ACTA ARITHMETICA
XLIX (1988)

Rational approximation vectors

by

GEORGE SZEKERES (Kensington) and VERA T.-SÓS* (Budapest)

To Paul Erdős, for his 75th birthday

1. Introduction. It is well known and an easy consequence of the theory of continued fractions that the "best" approximations

$$V_k = N_k\beta - a_k, \quad k = 1, 2, 3, \ldots, a_k \in \mathbf{Z}, N_k \in \mathbf{Z}_{>0}$$

of an irrational β change sign with each successive approximation, that is

$$V_k > 0 \Rightarrow V_{k+1} < 0 \Rightarrow V_{k+2} > 0.$$

Here V_k is called best (or closest) if $|V_k| < |N\beta - a|$ for all integers a, N, $0 < N < N_k$.

Little is known about the analogous problem in higher dimensions. One result by Rogers [3] will be mentioned below. Given $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n) \in \mathbf{R}^n$, the best approximation vectors

$$V_k = N_k\boldsymbol{\beta} - \boldsymbol{a}_k, \quad \boldsymbol{a}_k = (a_{k1}, \ldots, a_{kn}) \in \mathbf{Z}^n, \ N_k \in \mathbf{Z}_{>0}, \quad k = 1, 2, 3, \ldots$$

are characterized by the property that

$$\|N_k\boldsymbol{\beta} - \boldsymbol{a}_k\| < \|N\boldsymbol{\beta} - \boldsymbol{a}\| \quad \text{for all } \boldsymbol{a} \in \mathbf{Z}^n, \quad 0 < N < N_k$$

where $\|x\| = \max_j |x_j|$. For convenience we shall write (for irrational α)

$$\{\alpha\} = \alpha - a, \quad a \in \mathbf{Z}, \quad |\alpha - a| < 1/2,$$

and generally for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$

$$\{\boldsymbol{\alpha}\} = (\{\alpha_1\}, \ldots, \{\alpha_n\}).$$

The notation will also be used for rational α, provided that $|\alpha - a| \neq 1/2$.

* Research supported by Hungarian National Foundation for Scientific Research, Grant No. 1811.

In particular if each β_i is irrational then the best approximation vectors of $\boldsymbol{\beta}$ are

$$V_k = \{N_k \boldsymbol{\beta}\}, \quad \|\{N_k \boldsymbol{\beta}\}\| < \|\{N\boldsymbol{\beta}\}\| \quad \text{for all } 0 < N < N_k,$$
$$V_1 = \{\boldsymbol{\beta}\}, \quad N_1 = 1.$$

N_k is called the *denominator of the approximation* $p_k = a_k/N_k$, and $\|\{N_k \boldsymbol{\beta}\}\|$ the *norm of the approximation.*

Each $V_k = (V_{k1}, \ldots, V_{kn})$ determines an n-signature (briefly, a signature) $\sigma_k = (\eta_{k1}, \ldots, \eta_{kn})$, $\sigma_k^{(j)} = \eta_{kj} = +\text{or}-$, defined by

$$\eta_{kj} V_{kj} > 0, \quad j = 1, \ldots, n.$$

Thus with each irrational $\boldsymbol{\beta}$ we associate the sequence of signatures $\sigma_k = \sigma_k(\boldsymbol{\beta})$, $k = 1, 2, \ldots$ of its best approximation vectors. Sometimes it will be convenient to write $\sigma(k)$ for σ_k.

What are the possible sequences σ_k associated with some $\boldsymbol{\beta}$? First, it is easy to see that two consecutive best approximations cannot have the same signature (Rogers [3]). For suppose that $V_k = \{N_k \boldsymbol{\beta}\}$ and $V_{k+1} = \{N_{k+1} \boldsymbol{\beta}\}$ have the same signature (are in the same n-quadrant). Consider

$$V = \{(N_{k+1} - N_k)\boldsymbol{\beta}\} = (V_{k+1,j} - V_{kj},\ j = 1, \ldots, n).$$

Then clearly

$$\|V\| = \max_j |V_{k+1,j} - V_{kj}| < \max_j |V_{kj}| = \|V_k\|$$

since

$$\max_j |V_{k+1,j}| < \max_j |V_{kj}|$$

by the definition of best approximations. It follows that there is an $\{N\boldsymbol{\beta}\}$ with $N < N_{k+1}$ (namely $N = N_{k+1} - N_k$) which is better than V_k (hence $N > N_k$), contrary to the assumption that V_{k+1} is the next "best" approximation.

