NOTES ON FERGUSON AND FORCADE'S GENERALIZED EUCLIDEAN ALGORITHM

George M. Bergman

1. Background and introduction. Let us recall some of the important properties of the Euclidean algorithm on pairs of real numbers α, β.

If α and β are commensurable, the algorithm yields their "common measure", i.e. a single generator for the additive group they generate. The pair is commensurable if and only if α and β are linearly dependent over the integers, and the algorithm in fact yields an explicit equation of linear dependence. For arbitrary α and β, the coefficients arising in the application of the algorithm are the entries in the continued fraction expansion of β/α, which can be used to construct good rational approximations of this ratio. If β/α is a quadratic irrationality the sequence of coefficients is eventually periodic (proof recalled in §10 below), and from its periodic part one can construct units in the ring of algebraic integers in $Q(\beta/\alpha)$.

Given a family of more than 2 real numbers, $\alpha_1, \ldots, \alpha_n$, one can still use the Euclidean algorithm to find a common measure if they are commensurable: By repeatedly reducing one or another of the terms by a smaller nonzero one, one can eventually get a tuple of terms of arbitrarily small absolute values, unless at some step one's tuple has only one nonzero term, forcing one to stop. If the elements one began with were all commensurable, the process must clearly terminate in the above manner, and the final nonzero term is the common measure.

However, this version of the Euclidean algorithm does not have any of the other properties described above. For instance, if one starts with $(\alpha_1, \alpha_2, \alpha_3) =$

$(1, \sqrt{2}, 1+\sqrt{2})$, it is not hard to show that one can carry out a series of subtractions leading to arbitrarily small 3-tuples, but never one with a 0 term, so that one may never "discover" the linear relation $\alpha_1 + \alpha_2 - \alpha_3 = 0$. (E.g. first decrease α_2 by α_1, then decrease α_3 by six times the modified α_2, etc..)

The problem is that there is too much freedom as to which term to reduce by which. There have been attempts to remedy this by setting arbitrary rules. For instance, in a proposed generalization of the Euclidean algorithm which has been studied for many years, called the Jacobi-Perron algorithm, one starts with an n-tuple of positive real numbers indexed cyclically, and successively subtracts from each term the largest integral multiple of the preceding term which leaves a nonnegative remainder. It was hoped that if one started with an n-tuple of values forming a Q-basis of a number-field K of degree n, the process would become periodic, leading to a construction for units in the ring of algebraic integers in K. But in general, such algorithms turn out to work in some cases and fail in others. Introducing arbitrary regularity into the procedure does not give the same benefits as the essential unicity of the classical case.

See corrected statement on p.52

Recently, however, H. R. P. Ferguson and R. W. Forcade have found an algorithm which will find a $\mathbb{Z}$-linear dependence relation on an n-tuple of real numbers whenever one exists, and their approach changes our understanding of the nature of the whole problem. I found the details of their announcement [1] difficult to follow, but was able to extract two key ideas, and work out a "Ferguson-Forcade type" algorithm, similar in motivation but not in detail to theirs. I will develop it heuristically in §§2-4 below. §2 contains the Lemma which puts the problem in a new perspective. In §3 I apply this idea to produce a relatively simple algorithm which works for n = 3, 4, but unfortunately not for higher n. In §4 we introduce another idea from |1|, and with it construct an algorithm that works for arbitrary n.

The remaining sections explore various consequences and related ideas. In particular, in §6 we strengthen the algorithm, and in §§7, 8 note some consequences concerning subsets, subgroups and subspaces of real vector spaces. §10 contains a proof of the result of Lagrange, that the continued fraction expansion of a real quadratic irrationality is periodic, and recalls how this leads to a construction for units in quadratic number fields. In §11, I discuss the question of whether the present algorithm may have analogous properties for number fields of higher degree. §13 is an appendix in which I prove some curious results about finitely generated dense subgroups of real vector spaces, related to the discussion of §7.

This subject is far from my own field of work, so I would particularly welcome any comments on this material.*

2. Diophantine conditions. Let us write ${}^{p}\mathbb{R}$ for the space of all column-vectors of p real numbers, $\mathbb{R}^{q}$ for the space of row vectors of q real numbers, and ${}^{p}\mathbb{R}^{q}$ for the space of all $p \times q$ matrices of real numbers. We shall consider $\mathbb{R}^{p}$ and ${}^{p}\mathbb{R}$ dual spaces via the obvious pairing.

Given a finite sequence of real numbers $\alpha_1, \ldots, \alpha_n$, equivalently a vector $\alpha \in \mathbb{R}^{n}$, we wish to search for a $\mathbb{Z}$-linear relation

$$\alpha_1 b_{(1)} + \ldots + \alpha_n b_{(n)} = 0 \quad (b_{(i)} \in \mathbb{Z}, \text{ not all } 0). \tag{1}$$

We may generalize this problem by replacing the $\alpha_i \in \mathbb{R}$ by column vectors $\alpha_i \in {}^{m}\mathbb{R}$. We assume these linearly dependent over $\mathbb{R}$, and ask whether they satisfy a linear dependence relation (1) with coefficients in $\mathbb{Z}$. Here without loss of generality we may assume the α_i span ${}^{m}\mathbb{R}$ over $\mathbb{R}$ (since otherwise we can project ${}^{m}\mathbb{R}$ onto $m' < m$ appropriately chosen coordinates and get vectors $\alpha_1', \ldots, \alpha_n'$ satisfying the same linear dependence relations in ${}^{m'}\mathbb{R}$ as the α_i satisfy in ${}^{m}\mathbb{R}$.) In particular, $m < n$; let

$$n = m + r.$$

*I should also mention that this material has expanded to an unexpected extent as I have written it up, and I do not have the time free at present to do a polished job, so I apologize for the uneven exposition.

The given data $\alpha_1, \ldots, \alpha_n$ can be looked at as an $m \times n$ matrix

$$A \in {}^m\mathbb{R}^n.$$

The question of whether the columns of A satisfy a relation (1) can also be formulated in terms of the rows of A. Calling these $a_1, \ldots, a_m$, it asks: does the space spanned by the row vectors $a_1, \ldots, a_m$ lie in a hyperplane $H \subseteq \mathbb{R}^n$ described by a linear relation with integer coefficients?

