COLLOQUIUM MATHEMATICUM

VOL. VI DÉDIÉ À M. CASIMIR KURATOWSKI 1958

REMARKS ON THE THEORY OF DIOPHANTINE APPROXIMATION

BY

P. ERDÖS, P. SZÜSZ AND P. TURÁN (BUDAPEST)

The problems of the theory of diophantine approximation concern in general the solvability and non-solvability of systems of inequalities in rational integers (or integers of an algebraic extension $R(\vartheta)$ of the rational field). However, in the case of solvability, not very much is known about the localization of the solutions. The significance of this point of view concerning the classical theorems of Dirichlet and Kronecker was shown recently in a book [2], by the third of the present authors. In this note we shall discuss the localization-problem concerning the inequality

$$|\alpha - x/y| \leqslant A/y^2, \tag{1}$$

where A is a positive constant,

$$0 < \alpha < 1 \tag{2}$$

and x, y are integers subjected to

$$(x, y) = 1, \quad y > 1. \tag{3}$$

At a given A, as we know, even in the case of solvability, no interval I on the half-line $y > 1$ can be preassigned in such a way that the system (1)-(3) has certainly a solution with y in I for *all* α's in $0 < \alpha < 1$. However, if we drop the requirement $(x, y) = 1$, the situation changes. As the second of us proved (see [1]), there is a constant $N_0 > 1$ such that the inequality $|\alpha - x/y| \leqslant y^{-2}$ has a solution with $N \leqslant y \leqslant N^2$ for all α's in $0 < \alpha < 1$ if only $N > N_0$ and this is the best-possible in the sense that N^2 cannot be replaced by $o(N^2)$. Here we shall make the first step towards the solution of the

Problem I (P 241). *For fixed $A > 0$ and $c > 1$ we denote by $S(N, A, c)$ the set of those α's for which with an integer $N \geqslant 2$ the system* (1)-(3) *is solvable with an integer*

$$N \leqslant y \leqslant cN. \tag{4}$$

If $|S(N, A, c)|$ stands for the measure of $S(N, A, c)$, does

$$\lim_{N\to\infty} |S(N, A, c)| = f(A, c) \tag{5}$$

exist and, if it exists, what is its explicit form?

If we take into account the previous remarks, the localization (4) seems to be very strong and one might guess that $f(A, c) \equiv 0$. It is somewhat surprising that this is not the case. We shall prove

THEOREM I. *We have for $A > 0$, $c > 1$*

$$\lim_{N\to\infty} |S(N, A, c)| \geqslant \frac{3}{\pi^2}\left(1-\frac{1}{c^2}\right)\min(1, 2A).$$

THEOREM II. *For $A \geqslant 1$ and $c \geqslant 2$ we have the stronger estimation*

$$\lim_{N\to\infty} |S(N, A, c)| \geqslant \frac{3}{\pi^2}\left(\frac{5}{4}-\frac{2}{c^2}\right).$$

THEOREM III. *For $0 < A < c/(1+c^2)$ the limes exist and*

$$f(A, c) = \frac{12A}{\pi^2}\log c.$$

THEOREM IV. *For $A > 10$, $c > 10$ say, we have for all sufficiently large N*

$$|S(N, A, c)| < 1-\frac{1}{40A^4c^4\pi},$$

i. e. if the $\lim_{N\to\infty}|S(N, A, c)|$ exists, it is < 1.

A proof that $f(A, c)$ exists for $A > 0$, $c > 1$, seems to be rather difficult. Theorem I for $A = 1/2$ gives the first step towards the solution of the following problem of the metrical theory of continued fractions which was the starting point of the present investigations:

Problem II (P 242). *Denoting the set of those α's in $0 < \alpha < 1$, for which with an integer $N \geqslant 2$ and $c > 1$ the interval $N \leqslant y \leqslant cN$ contains at least one denominator q_ν of the regular continued fraction of α, by $R(N, c)$, does*

$$\lim_{N\to\infty} |R(N, c)| = \Phi(c) \tag{6}$$

exist and, if it exists, what is its explicit form?