We call a signature sequence *proper* if consecutive signatures are always distinct. Thus every signature sequence associated with an irrational $\boldsymbol{\beta}$ is necessarily proper. Is it true that conversely, every proper n-signature sequence is associated with some $\boldsymbol{\beta}$? In particular (in the 2-dimensional case) is there a $\boldsymbol{\beta} = (\beta_1, \beta_2)$ with the property that the successive closest approximations wander around in successive quadrants, clockwise or anticlockwise? The problem in this form came up in dynamical systems and was posed to us by Professor Sinai. We shall find that the answer is yes, in fact every proper n-signature sequence is associated with some $\boldsymbol{\beta} \in \boldsymbol{R}^n$.

THEOREM. *Given any proper sequence of n-signatures σ_k^* there is a $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ such that $\sigma_k(\boldsymbol{\beta}) = \sigma_k^*$ for all $k \geqslant 1$.*

It is interesting to note that if $\boldsymbol{\beta}$ lies on a rational line (one that passes through two points with rational coordinates) then $\sigma_k(\boldsymbol{\beta})$ must ultimately alternate between two opposite n-quadrants, namely the ones determined by the direction of the line. This follows from the following lemma due to John Mack:

MACK'S LEMMA ([1]). *Suppose that the n real numbers $\beta_1, \ldots, \beta_n$ satisfy r linear relations*

$$\sum_{j=1}^{n} B_{ij}\beta_j = A_i \quad (i = 1, \ldots, r)$$

where the B_{ij} are not all zero, and the B_{ij} and A_i are integers. Then there is a constant $c > 0$, depending on the B_{ij} only, such that the following is true:

If $a_1/N, \ldots, a_n/N$ $(a_j \in \boldsymbol{Z},\ N \in \boldsymbol{Z}_{>0})$ satisfy

$$\max_j |N\beta_j - a_j| < c$$

then

$$\sum_{j=1}^{n} B_{ij}\frac{a_j}{N} = A_i \quad (i = 1, \ldots, r).$$

That is, all reasonally good simultaneous rational approximations (certainly all best approximations with big denominators) must satisfy the same rational dependence relationships as $\boldsymbol{\beta}$ itself. In particular if $\boldsymbol{\beta}$ is on a rational line then all good approximations lie on the same line. We shall make substantial use of Mack's lemma in the next section.

Some interesting open questions remain. Suppose $\beta_1, \ldots, \beta_n$ are in a real algebraic number field of degree $n+1$. What signature sequences can such a $\boldsymbol{\beta}$ have? More generally if $\beta_1, \ldots, \beta_n$ are "badly approximable" numbers (e.g. they have bounded continued fraction digits, or $\min_i \liminf_{N \in \boldsymbol{Z}_{>0}} N^{\lambda} |\{N\beta_i\}|$ is positive for some $\lambda \geqslant 1$, or more appropriately $\liminf_{N \in \boldsymbol{Z}_{>0}} N^{\lambda} \min_i |\{N\beta_i\}|$ is positive for some $\lambda \geqslant 1/n$ etc.) what can one say about the signature sequence of $\boldsymbol{\beta}$? Our construction in the next section gives no information about these questions; the components of the constructed $\boldsymbol{\beta}$ are Liouville numbers (hence transcendental) with exceptionally good rational approximations.

