SIMULTANEOUS DIOPHANTINE APPROXIMATION

GORDON RAISBECK

Summary of results. The principal result of this paper is as follows: given any set of real numbers $z_1, z_2, \ldots, z_n$ and an integer t we can find an integer $q \leqslant t^{n-1}(t-1) - 2^{n-1} + 1$ and a set of integers $p_1, p_2, \ldots, p_n$ such that

$$|qz_j - p_j| \leqslant 1/t \qquad (j = 1, 2, \ldots, n). \tag{0.11}$$

Also, if n = 2, we can, given t, produce numbers z_1 and z_2 such that

$$|qz_j - p_j| > 1/t \qquad \text{for all } q < t^{n-1}(t-1) - 2^{n-1} + 1. \tag{0.12}$$

This supersedes the results of Nils Pipping (*Acta Aboensis*, vol. 13, no. 9, 1942) that there is a q satisfying (0.11) such that $q \leqslant t^n - 2^n + 1$, and also the classical result of Dirichlet that there is such a q less than t^n.

1.1. Suppose we are given a set of n positive real numbers $z_1, z_2, \ldots, z_n$. The problem of simultaneous Diophantine approximations is concerned with finding integers q and p_j such that the quantities

$$|qz_1 - p_1|, |qz_2 - p_2|, \ldots, |qz_n - p_n| \tag{1.11}$$

are small.

For convenience we shall use the following notations:

$$X \equiv (x_1, x_2, \ldots, x_n),$$
$$aX + bY \equiv (ax_1, + by_1, ax_2 + by_2, \ldots, ax_n + by_n),$$
$$f(X) \equiv f(x_1, x_2, \ldots, x_n).$$

Throughout this paper q, p_j, and t will be integers, and n will be a fixed integer.

In order to define what we mean by making the quantities (1.11) small, we shall suppose that there is a continuous positive real-valued function f of n real variables such that the equation

$$f(X) < t^{-1} \tag{1.12}$$

defines for every[1] $t > 0$ a simply connected star-shaped region containing the origin such that

$$f(aX) \leqslant f(X) \qquad \text{for } a \leqslant 1 \tag{1.13}$$

Received March 25, 1949. This paper contains material which will be presented by the author as part of the requirements for a Ph.D. degree at the Massachusetts Institute of Technology.

[1]Here, exceptionally, t need not be an integer.

and such that

$$f(X) = 0 \text{ implies } x_j = 0 \qquad (j = 1, 2, \ldots, n). \tag{1.14}$$

Then we will measure the smallness of the set of quantities (1.11) by the single quantity $f(qZ - P)$.

1.2. There are now two principal courses open. We may ask: *Given Z, for what infinite class of integers q do there exist integers P such that $f(qZ - P)$ is of the smallest order of magnitude?* Or we may ask: *Given t (generally, and in this paper, supposed to be an integer) and Z, what is the smallest value of q for which we can find integers P such that*

$$f(qZ - P) \leqslant t^{-1} \tag{1.21}$$

or

$$f(qZ - P) < t^{-1}? \tag{1.22}$$

We shall ask the question in the second form.

1.3. The particular values of $f(X)$ which have attracted most attention are:

$$f(X) = \max |x_j|, \tag{1.31}$$

$$f(X) = \sum_1^n |x_j|, \tag{1.32}$$

$$f(X) = [\sum_1^n x_j^2]^{\frac{1}{2}}, \tag{1.33}$$

$$f(X) = \prod_1^n |x_j|. \tag{1.34}$$

In the case (1.31) it was proved by Dirichlet that there is some $q \leqslant t^n$ for which (1.21) and (1.22) are true. Nils Pipping (*loc. cit.*) has shown that (1.21) is always satisfied with some $q \leqslant t^n - 2^n + 1$. We will show that there is a q for which (1.21) is satisfied such that $q \leqslant t^{n-1}(t - 1) - 2^{n-1} + 1$, and that in the case $n = 2$ this is the best possible result.

