARDY and WRIGHT. *An introduction to the theory of numbers*. Oxford, at the Clarendon Press, 1960 (4th edition).
OSTMANN. *Additive Zahlentheorie*. Berlin, Göttingen, Heidelber, Springer-Verlag, 1956.

(Reçu le 20 décembre 1971)

François Dress
U.E.R. de Mathématiques et d'Informatique
Université Bordeaux - I
33 - Talence (France)