2. Proof of the theorem. We denote by Γ_n $(= \boldsymbol{Z}^n)$ the set of lattice points with integer coordinates, by Σ_n the set of points with rational coordinates. If $\sigma = (\eta_1, \ldots, \eta_n)$ is an n-signature, its *opposite* is $\sigma^0 = (-\eta_1, \ldots, -\eta_n)$ (namely the signature of the opposite n-quadrant). If $\sigma \neq \sigma' \neq \sigma^0$, we say that σ' is *adjacent* to σ.

Suppose that we are given a proper n-signature sequence σ_k^*, $k = 1, 2, \ldots$ It determines uniquely a (possibly finite) sequence of integers

$k_1 = 1 < k_2 < k_3 < \ldots$ with the property: k_ν for $\nu > 1$ is the smallest $k_\nu > k_{\nu-1}$ such that $\sigma^*(k_\nu)$ is adjacent to $\sigma^*(k_\nu - 1)$. It is then also adjacent to $\sigma^*(k_{\nu-1})$. The sequence is finite, of length μ, if $\sigma^*_{k+1} = \sigma^{*0}_k$ for all $k \geqslant k_\mu$. For instance the k_ν-sequence is finite for all $\sigma_k(\beta)$ associated with a β on a rational line.

The proof of the theorem rests on the construction of a β whose components have, like Liouville's numbers, exceptionally good simultaneous rational approximations. The point β itself will be obtained as a fast converging limit of rational points $p(k) \in \Sigma_n$, $k = 1, 2, \ldots$, which will then be shown to be the best approximants of β.

To start with set $p(1) = 0$. Suppose we have already constructed $q = p(k_{\nu-1}) = a/N \in \Sigma_n$, $a \in \Gamma_n$, $N \in Z_{>0}$. We can obviously determine a point $q_0 = a_0 \in \Gamma_n$ so that $q_0 - q$ has the given signature $\sigma^*(k_{\nu-1})$. Moreover we can achieve that

$$\|q_0 - q\| = |a_{0s} - a_s/N|$$

for a certain index $s = s_\nu$ which is such that

$$\sigma^*(k_\nu)^{(s)} = (-1)^{d_\nu} \sigma^*(k_{\nu-1})^{(s)}, \quad d_\nu = k_\nu - k_{\nu-1}. \tag{1}$$

Such an s_ν certainly exists since $\sigma^*(k_\nu)$ is adjacent to $\sigma^*(k_{\nu-1})$. We want to construct the points $p(k)$ for $k_{\nu-1} < k \leqslant k_\nu$ so that they should all lie on the line segment L joining q and q_0. If β lies sufficiently close to $p(k_\nu)$ this will ensure that $\beta - p(k)$ for $k_{\nu-1} \leqslant k < k_\nu$ will have the correct signature.

Consider the points

$$q_K = (a_0 + Ka)/(1 + KN), \quad K = 1, 2, \ldots$$

(the iterated medians of q and q_0); clearly they all lie on L and approach q monotonically as $K \to \infty$. We now define

$$p(k_{\nu-1} + 1) = q_K$$

for a suitably large K; the exact conditions that K has to satisfy will be specified later, as we proceed with the construction. At this stage we merely require that K be so large that

$$\|q_K - q\| < \varepsilon_\nu$$

for some $\varepsilon_\nu > 0$ which has been specified in the previous steps of the construction. We may set at any rate $\varepsilon_1 = 1$.

Now set $d = d_\nu = k_\nu - k_{\nu-1}$ and define for $j = 2, \ldots, d$ (if $d > 1$)

$$p(k_{\nu-1} + j) = \frac{1}{Q_j}(F_j(a_0 + Ka) + F_{j-1}\, a)$$

where $F_{-1} = 1$, $F_0 = 0$, $F_1 = 1$, $F_2 = 1$, $F_3 = 2$, $F_4 = 3$, ... is the Fibonacci sequence ($F_{j+1} = F_j + F_{j-1}$) and

$$Q_j = F_j(1 + KN) + F_{j-1} N, \quad j = 0, 1, 2, \ldots$$

(In particular $Q_0 = N$, $Q_1 = 1+KN$.) The points $p(k_{\nu-1}+j)$ alternate on the segment L arround their limit point

$$p(\infty) = \left(a_0 + Ka + \frac{\sqrt{5}-1}{2}a\right) \Big/ \left(1 + KN + \frac{\sqrt{5}-1}{2}N\right).$$