To look at this in still another way, let us consider the space of all column vectors $b \in {}^n\mathbb{R}$ annihilating all rows of A. It is easy to find an $\mathbb{R}$-basis for these (using the row-reduced echelon form of A), say $b_1, \ldots, b_r \in {}^n\mathbb{R}$. Let us form out of these a matrix

$$B \in {}^n\mathbb{R}^r.$$

Then the question of whether a relation (1) with integer-valued $b_{(1)}, \ldots, b_{(n)}$ holds is equivalent to asking whether some nonzero linear combination of the columns of B has integer entries. Now a general linear combination of the columns of B has the form Bc ($c \in {}^r\mathbb{R}$). The vector c represents a linear functional on $\mathbb{R}^r$, so the given question takes on a fourth form: Does there exist a nonzero linear functional on $\mathbb{R}^r$ which gives integer values on all the rows of B? To summarize the above observations, let us make the following definitions:

A linear Diophantine relation satisfied by a family of elements of a real vector space will mean a $\mathbb{Z}$-linear dependence relation satisfied by these elements.

A Diophantine point of ${}^n\mathbb{R}$ (or $\mathbb{R}^n$) will mean a point with integer coordinates.

A Diophantine hyperplane in $\mathbb{R}^n$ (resp. ${}^n\mathbb{R}$) will mean the annihilator of a Diophantine point of ${}^n\mathbb{R}$ (resp. $\mathbb{R}^n$).

A linear co-Diophantine relation satisfied by a family of elements of a

real vector space will mean the condition that some specified nonzero real linear functional on the space assume integer values on all these elements.

Lemma 1. Let m, r be positive integers, and $A \in {}^{m}\mathbb{R}^{m+r}$, $B \in {}^{m+r}\mathbb{R}^{r}$ be matrices such that the rows of A form a basis for the left annihilator space of B, and the columns of B form a basis for the right annihilator space of A. Then the following four conditions are equivalent.

(i) The columns of A, $\alpha_1, \ldots, \alpha_{m+r}$, satisfy a linear Diophantine relation.

(ii) The vector space spanned by the rows of A, $a_1, \ldots, a_m$, lies in a Diophantine hyperplane in $\mathbb{R}^{r+m}$.

(iii) The vector space spanned by the columns of B, $b_1, \ldots, b_r$, contains a Diophantine point of ${}^{m+r}\mathbb{R}$.

(iv) The rows of B, $\beta_1, \ldots, \beta_{m+r}$, satisfy a linear co-Diophantine relation. ||

How does one picture condition (iv)? If $c \in {}^{r}\mathbb{R}$ is a fixed nonzero vector, then the solution-set of the associated co-Diophantine relation,

$$\{b \in \mathbb{R}^r \mid b\,c \in \mathbb{Z}\} \tag{2}$$

will be the union of a family of regularly spaced cosets of the hyperplane

$$\{b \mid b\,c = 0\}. \tag{3}$$

Clearly if $\beta_1, \ldots, \beta_n$ lie in the set (2), so does the whole additive group G that they generate. Note also that from our hypotheses relating A and B, it can be deduced that $\beta_1, \ldots, \beta_n$ span $\mathbb{R}^r$ as a vector space, so G will not lie entirely in the hyperplane (3).

Suppose now that we try to modify the generating set $\beta_1, \ldots, \beta_n$ of G, by adding to one β_i an integral linear combination of the others, and iterating

this process with the aim of getting "smaller" generators. Clearly there will always remain at least one generator not on the hyperplane (3), hence we can never get the maximum of the lengths of these generators less than the distance between successive cosets comprising (2).

What Ferguson and Forcade show by explicit construction is that conversely, if $\beta_1, \ldots, \beta_n$ do not satisfy a co-Diophantine relation, then one can get arbitrarily "small" generating sets for G. We shall do this in the next two sections.

When these constructions are applied to a family of vectors which do satisfy a co-Diophantine relation, they terminate by giving a family $\beta_1, \ldots, \beta_n$ such that

(4) all but one of the β_i lie in a common proper hyperplane (3).

In this case the linear functional c, normalized to have the value 1 on the exceptional β_i, yields the desired co-Diophantine relation.

In §§7,8 we shall note consequences and interpretations of this result in terms of all four viewpoints of Lemma 1.

Note that two points α_1, α_2 of the one-dimensional space $\mathbb{R}$ satisfy a linear Diophantine relation if and only if they satisfy a linear co-Diophantine relation. This is why the type of procedure that must be applied to B in the general case can in the special case $m = r = 1$ of the classical Euclidean algorithm be applied directly to A.

Remarks: Ferguson and Forcade |1| consider only the case $m = 1$. The equivalence (i) $\Leftrightarrow$ (iv) of Lemma 1 above is implicit in their method. I have introduced the additional conditions (ii) and (iii), and allowed general m, for the sake of symmetry and beauty. However, in §§3-5 we will also restrict ourselves

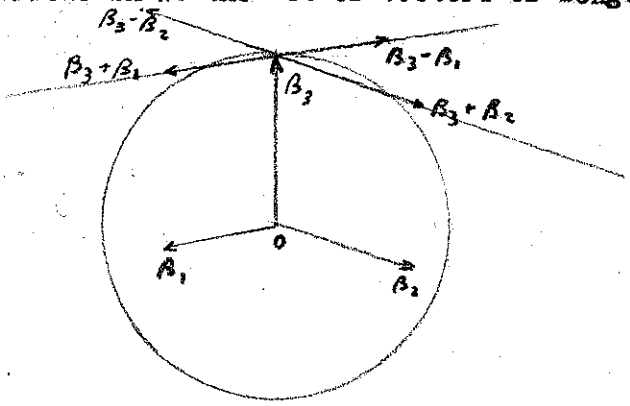
to the case m = 1 for simplicity, and work only with the viewpoint of (iv), i.e. that of trying to find co-Diophantine relations. In §6 the algorithm will be generalized to arbitrary m, and in §7 we consider consequences expressed in terms of the viewpoints of conditions (i)-(iii).

We also mention that as a measure of the "size" of a vector which they seek to reduce in the course of their algorithm, Ferguson and Forcade use the maximum of the absolute values of the coordinates. Here we shall instead us the <u>length</u> of the vector under the standard inner product norm on $\mathbb{R}^r$; we shall write $|x|$ for the length of x.

To avoid excessive complications with indices, I shall regularly abuse notation by using the same symbols for a system of vectors with which we start, e.g. $\beta_1, \ldots, \beta_{r+1}$, and the systems into which it is transformed at various stages of our algorithms. But these will be distinguished by context, e.g. "the original system $\beta_1, \ldots, \beta_{r+1}$", "the value of β_i after this step", etc.. Occasionally, when there is a need to distinguish these symbolically, I will do so using primes or superscripts.