In the case (1.32) our result will be that (1.21) and (1.22) are satisfied for some $q \leqslant t^n\, n!$ and that this is nearly the best possible result if $n = 2$, but that if $n \geqslant 3$ this estimate is probably not the best possible.

In the case (1.33) it will be shown that an inequality for q may be based on a knowledge of the maximum density to which equal spheres can be packed in Euclidean n-space.

In the case (1.34) our method will not apply.

2.1. We shall suppose we are given a set of n positive real numbers Z and a function $f(X)$ continuous in n variables such that $f(X) \leqslant c$ is a convex n-dimensional solid of diameter less than 1 symmetrical about $0, 0, \ldots, 0$ for any c such that $0 < c < \frac{1}{2}$, and such that $f(X) = 0$ implies $x_j = 0$ $(j = 1, 2, \ldots, n)$. We will investigate the values of q for which it is possible to find a set of integers P such that

$$f(qZ - P) \leqslant t^{-1}. \tag{2.11}$$

We shall refer to any such value of q as a "good q" and to the smallest (if any) as $\mathbf{q}$.

We define a convex solid as follows: if X' and X'' are interior points of the solid, so are $aX' + (1 - a)X''$, for all a such that $0 < a < 1$; and if X' and X'' are interior or boundary points of the solid, then so are $aX' + (1 - a)X''$ for $0 < a < 1$.

2.2. With a fixed q we will associate all the points $(qZ - P)$, where p_j runs through all integer values, positive or negative, for every j. This gives us a cubic lattice of points. We shall say that these points "belong to q."

About each of the points belonging to q we shall construct a solid

$$f[2(X - qZ + P)] \leqslant t^{-1}. \tag{2.21}$$

We shall call each of these solids a block, and say that each of the blocks defined above "belongs to q."

Suppose two blocks belonging to different q, say q' and q'', overlap or touch. We shall show in this case

$$f[(q' - q'')Z - (P' - P'')] \leqslant t^{-1}. \tag{2.22}$$

For let the centers of the blocks be $R = q'Z - P'$ and $S = q''Z - P''$. Let X be a point interior to or on the boundary of both blocks. Then:

$$f[2(R - X)] \leqslant t^{-1}, \tag{2.23}$$

$$f[2(S - X)] \leqslant t^{-1}; \tag{2.24}$$

by (2.23) and symmetry

$$f[2(X - R)] \leqslant t^{-1}; \tag{2.25}$$

by (2.24), (2.25), and the convexity property of the blocks

$$f\left[2\left(\frac{S - X}{2} + \frac{X - R}{2}\right)\right] \leqslant t^{-1}. \tag{2.26}$$

Hence

$$f[(q' - q'')Z - (P' - P'')] \leqslant t^{-1}, \tag{2.27}$$

which is (2.22).

2.3. Let us now give q the values $0, 1, 2, \ldots, m$ and construct the blocks belonging to these q. Let m be a value of q such that no two of the blocks so constructed overlap or touch.

Consider the unit cube $0 \leqslant x_j < 1$. From the construction above we can see that all the blocks or fractions of a block in this unit cube which belong to one q may be considered as a dissection of one single block. Hence, if the volume of one single block is V, the volume of all the blocks and parts of a block in the unit cube is $(m + 1)V$. But no two of the blocks overlap or touch. Hence, comparing the volume of the unit cube to the volume of the blocks it contains

$$(m + 1)V \leqslant 1.$$

Therefore m has a maximum which we may call m', and

$$m' + 1 = \mathbf{q} \leqslant V^{-1}.$$

If $f(X) = \max |x_j|$, then $V = t^{-n}$, and

$$\mathbf{q} \leqslant t^n, \tag{2.31}$$

which is the classic result of Dirichlet. If $f(X) = \sum_1^n |x_j|$, then $V = (n!\, t^n)^{-1}$, and

$$\mathbf{q} \leqslant t^n\, n!. \tag{2.32}$$

If $f(X) = [\sum_1^n |x_j^2|]^{\frac{1}{2}}$, then $V = k_n t^{-n}$, where $k_1 = 1$, $k_2 = \dfrac{\pi}{4}$, $k_n = \dfrac{\pi k_{n-2}}{2n}$; and

$$\mathbf{q} \leqslant t^n / k_n. \tag{2.33}$$

2.4. The estimate (2.31) cannot be improved if we insist on strict inequality in (2.21); for we may take $z_j = t^{j-n-1}$. If, however, we permit equality, we shall show in **3.1** that the result can be improved upon.