In particular

$$p(k_\nu) = p(k_{\nu-1}+d) = \frac{1}{Q_d}(F_d(a_0+Ka) + F_{d-1}a).$$

If the k_ν sequence is finite, of length μ, and $\nu = \mu+1$ then $d = \infty$ and $p(k_{\mu+1}) = p(\infty)$ above.

Note that

$$\{Nq_K\} = Nq_K - a = (Na_0 - a)/(1+KN) \tag{2}$$

hence as small as we like if K is large enough.

Since there are only a finite number of approximation vectors of points on L with denominator less than N, by taking K large enough we can achieve that $\{Nq_K\}$ is a best approximation vector of q_K and $\{Np(k_\nu)\}$ is a best approximation vector of $p(k_\nu)$. *A fortiori* it is best among approximation fractions which are confined to the line segment L.

It follows from a result of Mack ([2], p. 421) (which essentially states that the best approximations on L are obtained by the ordinary continued fraction process constrained to L) that the best approximations of $p(k_\nu)$ on L with denominators M, $N \leqslant M < Q_d$ are exactly the points $p(k)$, $k_{\nu-1} \leqslant k < k_\nu$, with the proviso (since these points are rational) that the point q_{K-1} (if $d = 1$) or $p(k_\nu - 2)$ (if $d > 1$) has the same approximation norm with respect to $p(k_\nu)$ as $p(k_\nu - 1)$ has. More specifically if $d = 1$

$$\begin{aligned} \{(1+(K-1)N)p(k_\nu)\} &= \{(1+(K-1)N)q_K\} \\ &= \frac{1+(K-1)N}{1+KN}(a_0+Ka) - (a_0+(K-1)a) \\ &= \frac{a-Na_0}{1+KN} = -\{Nq_K\} = -\{Np(k_\nu)\} \end{aligned} \tag{3}$$

by (2), and if $d > 1$

$$\begin{aligned} \{Q_{d-1}p(k_\nu)\} &= \frac{Q_{d-1}}{Q_d}(F_d(a_0+Ka)+F_{d-1}a) - (F_{d-1}(a_0+Ka)+F_{d-2}a) \\ &= (-1)^d(a-Na_0)/Q_d = -\{Q_{d-2}p(k_\nu)\}, \end{aligned} \tag{4}$$

as easily verified from $F_{d-1}F_{d-2} - F_dF_{d-3} = (-1)^{d-1}$.

Now the $p(k)$, $k_{\nu-1} \leqslant k < k_\nu$ are the best approximants of $p(k_\nu)$ even if we do not constrain them to the segment L; there are just no other best approximants off the segment, provided that K was chosen sufficiently large. For $k = k_{\nu-1}$ this is so by the assumption already made, and for $k_{\nu-1} < k < k_\nu$ it is an immediate consequence of Mack's lemma. The limit point $\boldsymbol{\beta} = \lim_{k\to\infty} p(k)$ is of course slightly displaced with respect to $p(k_\nu)$, and the approximation norms with respect to $\boldsymbol{\beta}$ are not quite the same as with respect to $p(k_\nu)$, but the $p(k)$, $k_{\nu-1} \leqslant k < k_\nu$ will remain the best approximants of $\boldsymbol{\beta}$ (and $\boldsymbol{\beta} - p(k)$ will have the same signature as $p(k_\nu) - p(k)$) provided that

