G. Poitou

Sur l'approximation simultanée de deux nombres réels

Séminaire Dubreil. Algèbre et théorie des nombres, tome 9 (1955-1956), exp. nº 13, p. 1-6
<http://www.numdam.org/item?id=SD_1955-1956__9__A8_0>

Faculté des Sciences de Paris
-:-:-:-
Séminaire P. DUBREIL et C. PISOT
(ALGÈBRE et THÉORIE DES NOMBRES)
Année 1955/56
-:-:-:-

13 février 1956

Exposé n° 13

SUR L'APPROXIMATION SIMULTANÉE DE DEUX NOMBRES RÉELS

par G. POITOU.

Un titre plus précis pour cet exposé serait "Sur l'approximation simultanée de deux nombres réels par des fractions de même dénominateur." En réalité, il s'agit aussi du problème de l'approximation simultanée de n nombres réels, mais le problème pour $n \geqslant 3$ est à peine dégrossi. L'exposé sera en grande partie historique, car il s'agit d'un problème déjà ancien et qui n'a toujours pas reçu de solution satisfaisante, mais qui a progressé lentement au cours des décades.

De résultats de Dirichlet, Hermite, etc..., généralisés par Kronecker (1884) résulte que, quels que soient les nombres réels $\alpha_1 , \alpha_2 , \dots , \alpha_n$, il existe une infinité de systèmes d'entiers $z , x_1 , x_2 , \dots , x_n$ tels que

$$|z| |x_i - \alpha_i z|^n < 1$$

La borne inférieure des constantes par lesquelles on peut remplacer 1 au second membre de cette inégalité sera notée $\frac{1}{C_n}$ (avec $C_n \geqslant 1$). En un sens restreint, le problème de l'approximation simultanée de n nombres réels consiste dans la détermination de cette constante C_n ; c'est à cet aspect, particulier mais crucial, que nous nous limiterons ici. On sait que ce problème est résolu depuis longtemps pour C_1 qui vaut $\sqrt{5}$ (approximation d'un seul nombre réel par une fraction).

Un important progrès est dû à Minkowski qui, dans "Geometrie der Zahlen", prouve que $C_n \geqslant (1 + \frac{1}{n})^n$ en considérant dans R^{n+1} le corps connexe défini par les inégalités

$$|x_i - \alpha_i z| + |\frac{z}{t}| < 1 .$$

En particulier, on a $C_2 \geqslant \frac{9}{4} = 2,25$.

Le même Minkowski, dans son mémoire de Göttingen (1904), utilise son résultat sur le déterminant critique d'un octaèdre (ou, ce qui revient au même sur le minimum de la somme des valeurs absolues de trois formes linéaires ternaires) pour porter la minoration de C_2 à $C_2 \geqslant \frac{19}{8} = 2,37 \dots$

Ce résultat n'a pas été amélioré pendant plus de quarante ans, les méthodes de Blichfeldt (Transactions Am. Math. Soc., 1914) n'apportant qu'une faible amélioration à la minoration de C_n, qui devient

$$C_n \geqslant (1 + \frac{1}{n})^n \left[1 + (\frac{n-1}{n+1})^{n+3} \right]$$

Par contre, on a obtenu dans les années vingt des majorations de C_n. Après un premier résultat assez grossier de Perron (Math. Annalen 1921), Furtwängler donne en 1927-28 aux Math. Annalen la majoration $C_n \leqslant \sqrt{|D_{n+1}|}$, où D_{n+1} est le discriminant d'un corps algébrique réel de degré $n+1$, minimum en valeur absolue. En particulier, on a $C_2 \leqslant \sqrt{23} = 4,79 \ldots$ (ce qui avait été annoncé par Blichfeldt en 1922 dans le Bulletin Am. Math. Soc.), $C_3 \leqslant \sqrt{275} \ldots$ Les valeurs de D_k pour $k \geqslant 5$ sont inconnues, on doit se contenter d'une majoration $D_k < 2^{k-1} k^k$ tirée de la considération de l'équation $x^k - 2 = 0$; remarquons que les travaux de Mr. Pisot sur les nombres entiers algébriques dont les conjugués sont inférieurs à 1 en module permettent de remplacer l'ordre de grandeur $(2k)^k$ par $(\frac{1+\sqrt{5}}{2} k)^k$. On notera que le discriminant minimum D_2 d'un corps quadratique réel est 5, ce qui montre que l'inégalité de Furtwängler s'écrit, pour $x = 1$, $C_1 \leqslant \sqrt{5}$, et est ainsi la meilleure possible, puisqu'on sait par ailleurs que $C_1 = \sqrt{5}$.