3. <u>Algorithms for small r.</u> Given $r+1$ vectors in $\mathbb{R}^r$, we wish to investigate whether they satisfy a linear co-Diophantine relation, by the method indicated above.

Consider the case of three vectors, β_1, β_2, β_3 in the plane. Say $|\beta_1| \leq |\beta_2| \leq |\beta_3|$. Now if β_1 and β_2 are colinear (pictured as arrows from O), then we already have the situation (4), which gives a co-Diophantine relation. So assume the contrary. It still may not be possible to reduce the length of β_3 by subtracting an integral multiple of β_1 or β_2. This can be seen in the illustration below, where the circle shows the set of vectors of length equal to that of β_3:



However, I claim that it <u>is</u> possible to reduce $|\beta_3|$ by subtracting some $\mathbb{Z}$-linear combination of β_1 and β_2. For consider the lattice $\mathbb{Z}\beta_1 + \mathbb{Z}\beta_2$. This dissects $\mathbb{R}^2$ into parallelograms with sides β_1 and β_2, and the vector β_3 will lie in (or on the boundary of) one of these. (To find this parallelogram computationally, one inverts the nonsingular 2×2 matrix $\binom{\beta_1}{\beta_2}$; the inverse matrix converts the lattice in question to the lattice $\mathbb{Z}^2 \subseteq \mathbb{R}^2$, and the square of this lattice containing the image of β_3 may be found from the integer parts of the two coordinates of this point.) Now the distance from β_3 to the <u>nearest</u> of the four vertices of the parallelogram containing it will be $\leq (\sqrt{2}/2)\,|\beta_2|$. We omit the details, simply noting that the worst case is when the paralleogram is a square and β_3 is its center.

Hence, if we write $n_1\beta_1 + n_2\beta_2$ for this lattice point, and $\beta_3' = \beta_3 - n_1\beta_1 - n_2\beta_2$, then $|\beta_3'| / |\beta_3| \leq |\beta_3'| / |\beta_2| \leq \sqrt{2}/2$. It is easily deduced that $(|\beta_1| + |\beta_2| + |\beta_3'|) / (|\beta_1| + |\beta_2| + |\beta_3|) \leq (2 + 2 + \sqrt{2})/(2 + 2 + 2)$. Hence by iterating this process, we can make $|\beta_1| + |\beta_2| + |\beta_3|$ arbitrarily small, or else discover a linear co-Diophantine relation.

The same argument works for $r = 3$; in our estimate we get $\sqrt{3}/2$ in place of $\sqrt{2}/2$; as this is still < 1, it is satisfactory. But for $r \geq 4$ we get $\sqrt{r}/2 \geq 1$, and the argument fails.

(Actually, in the marginal case $r = 4$ it <u>can</u> be saved. One first notes that if the lattice $\beta_1\mathbb{Z} + \ldots + \beta_4\mathbb{Z}$ differs greatly from a regular cubic lattice, or if β_5 deviates much from the center of its cell in this lattice, one can get a bound < 1 for $|\beta_5'| / |\beta_5|$. If the lattice <u>is</u> close to cubic and β_5 does lie near the center of its cell, but if this cell does not have 0 as a vertex, we can decrease the length of β_5 by translating it to a nearer cell. This leaves the

case where the cells are approximately cubic, and β_5 lies near the center of a cell with one vertex at 0. Thus (up to signs), $\beta_5 \approx \frac{1}{2}(\beta_1 + \beta_2 + \beta_3 + \beta_4)$. Now if we set $\beta_4' = \beta_1 + \beta_2 + \beta_3 + \beta_4 - 2\beta_5$, we have again reduced $\sum |\beta_i|$ by a factor which we can bound below 1. But we do not wish to go on looking for such ad hoc arguments for each higher value of r.)

See p.53 note 2 re earlier discovery of above algorithm, and same page, note 3, for correction to next three paragraphs.

Thus we need a different approach.

4. A general construction. Let r now be an arbitrary positive integer. Given a family of vectors $\beta_1, \ldots, \beta_{r+1} \in \mathbb{R}^r$, we seek an algorithm which, by repeatedly modifying one and another member of this system by linear combinations of the others, either produces systems of vectors of arbitrarily small lengths, or terminates in a situation (4), establishing a co-Diophantine relation.

The second key idea which I adapt from Ferguson and Forcade is to begin as follows. Assume inductively that such an algorithm exists for r-1. Let β_1 be the shortest of the vectors β_i, project $\beta_2, \ldots, \beta_{r+1}$ onto the orthogonal complement of the subspace of $\mathbb{R}^r$ spanned by β_1, and apply the (r-1)-algorithm to these projections.

We remark again that Ferguson and Forcade use the max norm rather than the length function. Correspondingly, rather than project onto the orthogonal complement of $\beta_1\mathbb{R}$, they let the largest coordinate of β_1 be the j^{th}, and project $\beta_2, \ldots, \beta_{r+1}$ to their r-1 coordinates other than the j^{th}.

Further, while Ferguson and Forcade follow precisely the approach just stated, I find it more elegant to give a single construction in which the algorithms for smaller dimensions are implicitly incorporated.

To describe it, we will use the following notation. If $\beta_1, \ldots, \beta_{r+1}$ are vectors in $\mathbb{R}^r$ no proper subset of which are linearly dependent, define $S^{(i)}$ $(0 \leq i \leq r)$ to be the orthogonal complement in $\mathbb{R}^r$ of the space spanned by $\beta_1, \ldots, \beta_i$. Thus

$$\mathbb{R}^r = S^{(0)} \supset \ldots \supset S^{(r)} = \{0\}.$$

For any vector x, we let $x^{(i)}$ denote the orthogonal projection of x in $S^{(i)}$. Note that for $i < j$, the orthogonal complement in $S^{(i)}$ of the space spanned by $\beta_{i+1}^{(i)}, \ldots, \beta_j^{(i)}$ is $S^{(j)}$, and for any x, $(x^{(i)})^{(j)} = x^{(j)}$. In our construction, whenever we modify a system of vectors $\beta_1, \ldots, \beta_{r+1}$, it will be understood that the system of subspaces $S^{(i)}$, is modified accordingly. It will also be understood that if our system of $r+1$ vectors ever has the property that the first r are linearly dependent, then we have a co-Diophantine relation and can terminate our construction. Hence we may always assume the contrary is the case in describing a step of the construction.