In the case (1.32), the estimate (2.32) cannot be improved for $n = 2$ if we insist on strict inequality in (2.11); for let

$$z_1 = \frac{1}{2t} - \frac{1}{4t^2}, \qquad z_2 = \frac{1}{2t} + \frac{1}{4t^2}.$$

If n is greater than 2, however, the estimate leaves something to be desired; for, although our proof tells us that even if the blocks fill up space without overlapping, (2.32) follows, it does not show us how to fill space with given blocks. In fact, space cannot be completely filled with equal regular octahedra all oriented in the same direction, as our construction would suggest.

In the case (1.33) our blocks are spheres, and obviously space cannot be filled with equal spheres. Hence (2.33) is not the best possible estimate. However, if we know how densely equal spheres can be packed we can derive an upper bound for $\mathbf{q}$.

3.1. We shall now deal in particular with the case

$$f(X) = \max |x_j|$$

and the inequality

$$f(X) \leqslant t^{-1}$$

where t is an integer. We will show in this case that there exists a good q less than $t^{n-1}(t-1) - 2^{n-1} + 1$.

We perform the same construction as in **2.3**. Now take the t^{n-1} line segments

$$0 \leqslant x_1 \leqslant 1, \quad x_j = (k_j + \tfrac{1}{2})/t, \qquad (j = 2, 3, \ldots, n; \; k = 0, 1, \ldots, t-1).$$

Every block in the unit cube is intersected by at least one of these lines. If a block is intersected by one of these lines, the line segment common to the block and the line has length t^{-1}. Inasmuch as no two of these blocks overlap or touch, no two of these line segments overlap or touch. Hence, not more than $(t-1)$ blocks can be pierced by the given line. Hence

$$m' + 1 \leqslant t^{n-1}(t-1).$$

But the block corresponding to $q = 0$ is counted 2^{n-1} times. Therefore

$$m' + 1 = \mathbf{q} \leqslant t^{n-1}(t-1) - 2^{n-1} + 1.$$

If $n = 2$, this cannot be improved; for we can let

$$z_1 = \frac{1}{t(t-1) - 1}, \quad z_2 = \frac{t-1}{t(t-1) - 1}.$$

It may be observed that we have not used the full force of the hypothesis

$$|qz_j - p_j| > t^{-1} \qquad (j = 1, 2, \ldots, n; \text{ all } q \text{ such that } q < \mathbf{q}; \text{ all } p_j)$$

but only in fact

$$\begin{aligned} |qz_1 - p_1| &> t^{-1} \\ |qz_j - p_j| &\geqslant t^{-1} \end{aligned} \qquad (j = 2, 3, \ldots, n; \text{ all } q \text{ such that } q < \mathbf{q}; \text{ all } p_1 \text{ and } p_j).$$

This follows from the fact that the lines with which we skewered the blocks inside the unit cube are parallel in the direction of increasing x_1, and we use only the fact that blocks on the same skewer do not touch, whereas we say nothing about blocks on different skewers. With the hypothesis thus weakened it is easy to show still that

$$\mathbf{q} \leqslant t^{n-1}(t-1).$$

3.2. In the case

$$f(X) = \sum_{1}^{n} |x_j| \leqslant t^{-1}$$

with $n = 2$, we can show by similar methods that

$$\mathfrak{q} \leqslant t(2t-1) - 1.$$

We do this by considering the intersections of the lines $x_1 + x_2 = (k+\frac{1}{2})/t$ with the blocks.

Bell Telephone Laboratories,
Murray Hill, N.J.