On pouvait se demander si l'inégalité de Furtwängler ne donnait pas toujours la borne exacte ; supposition renforcée par la démonstration en 1935 dans les Monatshefte par Hofreiter d'une inégalité analogue pour l'approximation simultanée de n nombres complexes par des entiers d'un corps quadratique imaginaire, sans idéaux non principaux, de discriminant $-d$; selon cette inégalité la constante $C_n(d)$ d'approximation simultanée de n nombres complexes sur ce corps (définie de façon tout à fait analogue à C_n) ne surpasse pas $\sqrt{|D_{n+1}(d)|}$, où $D_k(d)$ est le discriminant relatif d'un surcorps de degré relatif k du corps de $\sqrt{-d}$, minimum en module. Naturellement, les constantes d'approximation vraiment simultanée étaient aussi inconnues sur ces corps, mais on connaissait en 1936 les constantes $C_1(d)$ dans les quatre cas les plus simples $(d = 3, 4, 7, 8)$, et on constatait qu'elles étaient toutes égales à $\sqrt{D_2(d)}$.

Le rôle des discriminants de corps algébriques dans ces questions parut encore s'accroître en 1946, après la démonstration par Davenport et Mahler dans le Duke Math. J. du théorème suivant : "Quels que soient les nombres réels α et β

il existe une infinité de triplets d'entiers x , y , z tels que $|z| \left\{ (x-\alpha z)^2 + (y - \beta z)^2 \right\} \leqslant \frac{2}{\sqrt{23}}$, et la constante du second membre est la plus petite possible". De ce théorème résulte évidemment $C_2 \geqslant \frac{\sqrt{23}}{2} = 2,39 \ldots$

Les suppositions sur le rôle du corps cubique de discriminant -23 dans l'approximation simultanée ont été infirmées. Déjà en 1950, Descombes et moi-même avions démontré, en reprenant les problèmes d'approximation de nombres complexes, que $C_1(11) < \sqrt{D_2(11)}$ (Comptes-Rendus) , et la même année, Mullender avait démontré dans les Annales que $C_2 \geqslant \frac{23^{\frac{1}{4}} 3^{\frac{9}{2}}}{2^7}$, quantité de très peu supérieure à $\frac{\sqrt{23}}{2}$. Davenport, reprenant la méthode de Mullender, obtint en 1952 (Proceedings London M. Soc.) l'inégalité $C_2 \geqslant 46^{1/4} = 2,60 \ldots$ qui constitue la meilleure minoration actuellement connue de C_2 . Enfin, la majoration de Furtwängler $C_2 \leqslant \sqrt{23}$ a elle-même été améliorée par Cassels en 1955 (Journal London M. Soc.) qui obtient $C_2 \leqslant \frac{7}{2} = 3,5$. Le résultat de Cassels a été intégré par Davenport dans le fascicule suivant du même journal à un exposé synthétique dont nous donnerons plus loin une idée. Tel est l'état actuel de la question. Nous allons examiner maintenant les méthodes utilisées.

Les démonstrations les plus naturelles de l'égalité $C_1 = \sqrt{5}$ s'appuient sur l'algorithme des fractions continues ; on a cherché pendant longtemps des algorithmes analogues pour l'approximation simultanée, mais en vain. Hermite avait essayé d'appliquer à ce problème la réduction continuelle de la forme quadratique $a(x - \alpha z)^2 + b(y - \beta z)^2 + z^2$ $(a > 0 , b > 0)$, sans obtenir "l'algorithme régulier" qu'il appelle de ses voeux dans cette lettre du 2 mai 1894 , adressée à Stieltjes, qui serait à citer longuement et dont j'extrais cette phrase célèbre : "La recherche des fractions $\frac{P'}{P}$, $\frac{P''}{P}$ qui approchent le plus de deux nombres donnés n'a cessé, depuis plus de 50 ans, de me préoccuper et aussi de me désespérer".