Our algorithm will consist of the alternate application of two steps:

"Adjustment". The idea of this step will be to add to each β_k $(1 < k \leq r+1)$ that integral linear combination of $\beta_1, \ldots, \beta_{k-1}$ which will minimize the distance from $\beta_k^{(j-1)}$ to $S^{(j)}$ (i.e. to $\beta_k^{(j)}$) for $j = 1, \ldots, k-1$. Note that

(5) Adding any multiple of β_j to β_k $(j < k)$ does not change the system of subspaces $S^{(0)}, \ldots, S^{(r)}$, nor the projections $\beta_i^{(i-1)}$ $(1 \leq i \leq r+1)$.

A precise statement of the operation is as follows. Letting j run through the values $r, r-1, \ldots, 1$ in that order, we successively

(6) replace every vector β_k $(j < k \leq r+1)$ by $\beta_k - n_{kj}\beta_j$,

where

$$(7) \qquad n_{kj} = \left\{ \frac{\beta_k^{(j-1)} \cdot \beta_j^{(j-1)}}{\beta_j^{(j-1)} \cdot \beta_j^{(j-1)}} \right\}.$$

Here dots denote inner product of vectors, and brackets denote the "nearest integer" function (with "rounding down" when the argument is a half-integer.) We see that after an application of (6) for some value of j, the new vectors β_k $(j < k \leq r+1)$ will satisfy $\left| \frac{\beta_k^{(j-1)} \cdot \beta_j^{(j-1)}}{\beta_j^{(j-1)} \cdot \beta_j^{(j-1)}} \right| \leq \frac{1}{2}$, which can be translated as:

(8) The perpendicular projection of $\beta_k^{(j-1)}$ on $\beta_j^{(j-1)}$, namely $\beta_k^{(j-1)} - \beta_k^{(j)}$, is at most half the length of $\beta_j^{(j-1)}$.

Further, the condition (8) is not disturbed when (6) is subsequently applied with j replaced by values $i < j$, since β_i has zero projection on $S^{(j-1)}$. Hence after applying (6) successively for $j = r, \ldots, 1$, we get a system of vectors which satisfies (8) for all $j < k$. Let us make the

(9) Definition. A system of vectors $\beta_1, \ldots, \beta_{r+1} \in \mathbb{R}^r$ such that (8) holds for $1 \leq j < k \leq r+1$ will be called adjusted.

Now roughly speaking, it is best to perform this operation of adjustment on systems $\beta_1, \ldots, \beta_{r+1}$ having smaller vectors to the left. The process itself tends to make the vectors on the right smaller. Hence we follow each adjustment step by one of

Reindexing. The principle here is

(10) If $|\beta_i^{(i-1)}| > |\beta_{i+1}^{(i-1)}|$, interchange the labels of β_i and β_{i+1}.

However, things are complicated by the fact that there may be more than one value of i for which the hypothesis of (10) holds, and (as we shall show by an example at the end of the next section) the effect of a choice of which interchange

to perform first cannot always be neutralized by later choices. So I shall simply state one strategy of reindexing for which our proof of the algorithm will go through, and we shall note in the next section that this can be varied in many ways:

(11) Perform (10) successively for $i = r, r-1, \ldots, 1$ (then stop).

We claim that the operations of adjustment and reindexing described above, performed alternately on an $(r+1)$-tuple $\beta_1, \ldots, \beta_{r+1}$, will either terminate in a situation where some $\beta_1, \ldots, \beta_i$ $(i \leq r)$ are linearly dependent (so that, if i is the least such value, $\beta_i^{(i-1)} = 0$), giving a co-Diophantine relation, or will produce systems of arbitrarily small vectors.

We first note that for an adjusted system of vectors, the lengths of the β_k can be estimated in terms of the lengths of the $\beta_j^{(j-1)}$ $(j \leq k)$. Indeed, from (8) and the Pythagorean Theorem we have

$$|\beta_k|^2 \leq \tfrac{1}{4}|\beta_1^{(0)}|^2 + \ldots + \tfrac{1}{4}|\beta_{k-1}^{(k-2)}|^2 + |\beta_k^{(k-1)}|^2 \quad (1 \leq k \leq r+1). \tag{12a}$$

Ignoring the "1/4"'s for simplicity, we get the bound

$$\sum |\beta_k|^2 \leq (r+1)|\beta_1^{(0)}|^2 + r\,|\beta_2^{(1)}|^2 + \ldots + |\beta_{r+1}^{(r)}|^2. \tag{12}$$

Let us make the

(13) Definition. The right hand side of (12) will be denoted $m(\beta_1, \ldots, \beta_{r+1})$.

We shall prove that if our algorithm continues without terminating, this function converges to zero. We first note (cf.(5)):

(14) An operation (10) changes none of the spaces $S^{(k)}$ except $S^{(i)}$, and none of the projections $\beta_k^{(k-1)}$ except $\beta_i^{(i-1)}$ and $\beta_{i+1}^{(i)}$.

Assuming the hypothesis of (10) satisfied, let us describe the effect of that operation explicitly by writing $\beta_i' = \beta_{i+1}$, $\beta_{i+1}' = \beta_i$, $S^{(i)'} =$ orthogonal complement of space spanned by $\beta_1, \ldots, \beta_{i-1}, \beta_i'$. If we let s denote the sine of the angle between $\beta_i^{(i-1)}$ and $\beta_{i+1}^{(i-1)}$ (which are also $\beta_{i+1}'^{(i-1)}$ and $\beta_i'^{(i-1)}$) in $S^{(i-1)}$, then we have $|\beta_{i+1}^{(i)}| = |s|\ |\beta_{i+1}^{(i-1)}|$. We now compare the lengths of the i^{th} and $i+1^{st}$ of the $\beta_k^{(k-1)}$ before and after (10):

(15)

old lengths	new lengths
$\lvert\beta_i^{(i-1)}\rvert$	$\lvert\beta_i'^{(i-1)}\rvert = \lvert\beta_{i+1}^{(i-1)}\rvert$
$\lvert\beta_{i+1}^{(i)}\rvert = \lvert s\rvert\ \lvert\beta_{i+1}^{(i-1)}\rvert$	$\lvert\beta_{i+1}'^{(i)},\rvert = \lvert s\rvert\ \lvert\beta_{i+1}'^{(i-1)}\rvert = \lvert s\rvert\ \lvert\beta_i^{(i-1)}\rvert$.

Were it not for the factor $|s|$, we see that the sum of the old values would equal the sum of the new values. Since $|s| \leq 1$ and $|\beta_{i+1}^{(i-1)}| \leq |\beta_i^{(i-1)}|$, the sum of the new values is actually less than or equal to the sum of the old values. The same argument applies to the squares of the lengths, giving

(16) When the hypothesis of (10) is satisfied, the operation there described decreases $|\beta_i^{(i+1)}|^2$, and increases $|\beta_{i+1}^{(i)}|^2$ by at most the same amount.