Namely, since any fraction x/y with $(x, y) = 1$ and

$$|\alpha - x/y| < 1/2y^2$$

is a convergent of α, theorem I gives immediately the following

COROLLARY. *For the above defined $R(N, c)$-set we have*

$$\lim_{N\to\infty} |R(N, c)| \geqslant \frac{3}{\pi^2}\left(1-\frac{1}{c^2}\right).$$

Next we pass to the proofs of the above theorems. In order to prove theorem I let g be an integer with

$$N \leqslant g \leqslant [cN]-1 \tag{7}$$

and h an integer with

$$1 \leqslant h \leqslant g-1, \quad (h, g) = 1. \tag{8}$$

First we assert that for two different pairs of such integers we have

$$\left|\frac{h_1}{g_1}-\frac{h_2}{g_2}\right| > \frac{1}{[cN]^2}. \tag{9}$$

For if not and we had

$$h_1/g_1 < h_2/g_2, \tag{10}$$

then we had

$$|h_1g_2-h_2g_1| \leqslant g_1g_2/[cN]^2 < 1$$

which contradicts to (10). Then we construct open intervals $I(h/g)$ around each of our fractions h/g as centres of the length

$$\frac{\min(1, 2A)}{[cN]^2}.$$

It follows from (9) that no two of these intervals have common points; further for all α's in each $I(h/g)$ we have

$$|\alpha-h/g| \leqslant A/[cN]^2 < A/g^2,$$

i. e. (1)-(3)-(4) are satisfied as well. Hence

$$|S(N, A, c)| > \frac{\min(1, 2A)}{[cN]^2}\sum_{g=N}^{[cN]-1}\varphi(g),$$

where $\varphi(g)$ stands for the usual Euler number-theoretical function. Since, as we know,

$$\lim_{x\to\infty}\frac{1}{x^2}\sum_{n\leqslant x}\varphi(n) = \frac{3}{\pi^2}, \tag{11}$$

we have

$$\lim_{N\to\infty}\frac{1}{N^2}\sum_{N\leqslant g\leqslant[cN]-1}\varphi(g)=\frac{3}{\pi^2}(c^2-1),$$

and theorem I follows.

In order to prove theorem II we start from the following remark. Let $I_1,\dots,I_k$ be finitely many intervals, which might have common parts, J the union of all I_ν's and $I^{(l)}$ the subset of J, which is covered by the I_ν's at least l times. Then we have $l\leqslant k$ and

$$|J|=\sum_\nu|I_\nu|-\sum_{l=2}^{k}|I^{(l)}|. \tag{12}$$

We again consider the points h/g with (7) and (8) and construct around each h/g as centre the open interval $I^*(h/g)$ with the length $2/[cN]^2$. For the α's of $I^*(h/g)$ we have

$$|\alpha-h/g|<1/[cN]^2\leqslant A/g^2$$

owing to (7) and $A\geqslant 1$; from (8) it follows that

$$I^*(h/g)\subset S(N,A,c). \tag{13}$$

The intervals $I^*(h/g)$ may now have common parts; we assert, however, that *no three* of them have a common point. Indeed, if $h_1/g_1<h_2/g_2<h_3/g_3$ are any three consecutives of our fractions (7)-(8), then we have from (9)

$$\frac{h_3}{g_3}-\frac{h_1}{g_1}>\frac{2}{[cN]^2},$$

i. e. $I^*(h_1/g_1)$ and $I^*(h_3/g_3)$ cannot have common points. Hence (12) and (13) give

$$|S(N,A,c)|\geqslant\sum_{h,g}|I^*(h/g)|-|I^{(2)}|. \tag{14}$$

In order to estimate the right-hand side of (14) from below let our fractions be

$$0<h_1/g_1<h_2/g_2<\dots<1 \quad\text{and}\quad \frac{h_{\nu+1}}{g_{\nu+1}}-\frac{h_\nu}{g_\nu}=\delta_\nu.$$

Obviously, the intervals $I^*(h_\nu/g_\nu)$ and $I^*(h_{\nu+1}/g_{\nu+1})$ have a common part if and only if $\delta_\nu<2/[cN]^2$ and their contribution to $I^{(2)}$ is $2/[cN]^2-\delta_\nu$. Thus from (14) it follows that

$$|S(N,A,c)|\geqslant\sum_\nu\frac{2}{[cN]^2}-\sum_{\delta_\nu<2/[cN]^2}\left(\frac{2}{[cN]^2}-\delta_\nu\right)$$

$$=\sum_{\delta_\nu\geqslant 2/[cN]^2}\frac{2}{[cN]^2}+\sum_{\delta_\nu<2/[cN]^2}\delta_\nu.$$

From (9) we have $\delta_\nu>1/[cN]^2$ and therefore

$$|S(N,A,c)|>\frac{1}{[cN]^2}\sum_{g=N}^{[cN]-1}\varphi(g)+\frac{1}{[cN]^2}\sum_{\delta_\nu\geqslant 2/[cN]^2}1. \tag{15}$$