Les essais ultérieurs ont confirmé la difficulté de cette voie, et la plupart des algorithmes proposés en généralisation des fractions continues ne fournissent pas de bonnes approximations simultanées. L'algorithme de Jacobi, par exemple, donne des entiers z , x_1 , ... , x_n , tels qu'il n'est même pas sûr que $z\alpha_i - x_i$ tende vers 0 , et la modification proposée par Chabauty et Pisot, si elle écarte cette anomalie, n'assure pas qu'on obtienne les meilleures approximations simultanées. D'ailleurs, le succès des fractions continues tient aussi aux propriétés de périodicité qui sont ici, soit hypothétiques, soit même incompatibles avec de bonnes approximations simultanées. Pourtant, il me semble que l'algorithme proposé par Furtwängler dans ses mémoires déjà cités mériterait d'être examiné plus à fond. Furtwängler définit (par exemple pour $n = 2$) a priori les triplets de meilleurs approximation (x, y, z) des deux nombres α et β par la condition que le système

$$|Z| < |z| \quad , \quad \max |X - \alpha Z|, \ |Y - \beta Z| < \max |x - \alpha z|, \ |y - \beta z|$$

est impossible en entiers X, Y, Z (avec des précisions pour les signes $\leq$), et il montre que si on range ces triplets par z croissants, 3 triplets consécutifs ont toujours un déterminant de valeur absolue $0, 1$ ou 2. On remarque que Descombes et moi-même nous sommes inspirés de cette idée pour définir les fractions continues complexes que j'ai utilisées dans ma thèse. Notons encore que Davenport a récemment consacré un article à la généralisation, pour n quelconque, de cette majoration de déterminant.

Devant l'échec des méthodes algorithmes, ce sont les méthodes géométriques qui ont formé les résultats mentionnés plus haut. Les progrés réalisés l'ont été, comme le montre clairement le dernier article de Davenport, tant en prenant conscience de l'égalité de la constante C_n avec un certain déterminant critique, qu'en améliorant les procédés d'estimation de celui-ci (même avant que l'identité précédente ait été démontrée). Cette identité de C_n avec un déterminant critique n'est énoncé et démontré explicitement que par Cassels (pour C_2) et Davenport (pour C_n) dans leurs articles cités de 1955, mais elle était "dans l'air" depuis quelque temps, surtout depuis l'article de Davenport-Mahler de 1946.

Nous allons donner une idée de la démonstration de cette égalité, en la décomposant en deux inégalités dont aucune n'est d'ailleurs évidente. Précisons-la d'abord en disant qu'il s'agit du déterminant critique de la portion de R^{n+1} définie par $|X_{n+1}| \max\limits_{1 \leq i \leq n} |X_i|^n < 1$ c'est-à-dire comme on le sait de la borne inférieure du déterminant d'un réseau passant par 0 et n'ayant pas d'autre point dans la région précédente. Désignons par Δ_n ce déterminant critique.

<u>Démonstration de</u> $C_n \geqslant \Delta_n$. Si $\gamma < \Delta_n$, des formes linéaires de déterminant 1 ont toujours des valeurs telles que $|X_{n+1}| \max_{1\leqslant i\leqslant n} |X_i|^n < \frac{1}{\gamma}$; c'est le cas en particulier pour les formes $X_i = x_i - \alpha_i z$ $(1 \leq i \leq n)$ $X_{i+1} = z$; donc tout serait démontré si l'on savait remplacer "des valeurs" par "une infinité de valeurs". Il suffirait de pouvoir satisfaire outre l'inégalité précédente, les inégalités $|X_i| < \varepsilon$, où ε est un nombre positif arbitraire et $i = 1 , 2 , \ldots , n$; donc tout revient à démontrer que les régions

$$K \quad : \quad |X_{n+1}| \max_{1\leqslant i\leqslant n} |X_i|^n < 1 \qquad \text{et}$$

$$K_\varepsilon \quad : \quad |X_{n+1}| \max_{1\leqslant i\leqslant n} |X_i|^n < 1 \quad , |X_i| < \varepsilon \quad (i = 1 , \ldots n)$$

ont même déterminant critique, soit $\Delta(K) = \Delta(K_\varepsilon)$; mais ceci résulte des faits suivants