Since in the definition of $m(\beta_1, \ldots, \beta_{r+1})$, the summand $|\beta_i^{(i-1)}|^2$ has a larger coefficient than $|\beta_{i+1}^{(i)}|^2$ (larger by 1), and since the value which replaces $|\beta_i^{(i-1)}|^2$ after (10) is the former value of $|\beta_{i+1}^{(i-1)}|^2$, we get from (16)

(17) When the hypothesis of (10) is satisfied, the operation there described <u>decreases</u> the value of $m(\beta_1, \ldots, \beta_{r+1})$ by at least the value of $|\beta_i^{(i-1)}|^2 - |\beta_{i+1}^{(i-1)}|^2$ before the application of the operation.

To conclude that $m(\beta_1, \ldots, \beta_{r+1})$ converges to 0, we need to know that when we follow (11), at least one fairly large decrease occurs. Now one can show that for $i = r$, the hypothesis of (10) is always satisfied, and that $|\beta_{r+1}^{(r-1)}| \leq \frac{1}{2} |\beta_r^{(r-1)}|$. (Indeed, since $S^{(r-1)}$ is one-dimensional, "adjustment" for $j=r$, $k=r+1$ is like a step of the ordinary Euclidean algorithm.) However, these numbers may be very small compared with some of the other $|\beta_i^{(i-1)}|$, and so not cause $m(\beta_1, \ldots, \beta_{r+1})$ to decrease by any stated fraction of its original value. On the other hand, if we look at the largest values of $|\beta_i^{(i-1)}|$, though these do contribute a large fraction of $m(\beta_1, \ldots, \beta_{r+1})$, the differences $|\beta_i^{(i-1)}|^2 - |\beta_{i+1}^{(i-1)}|^2$ may be small, if the hypothesis of (10) is satisfied at all. To get a satisfactory compromise, let us choose i so as to maximize $2^i |\beta_i^{(i-1)}|$. This i will be less than $r+1$ (since $S^{(r)} = 0$ and hence $\beta_{r+1}^{(r)} = 0$). Then

$$|\beta_{i+1}^{(i)}| \leq \frac{1}{2} |\beta_i^{(i-1)}|. \tag{18}$$

If the system $\beta_1, \ldots, \beta_{r+1}$ is adjusted (9), or more generally if

(19) Other applications (10) performed since the preceding adjustment have left intact the case $j = i$, $k = i+1$ of (8),

then we may conclude by the Pythagorean theorem that

$$\begin{aligned} |\beta_{i+1}^{(i-1)}|^2 &\leq |\beta_{i+1}^{(i)}|^2 + |\beta_{i+1}^{(i-1)} - \beta_{i+1}^{(i)}|^2 \\ &\leq (\tfrac{1}{2} |\beta_i^{(i-1)}|)^2 + (\tfrac{1}{2} |\beta_i^{(i-1)}|)^2 = \tfrac{1}{2} |\beta_i^{(i-1)}|^2. \end{aligned} \tag{20}$$

Hence the hypothesis of (10) is satisfied for this i, and the amount by which $m(\beta_1, \ldots, \beta_{r+1})$ is changed is at least $\frac{1}{2} |\beta_i^{(i-1)}|^2$. But

$$(21) \qquad |\beta_i^{(i-1)}| = \frac{1}{2^i} \max(2^j |\beta_j^{(j-1)}|) \geq \frac{1}{2^r} \max(|\beta_j^{(j-1)}|),$$

while

$$(22) \qquad m(\beta_1, \ldots, \beta_{r+1}) \leq \frac{(r+1)(r+2)}{2} (\max |\beta_j^{(j-1)}|))^2.$$

We can now get the needed estimate. Starting with an adjusted system $\beta_1, \ldots, \beta_r$, we choose i as above to maximize $2^i |\beta_i^{(i-1)}|$. Thus from (21) and (22) we can deduce

$$(23) \qquad \frac{1}{2} |\beta_i^{(i-1)}|^2 \geq \frac{1}{4^r(r+1)(r+2)} m(\beta_1, \ldots, \beta_{r+1}).$$

Now we start reindexing according to (11), until our chosen value of i is reached. The previous reindexings will not have changed $\beta_i^{(i-1)}$, and can at most decrease $m(\beta_1, \ldots, \beta_{r+1})$, so (23) remains true. Further, if β_{i+1} has been changed at all, it has been replaced by some β_k, $k > i$, so (19) still holds, and we get (20). Hence for the chosen value of i the hypothesis of (10) holds and by (17), $m(\beta_1, \ldots, \beta_{r+1})$ is decreased by at least the amount described by (23). Subsequent applications of (10) can only decrease it further. So if we let the reindexed system after application of (11) be denoted $\beta_1', \ldots, \beta_{r+1}'$, we have

$$(24) \qquad m(\beta_1', \ldots, \beta_{r+1}') \leq \left(1 - \frac{1}{4^r(r+1)(r+2)}\right) m(\beta_1, \ldots, \beta_r).$$

Clearly, then, the function m and hence our system of vectors converges to 0.

5. Variants, and related observations. Let us note some modifications of (11) for which the above argument, with the same estimate (24), is still valid.

We might replace (11) by

(25) Find a value of i maximizing $2^i|\beta_i^{(i-1)}|$, and interchange the labels of β_i and β_{i+1}.

This in fact would simplify the final argument leading to (24). Alternatively, we could change the words "(then stop)" at the end of (11) to "and iterate this until no such values of i remain". This would terminate after finitely many iterations, since it is decreasing on $m(\beta_1, \ldots, \beta_{r+1})$, and there are only finitely many permutations of the β_i's.

It might appear that we could make a simplification in the "adjustment" step corresponding to (25), getting a combined "adjustment and reindexing" operation of the form

(26) Find the value of i maximizing $2^i|\beta_i|$, replace β_{i+1} by $\beta_{i+1} - n_{i+1,\,i}\,\beta_i$, then interchange the labels of β_i and β_{i+1}.

This would indeed get $m(\beta_1, \ldots, \beta_{r+1})$ to converge to 0 according to (24); the difficulty is that the estimate (12) is only valid for adjusted systems. However, since adjustment does not change the value of $m(\beta_1, \ldots, \beta_{r+1})$, we could iterate (26) till we had $m(\beta_1, \ldots, \beta_{r+1})$ as small as we liked, then perform a "complete adjustment" at the end. This is not an iterative algorithm in the strictest sense, of course. Whether it is a better algorithm than the one described I don't know. It would involve fewer arithmetic operations,

but might also involve the build-up of large numbers.