As to the second sum in (15) we may observe that for all fractions h_ν/g_ν with

$$N\leqslant g_\nu\leqslant\left[\frac{c}{2}N\right]-1$$

the condition $\delta_\nu\geqslant 2/[cN]^2$ is fulfilled. Indeed, owing to

$$\left[\frac{c}{2}N\right]-1\leqslant\frac{[cN]-1}{2}<\frac{[cN]}{2},$$

we have the inequality

$$\delta_\nu=\frac{h_{\nu+1}}{g_{\nu+1}}-\frac{h_\nu}{g_\nu}\geqslant\frac{1}{g_\nu g_{\nu+1}}>\frac{1}{\left(\left[\frac{c}{2}N\right]-1\right)[cN]}>\frac{2}{[cN]^2}.$$

Thus the second sum in (15) is greater than

$$\frac{1}{[cN]^2}\sum_{N\leqslant g\leqslant[cN/2]-1}\varphi(g).$$

Using this and (11) theorem II follows from (15).

Next we turn to the proof of theorem III. Around each of our h/g's with (7) and (8), as centres, we construct an interval of the length $2A/g^2$. If we can prove that no two of these intervals have a common point, then we obviously have

$$|S(N,A,c)|=2A\sum_{N\leqslant g\leqslant[cN]-1}\frac{\varphi(g)}{g^2}. \tag{16}$$

In order to show that no two intervals of the above type have a common point, let $h_\nu/g_\nu < h_{\nu+1}/g_{\nu+1}$ be two consecutive ones of our fractions; then there is no overlapping indeed if we can prove that

$$\left(\frac{h_{\nu+1}}{g_{\nu+1}} - \frac{A}{g_{\nu+1}^2}\right) - \left(\frac{h_\nu}{g_\nu} + \frac{A}{g_\nu^2}\right) > 0.$$

But this is true indeed, since the difference on the left is not smaller than

$$\frac{1}{g_\nu g_{\nu+1}} - \frac{A}{g_\nu^2} - \frac{A}{g_{\nu+1}^2} = \frac{1}{g_\nu^2}\left\{\frac{g_\nu}{g_{\nu+1}} - A - A\left(\frac{g_\nu}{g_{\nu+1}}\right)^2\right\},$$

further $g_\nu/g_{\nu+1}$ is certainly between c and $1/c$ and the quadratic function $y - A - Ay^2$ is non-negative for $1/c \leqslant y \leqslant c$. Since partial summation from (11) gives at once

$$\sum_{n \leqslant x} \frac{\varphi(n)}{n^2} \sim \frac{6}{\pi^2} \log x$$

for $x \to \infty$, theorem III follows from (16).

Finally we prove theorem IV. We shall prove it in a twofold sharper form; denoting by $S^*(N, A, c)$ the set of α's with the property that

$$|\alpha - x/y| \leqslant A/N^2, \tag{17}$$

is solvable with integer x and y satisfying

$$N \leqslant y \leqslant cN \tag{18}$$

(i. e. dropping the restriction $(x, y) = 1$) we obviously have $S(N, A, c) \subset S^*(N, A, c)$ and we assert that the inequality

$$|S^*(N, A, c)| \leqslant 1 - \frac{1}{40\pi A^4 c^4} \tag{19}$$

holds for all sufficiently large N's. To prove (19) we consider the intervals

$$\frac{a}{b} + \frac{A^2c^2}{20N^2} \leqslant \alpha \leqslant \frac{a}{b} + \frac{A^2c^2}{10N^2}, \tag{20}$$

where

$$N/2A^3c^3 \leqslant b \leqslant N/A^3c^3 \tag{21}$$

and

$$1 \leqslant a < b, \quad (a, b) = 1. \tag{22}$$

If $a/b < a'/b'$ are two consecutive ones of our fractions, we have from (21)

$$\frac{a}{b} + \frac{A^2c^2}{10N^2} = \frac{a'}{b'} + \left(\frac{a}{b} - \frac{a'}{b'}\right) + \frac{A^2c^2}{10N^2} \leqslant \frac{a'}{b'} - \frac{1}{bb'} + \frac{A^2c^2}{10N^2}$$

$$< \frac{a'}{b'} - \frac{A^6c^6}{N^2} + \frac{A^2c^2}{10N^2} < \frac{a'}{b'} + \frac{A^2c^2}{20N^2},$$

i. e. the intervals (20) do not overlap. Their total length is for sufficiently large N's

$$\frac{A^2c^2}{20N^2} \sum \varphi(b) > \frac{1}{40\pi A^4 c^4},$$

using (21) and (11). Hence, if we succeed in proving that for the α's in (20) the inequalities (17)-(18) are not solvable, the proof of theorem IV will be finished.