1°) d'après la théorie de Mahler (voir un exposé par exemple dans Chabauty, Annales ENS LXVI) $\Delta(K_\varepsilon) \longrightarrow \Delta(K)$ quand $\varepsilon \longrightarrow \infty$

2°) les K_ε se déduisent les uns des autres par des transformations unimodulaires, donc ont le même déterminant critique. Ces transformations sont définies par

$$X'_i = X_i \frac{\varepsilon_1}{\varepsilon_2} \qquad X'_{n+1} = X_{n+1} \left(\frac{\varepsilon_2}{\varepsilon_1}\right)^n \qquad (i = 1 , 2 , \ldots , n) .$$

<u>Démonstration de</u> $C_n \leq \Delta_n$. Elle est nécessaire parce que les formes X_i qui interviennent dans l'approximation simultanée ne sont pas des formes quelconques ; leur matrice présente, après suppression de la dernière ligne et de la dernière colonne, une matrice unité à n lignes et n colonnes ; la démonstration s'appuie sur un lemme de Davenport inspiré de Furtwängler qui dit qu'une matrice d'ordre $n+1$ quelconque peut être ramenée, par une substitution unimodulaire des variables, à une forme telle qu'après suppression de la dernière ligne et de la dernière colonne, on obtienne une matrice d'ordre n qui est approximativement un multiple de la matrice unité ; précisément, ε étant donné, on peut trouver N tel que la matrice ait la forme (a_{ij}) avec
$|a_{ii} - N| < N^\varepsilon$ pour $i \leqslant n$ et $|a_{ij}| < N^\varepsilon$ pour $i \neq j$, i et $j \leqslant n$.

A l'aide de ce lemme, de démonstration assez technique, on prouve sans difficulté l'inégalité annoncée.

<u>Théorème de Furtwängler</u>. Etant donné un corps algébrique de degré k de discriminant d, il existe dans ce corps et ses conjugués des formes linéaires $M_1, \ldots, M_k$ de déterminant $\sqrt{d}$ telles que $|M_1, \ldots, M_k| \geqslant 1$ pour tout système de variables entières non toutes nulles. Si $k = r + 2s$, r étant le nombre de conjugués réels du corps, et si $M_{2s+1}, M_{2s+2}, \ldots, M_k$ sont réelles, M_i et M_{i+s} étant imaginaires conjugués, posons $X_i + X_{s+i}\sqrt{-1} = M_i\sqrt{2}$ pour $i \leqslant s$ et $X_i = M_i$ pour $i \geqslant 2s+1$

Les formes $X_1, \ldots, X_k$ sont réelles et ont pour déterminant $\sqrt{|d|}$, et on a évidemment

$$X_k \max_{1 \leq i \leq k-1} |X_i|^{k-1} \geqslant |M_1, \ldots, M_k| \geqslant 1$$

pour des valeurs entières non toutes nulles des variables, puisque

$$\max |X_i|^2 , |X_{s+1}{}^2| \geqslant \frac{X_i^2 + X_{s+i}^2}{2} = M_i \, M_{i+s}$$

Donc $\Delta_{k-1} \leq \sqrt{|d|}$: c'est l'inégalité de Furtwängler.

<u>Théorème de Cassels</u>. Dans le corps cubique réel de discriminant 49, il existe trois formes linéaires L_1, L_2, L_3 de déterminant 7 telles que $|L_1 \, L_2 \, L_3| \geqslant 1$ pour tout système de variables entières non toutes nulles.
Posons

$$X_3 = L_3$$
$$X_1 = \frac{L_1 + L_2}{2}$$
$$X_2 = \frac{L_1 - L_2}{2}$$

Alors $|X_3| \max (X_1^2, X_2^2) = |X_3| \left(\frac{|L_1| + |L_2|}{2}\right)^2 \geqslant |L_1 \, L_2 \, L_3| \geqslant 1$. Dans les mêmes conditions ; comme X_1, X_2, X_3 ont pour déterminant $\frac{7}{2}$, ceci prouve que $\Delta_2 \leq \frac{7}{2}$ et par suite $C_2 \leq \frac{7^2}{2}$.