My feeling is that the algorithm with complete adjustment as described in the last section, but with reindexing iterated as long as possible, is the least arbitrary, and so may in some way be the best. If we divide our estimate of its rate of convergence by the number of arithmetic operations performed, the result is of course far poorer than for (25). However, it's actual convergence may be much better than the estimate, especially for "typical" rather than "worst" cases. This might be studied by working some examples on a computer.

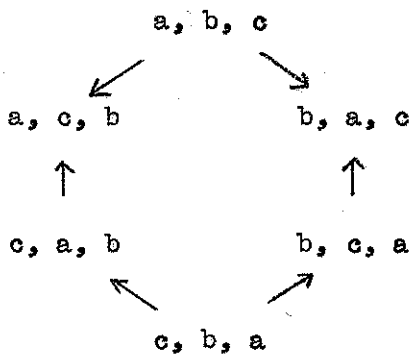
We remark that the same construction can be applied in $\mathbb{C}^r$, with the Gaussian integers $\mathbb{Z}[i]$ in place of $\mathbb{Z}$. The one significant change is that $1/\sqrt{2}$ replaces $1/2$ in (8) and the sentence immediately preceding. This changes the coefficient on the right hand side of (20) from $(\frac{1}{2})^2 + (\frac{1}{2})^2 = \frac{1}{2}$ to $(\frac{1}{2})^2 + (\frac{1}{\sqrt{2}})^2 = \frac{3}{4}$; hence the coefficient we want to put on the left-hand-side in (23) is not $1 - \frac{1}{2} = \frac{1}{2}$ but $1 - \frac{3}{4} = \frac{1}{4}$, and the final effect is to put an extra factor of 2 into the denominator occurring in (24); i.e. the convergence as estimated is essentially half as fast. Similarly if ω is a primitive cube root of unity, and we use $\mathbb{Z}[\omega]$ in place of $\mathbb{Z}$, we get $1/\sqrt{3}$ in place of $1/2$ in (8), $\frac{1}{4} + \frac{1}{3} = \frac{7}{12}$ in (20), and convergence essentially $(1 - \frac{7}{12})/(1 - \frac{1}{2}) = 5/6$ as fast as for $\mathbb{Z}$. One can also put a polynomial ring $k[t]$ in place of $\mathbb{Z}$ and the formal Laurent series field $k(\!(t)\!)$ in place of $\mathbb{R}$; but because of the ultrametric absolute value, one must make greater changes in the formalism of the construction, and we will not discuss this case.

We remark that when we chose i to maximize $2^i\,|A_i^{(i-1)}|$, the factors 2^i were chosen for simplicity, but a different system of coefficients would have

led to different estimates. If we replace 2^i by c^i, I believe the estimates improve (for large r) as c decreases toward $2/\sqrt{3}$. When $c = 2/\sqrt{3}$, the analog of (20) gives the useless estimate $|\beta_{i+1}^{(i-1)}| \leq |\beta_i^{(i-1)}|$. However, if we use not just the fact that for the chosen value of i, $(2/\sqrt{3})^i|\beta_i^{(i-1)}| - (2/\sqrt{3})^{i+1}|\beta_{i+1}^{(i)}| \geq 0$, but the fact that the sum of all the positive values assumed by $(2/\sqrt{3})^j|\beta_j^{(j-1)}|^2 - (2/\sqrt{3})^{j+1}|\beta_{j+1}^{(j)}|^2$ for $j \geq i$ is at least $(2/\sqrt{3})^i|\beta_i^{(i-1)}|$, I believe we get quite a good estimate. The possibility of improving our estimates by modifying the indicator function m should also be looked into.

Here is the promised example showing that a choice of which operation (10) to perform first may make a difference. Let $a = (3,0)$, $b = (0,2)$, $c = (4,1)$ in $\mathbb{R}^2$. Then the arrows in the diagram below show the operations on systems $\beta_1, \beta_2, \beta_3$ composed of these three vectors in some order, allowed by the hypothesis of (10).

(27)



Thus, starting at a, b, c (or at c, b, a) one can arrive at either of the two "stable" systems.

(Given $r \geq 2$, suppose G_{r+1} denotes the graph with $(r+1)!$ vertices corresponding to the permutations of $1, \ldots, r+1$, and an edge connecting each pair of permutations which differ by a transposition of two adjacent terms. E.g., G_3 is a hexagon. It would be amusing to try to determine what systems of

orientations of the edges of G_{r+1} can be realized as in (27) by allowed operations (10) on some system of vectors $\beta_1, \ldots, \beta_{r+1} \in \mathbb{R}^r$. There are two obvious restrictions: there can be no cycles by (17), and a second condition arises, for $r+1 \geq 4$, from the observation that whether one transposes a given two adjacent terms does not affect whether one may transpose a pair of adjacent terms disjoint from them.)

In connection with the "low-dimensional" construction of §3, it would be interesting to know what is the least integer, r_0, such that there exist nonzero vectors $\beta_1, \ldots, \beta_{r_0+1} \in \mathbb{R}^{r_0}$ such that none of the lengths $|\beta_i|$ can be decreased by subtracting from β_i a $\mathbb{Z}$-linear combination of $\{\beta_j \mid j \neq i\}$. (From the parenthetical paragraph at the end of that section we can say that $r_0 > 4$. On the other hand, one can see that $r_0 \leq 14$ by considering in $\mathbb{R}^{14}$ the 14 standard basis vectors $\beta_1 = (1,0,\ldots,0)$, ..., $\beta_{14} = (0,\ldots, 0, 1)$, and a vector β_{15} having five coordinates equal to $2/10$, four coordinates equal to $3/10$, and five coordinates equal to $4/10$.)

6. Recycling the algorithm. Suppose $\beta_1, \ldots, \beta_{r+1}$ span $\mathbb{R}^r$, and satisfy a linear co-Diophantine relation. Then if we apply the algorithm of §4 to this system, we eventually obtain a system such that for some $r' < r$, $\beta_{r'+1}^{(r')} = 0$, i.e. such that $\beta_{r'+1}$ lies in the subspace $T \subseteq \mathbb{R}^r$ spanned by $\beta_1, \ldots, \beta_{r'}$. But let us not stop there. Note that since our system involves just one more vector than the dimension of $\mathbb{R}^r$, T must be exactly r'-dimensional. Hence identifying T with $\mathbb{R}^{r'}$, we can now apply the algorithm to this system of $r' + 1$ elements. If these satisfy a co-Diophantine relation in T, we eventually obtain in the same manner a system of still fewer vectors. This process can be repeated only finitely many times. Hence our algorithm, iterated in this manner,

eventually transform any system of vectors $\beta_1, \ldots, \beta_{r+1}$ spanning $\mathbb{R}^r$ into one such that for some s $(0 \leq s \leq r)$, $\beta_1, \ldots, \beta_{s+1}$ span an s-dimensional subspace U, but satisfy no linear co-Diophantine relation in U. Continued application of the algorithm to $\beta_1, \ldots, \beta_{s+1}$ will lead to systems of $s+1$ vectors of arbitrarily small lengths.