In order to prove this assertion we show first that fixing α in (20) the solution x/y of (17) cannot be chosen as a/b. The assumption $x/y = a/b$ would yield owing to (20) and (17)

$$\frac{A^2c^2}{20N^2} \leqslant \alpha - \frac{a}{b} = \alpha - \frac{x}{y} = \left|\alpha - \frac{x}{y}\right| < \frac{A}{N^2},$$

which is false owing to $A > 10$, $c > 10$. If finally $x/y \neq a/b$, then owing to (18), (21) and (17) we have

$$\frac{A^3c^2}{N^2} = \frac{1}{(cN)(N/A^3c^3)} < \frac{1}{yb} \leqslant \left(\frac{x}{y} - \frac{a}{b}\right)$$

$$\leqslant \left|\frac{x}{y} - \alpha\right| + \left|\alpha - \frac{a}{b}\right| \leqslant \frac{A}{N^2} + \frac{A^2c^2}{10N^2},$$

which is again false owing to $A > 10$, $c > 10$.

Added in proof. We can prove the following theorem: *Let A and ε be arbitrary posittve numbers. Then there exist $c_0 = c_0(A, \varepsilon)$ and $N_0 = N_0(A, \varepsilon)$ so that for $c > c_0$ and $N > N_0$*

$$S(N, A, c) > 1 - \varepsilon.$$

As a corollary we obtain: for each $\varepsilon > 0$ there is a c_0 and an N_0 such that the set of those numbers in $(0, 1)$ to which there is a con-

vergent with denominator q_k satisfying

$$N < q_k < eN \quad (c < c_0, N < N_0)$$

has a measure greater than $1-\varepsilon$.

We shall return to this subject elsewhere.

REFERENCES

[1] P. Szüsz, *Bemerkungen zur Approximation einer reellen Zahl durch Brüche*, Acta Mathematica Academiae Scientiarum Hungaricae 6 (1955), p. 203-212.

[2] P. Turán, *Eine neue Methode in der Analysis und deren Anwendungen*, Budapest 1953 (for an enlarged version see the Chinese edition, Peking 1956).

Reçu par la Rédaction le 16. 11. 1957, en version modifiée le 10. 6. 1958

COLLOQUIUM MATHEMATICUM

VOL. VI DÉDIÉ À M. CASIMIR KURATOWSKI 1958

ON THE APPROXIMATE SOLUTIONS OF FUNCTIONAL EQUATIONS IN L^p SPACES

BY

M. ALTMAN (WARSAW)

In papers [1] and [2] we have suggested an iterative method for solving non-linear functional equations in Banach spaces. This method may also be regarded as a generalization of Newton's well-known classical method. But this generalization is essentially different from that given by L. V. Kantorovitch [5].

The present paper contains a specification for the case of the real L^p-spaces and a real Hilbert space of the iterative method defined in paper [2]. An application to approximate solutions of operator equations in this space is also given. In particular we consider in the L^p-spaces an analogue of the method of steepest descent for non-linear operator equations.

The iterative process for solving non-linear functional equations is defined in papers [1], [2] as follows:

Let X be a Banach space and let $F(x)$, $x \in X$, be a non-linear continuous functional which is differentiable in the sense of Fréchet. Then the approximate process for solving the non-linear functional equation

$$F(x) = 0 \tag{1}$$

is defined by the formula

$$x_1 = x_0 - \frac{F(x_0)}{f_0(y_0)} y_0, \qquad x_{n+1} = x_n - \frac{F(x_n)}{f_n(y_n)} y_n, \tag{2}$$

where x_0 is the initial approximate solution, $f_n = F'(x_n)$ for $n = 0, 1, 2, \ldots$ denotes the Fréchet differential of $F(x)$ at the point $x = x_n$, and y_n are elements appropriately chosen in X, *i. e.* $\|y_n\| = 1$, $f_n(y_n) = \|f_n\|$, $n = 0, 1, 2, \ldots$, provided that such a choice is possible.

The specification for the case of the real L^p-spaces and the real Hilbert space consists in the appropriate choice of the elements y_n. It appears that in this case the choice of the elements y_n is effective and may be realized in a simple manner.