(i) $\|\boldsymbol{\beta} - p(k_\nu)\|$ is sufficiently small, and

(ii) the following condition is satisfied (to circumvent the ambiguity of approximation norms mentioned in (3), (4) above):

$$\|\{(1+(K-1)N)\boldsymbol{\beta}\}\| > \|\{(1+(K-1)N)p(k_\nu)\}\| = \|\{Np(k_\nu)\}\| > \|\{N\boldsymbol{\beta}\}\| \quad \text{if} \quad d = 1, \tag{5}$$

$$\|\{Q_{d-2}\boldsymbol{\beta}\}\| > \|\{Q_{d-2}p(k_\nu)\}\| = \|\{Q_{d-1}p(k_\nu)\}\| > \|\{Q_{d-1}\boldsymbol{\beta}\}\| \quad \text{if} \quad d > 1. \tag{6}$$

Condition (i) can be achieved if we prescribe sufficiently small values for $\varepsilon_{\nu+1}, \varepsilon_{\nu+2}, \ldots$ in the subsequent steps of the construction. Condition (ii) is fulfilled because of definition (1) of the index $s = s_\nu$ which specifies the norms of vectors in the direction determined by L. For let us write $p(k_{\nu-1}+i) - p(k_{\nu-1}) = (p_{i1}, \ldots, p_{in})$, $i = 1, \ldots, d$ and suppose that $\sigma^*(k_\nu)^{(s)} = +$ (if it is $-$, the proof goes in exactly the same fashion); then for $1 \leqslant i < d$

$$p_{is} > p_{i+1,s} > 0 \quad \text{if} \quad i \text{ is odd},$$

$$0 < p_{is} < p_{i+1,s} \quad \text{if} \quad i \text{ is even}$$

and in particular $p_{ds} > p_{d-1,s}$ if d is odd, $p_{ds} < p_{d-1,s}$ if d is even. Hence if we set

$$\boldsymbol{\beta} = p(k_\nu) + \boldsymbol{\delta}, \qquad \boldsymbol{\delta} = (\delta_1, \ldots, \delta_n),$$

then condition (1) requires

$$(-1)^d \delta_s > 0. \tag{7}$$

To show (5) suppose $d = 1$, then

$$\{(1+(K-1)N)\boldsymbol{\beta}\} = \{(1+(K-1)N)p(k_\nu)\} + (1+(K-1)N)\boldsymbol{\delta} \tag{8}$$

provided that $\|\boldsymbol{\delta}\|$ is sufficiently small (this is part of the specifications for condition (i) above). But then the s-components of the two terms on the right-hand side of (8) are both negative, by (3) and (7), and the first inequality

in (5) is proved. Similarly

$$\{N\beta\} = \{Np(k_\nu)\} + N\delta$$

for sufficiently small $\|\delta\|$, and the s-component of the first term on the right is positive, by (3), proving the second inequality in (5) (again provided that $\|\delta\|$ is small enough).

Next suppose $d > 1$ and even. Then

$$\{Q_{d-2}\beta\} = \{Q_{d-2}p(k_\nu)\} + Q_{d-2}\delta$$

and the s-components of both terms on the right are positive, proving the first inequality in (6). Similarly

$$\{Q_{d-1}\beta\} = \{Q_{d-1}p(k_\nu)\} + Q_{d-1}\delta$$

and the terms on the right have opposite signs, proving the second inequality in (6). The proofs are the same if $d > 1$ and odd.

This concludes the construction and the theorem is proved. It would be difficult to write down the conditions for δ, the ε_ν and the K explicitly, but it is clear from the construction that the ε_ν must decrease extremely fast, and the components of β are strongly Liouvillean.

References

[1] J. M. Mack, *A note on simultaneous approximation*, Bull. Austral. Math. Soc. 3 (1970), pp. 81–83.

[2] — *On the continued fraction algorithm*, ibid. 3 (1970), pp. 413–422.

[3] C. A. Rogers, *The signatures of errors of simultaneous diophantine approximations*, Proc. London Math. Soc. (2) 52 (1950), pp. 186–190.

SCHOOL OF MATHEMATICS
UNIVERSITY OF NEW SOUTH WALES
Kensington NSW, Australia

MATHEMATICAL INSTITUTE OF THE
HUNGARIAN ACADEMY OF SCIENCE
Budapest, Hungary

Received on 8.4.1986 (1620)