In the above situation, (indeed, even without the assumption that we have reached the final stage, where $\beta_1, \ldots, \beta_{s+1}$ satisfy no co-Diophantine relation), we can see by looking at dimensions that $\beta_{s+2}, \ldots, \beta_{r+1}$ will be linearly independent modulo U, so that their projections $\beta_{s+2}^{(s-1)}, \ldots, \beta_{r+1}^{(s-1)}$ on the orthogonal complement $S^{(s-1)}$ of U will form a basis of $S^{(s-1)}$.

Now suppose that we have in our pocket an $r+2^{nd}$ vector, β_{r+2}, which we have not done anything with up to this point. If we bring in its projection $\beta_{r+2}^{(s+1)}$, then we have $r-s+1$ vectors $\beta_{s+2}^{(s+1)}, \ldots, \beta_{r+2}^{(s+1)}$ spanning the $(r-s)$-dimensional space $S^{(s+1)}$, and we can apply the algorithm to these. This process can, in turn, be repeated if more additional vectors $\beta_{r+3}, \ldots, \beta_{r+m}$ are given.

The process sketched above involves a discrete succession of applications of the algorithm of §4. But that algorithm can, in fact, be modified in a very minor way so as to do all this at once. So let us consider the above remarks as heuristic only, and now go on to a precise description of the extended algorithm.

7. The extended algorithm. Given any system of vectors $\beta_1, \ldots, \beta_{r+m}$ spanning $\mathbb{R}^r$ $(m, r \geq 0)$, let us, as in §4, define $S^{(i)}$ $(0 \leq i \leq r+m)$ to be the orthogonal complement in $\mathbb{R}^r$ of the subspace spanned by $\beta_1, \ldots, \beta_i$, and for any $x \in \mathbb{R}^r$ define $x^{(i)}$ to be the projection of x on $S^{(i)}$. Thus, $\mathbb{R}^r = S^{(0)} \supseteq \ldots \supseteq S^{(r+m)} = \{0\}$; but note that in this general case not all the inclusions will be strict.

In fact, for precisely m values of i we must have $S^{(i-1)} = S^{(i)}$, equivalently, $\beta_i^{(i-1)} = 0$. Let us call the index i weak if the above equalities hold, strong otherwise. Those indices (necessarily strong if any) which are greater than all weak indices will be called the terminal indices. The status of an index will, of course, change as we modify our system.

For such a system of vectors, let us define the process of adjustment exactly as in (6), except that k now ranges only over non-terminal indices, an j only over the strong indices preceding k. The second restriction is clearly needed, because (7) is undefined when j is weak. The first is included more for reasons of elegance: We shall see that the vectors indexed by terminal indices cannot be made arbitrarily small; so we may as well keep them fixed, rather than changing them slightly at each step. Note that after adjustment, (8) will hold for all nonterminal k and all j: If j is strong it holds because of the effect of the adjustment as in §4; if j is weak, because both lengths are 0.

The operation of reindexing we define as in (10) and (11) with only the restriction (again for elegance) that i be nonterminal. Note that the hypothesis of (10) can never be satisfied when i is weak, since then the left-hand side of the indicated inequality is 0.

As in §4, adjustment and reindexing are to be performed alternately.

Let us see when, in this process, the status (weak vs. strong) of an index can change. This cannot happen at an adjustment step by (5). At a step (10), it still cannot happen if both i and i+1 are strong. For in that case, the space spanned by $\beta_1, \ldots, \beta_{i+1}$ is of dimension 2 more than that spanned by $\beta_1, \ldots, \beta_{i-1}$, i.e. the dimension goes up by 1 when each of β_i, β_{i+1} is added. This property is retained if we reverse the order of the two vectors, hence after the reordering the indices remain strong, and clearly the status of no earlier or later index is affected (14). We have seen that the hypothesis of (10) cannot be

satisfied if i is weak. There remains only the case where i is strong and $i+1$ is weak. This means that $\beta_i^{(i-1)}$ is nonzero, and $\beta_{i+1}^{(i-1)}$ is linearly dependent thereon. Now if $\beta_{i+1}^{(i-1)}$ is also nonzero, then after reversing the labels of β_i and β_{i+1}, i will again be strong, and $i+1$ again weak. But if $\beta_{i+1}^{(i-1)} = 0$ then after this reindexing step i is weak and $i+1$ strong. Thus we see that the only way indices change status during our algorithm is by the location of a weak index moving one step to the left.

Note in particular that once an index i has become terminal, its status cannot change, and the vector β_i itself remains fixed as well.

Now say that at a particular stage in this process, $s+m$ is the last weak index, $(0 \leq s \leq r)$, and we let U denote the subspace spanned by $\beta_1, \ldots, \beta_{s+m}$. Thus, U will be s-dimensional, and the remaining $r-s$ vectors $\beta_{s+m+1}, \ldots, \beta_{r+m}$ will span $\mathbb{R}^r$ over U, and be linearly independent module U. Thus, their projections $\beta_{s+m+1}^{(s+m)}, \ldots, \beta_{r+m}^{(s+m)}$ on $S^{(s+m)}$ will be a basis thereof, and hence generate a discrete additive subgroup. Hence the subgroup of $\mathbb{R}^r$ generated by all of $\beta_1, \ldots, \beta_{r+m}$ will lie in the discrete union of cosets of U by the elements of that free abelian group of rank $r-s$. This is a generalization of the picture that we gave of a co-Diophantine relation, following Lemma 1. In fact, each terminal index corresponds to an independent co-Diophantine relation satisfied by our system of vectors, and whenever the last weak index moves a step to the left, this corresponds to the discovery of another such relation.

We now claim that as we apply our algorithm, the maximum length of the vectors associated with <u>nonterminal</u> indices goes to 0. To see this, we define

$$m(\beta_1, \ldots, \beta_{r+m}) = \sum_{i \text{ nonterminal}} (r+m+1-i)|\beta_i^{(i-1)}|^2. \tag{28}$$

Then precisely the same calculations as before show that if this is nonzero, it decreases at each reindexing step by a factor which we can bound (by (24) with r+m-1 in place of r). The key point to note is that because we have restricted the sum (28) to nonterminal i, the last term $|\beta_i^{(i-1)}|^2$ which appears will correspond to the last weak index i, and hence will be zero. Hence the term which maximizes $2^i|\beta_i^{(i-1)}|$ will not be the last term, and our argument about the effect of the next reindexing still holds.

Now if we apply our algorithm long enough to a given system of vectors, the position of the last weak index s+m must eventually stabilize. The lengths of $\beta_1, \ldots, \beta_{s+m}$ will thus tend to 0, hence these vectors cannot satisfy any linear co-Diophantine relation in the space they span. Thus all linear co-Diophantine relations satisfied by the original system have been determined. Specifically, if we let G denote the additive subgroup generated by the given vectors, which is not changed by our algorithm, then all linear functionals $f: \mathbb{R}^r \to \mathbb{R}$ satisfying $f(G) \subseteq \mathbb{Z}$ are determined by specifying arbitrary integer values on $\beta_{s+m+1}, \ldots, \beta_{r+m}$, and the value 0 on the s-dimensional space spanned by $\beta_1, \ldots, \beta_{s+m}$.

Of course, we cannot know in practice when the last weak index has moved to the left for the last time. But we can give the following finitary version of our result.

Lemma 2. Let r, m be nonnegative integers, $\beta_1, \ldots, \beta_{r+m}$ a system of vectors spanning $\mathbb{R}^r$, and ε a positive number. Then in a number of steps which can be explicitly bounded the algorithm described above will transform the given system into one having the following properties:

(i) For some $s \geq 0$, $\beta_1, \ldots, \beta_{s+m}$ all have length $< \varepsilon$, and span an s-dimensional subspace $U \subseteq \mathbb{R}^r$, while the remaining $r-s$ vectors are (necessarily) linearly independent modulo this subspace. When this holds we also have

(ii) If we let H denote the additive subgroup of U spanned by those s vectors β_i $(1 \leq i \leq s)$ such that $\beta_i^{(i-1)} \neq 0$, then every point of U lies within distance $\frac{\sqrt{n}}{2}\varepsilon$ of a point of H.

(iii). Every linear functional $f: \mathbb{R}^r \to \mathbb{R}$ of norm $\leq 1/\varepsilon$ which assumes integer values on $\beta_1, \ldots, \beta_{r+s}$ is an integral linear combination of the $r-s$ functional defined to have value 1 at <u>one</u> of $\beta_{s+m+1}, \ldots, \beta_{r+m}$, and 0 at all other β_i.

<u>Proof that</u> (i) $\Rightarrow$ (ii) <u>and</u> (iii). From (i) we clearly have $|\beta_i^{(i-1)}| < \varepsilon$ for all nonterminal i. Given $x \in U$, we may "adjust" x with respect to the set of β_i such that i is nonterminal and strong, i.e. we can form a vector $y = x - \sum n_i \beta_i$ which satisfies $|y^{(i-1)} - y^{(i)}| \leq |\beta_i^{(i-1)}|/2$ (cf. (8)). Then by the Pythagorean theorem $|y| \leq \frac{\sqrt{n}}{2}\varepsilon$, i.e. x lies within that distance of $\sum n_i \beta_i \in A$, establishing (ii). Given f as in (iii) which is integer-valued on all β_i and has norm $\leq 1/\varepsilon$, we see that its value on each β_i $(i \leq s+m)$ will be $< (1/\varepsilon)\varepsilon = 1$, hence, being an integer, must be 0. The characterization of such functionals follows. ‖

If G is any additive subgroup of $\mathbb{R}^r$, then its closure $\mathrm{cl}(G)$ will be a closed subgroup. There is a simple description of the closed subgroups of finite-dimensional real vector spaces. They are direct products of vector spaces $\mathbb{R}^s$ and finitely generated discrete groups $\mathbb{Z}^t$. Thus, our algorithm can be thought of as finding (or approximating to any desired degree of accuracy) this decomposition for $\mathrm{cl}(G)$ when $G \subseteq \mathbb{R}^r$ is finitely generated (and for convenience, spans $\mathbb{R}^r$.)

We remark that to state Lemma 2 properly, one should give the bound on the number of steps, rather than merely saying that it "can be explicitly bounded". For some of the assertions of the Lemma, the existence of an algorithm whose length can be bounded is trivial. E.g. to find all functionals as in (iii), choose from the original system a basis $\{\beta_i \mid i \in I\}$ of $\mathbb{R}^r$, go through all ways of mapping each β_i $(i \in I)$ to an integer less than $|\beta_i|/\varepsilon$, and see which of these induce functionals that are also integer valued on the remaining β_j $(j \notin I)$. The point is that the "Euclidean" technique is so much faster. Thus, Ferguson and Forcade announce several results such as: [1] Theorem 5(a). Any polynomial of degree ≤ 5 with integer coefficients satisfied by the Euler constant γ must have at least one coefficient $\geq 10^{50}$. This they presumably obtain by applying their algorithm to search for $\mathbb{Z}$-linear dependence among $1, \gamma, \ldots, \gamma^5$, using a sufficiently good decimal approximation of γ. Clearly, the algorithm of exhaustive search could not give an estimate like this in any practical time!

However, we shall say no more about explicit bounds in this note. Such bounds can easily be obtained from (24) and related discussion following; we leave these to the expert to study. Let us note, instead, some information that Lemma 2 contains which is independent of algorithmic considerations.

Corollary 3. Let $\beta_1, \ldots, \beta_n$ be vectors spanning $\mathbb{R}^r$ $(r \leq n)$, and $B \in {}^n\mathbb{R}^r$ the matrix having the β_i as rows. Then the following conditions are equivalent:
(i) $\beta_1, \ldots, \beta_n$ satisfy no linear co-Diophantine relation in $\mathbb{R}^r$.
(ii) The additive subgroup $G \subseteq \mathbb{R}^r$ generated by $\beta_1, \ldots, \beta_n$ is dense.
(iii) For any $\varepsilon > 0$ we can find $u \in GL(n,\mathbb{Z})$ such that uB has all entries $< \varepsilon$. ||

In particular, the implication (i) $\Rightarrow$ (iii), applied to the case where $\beta_1, \ldots, \beta_n$ form a basis for the free abelian group G, gives

Corollary 4. (Ferguson and Forcade |1|, Theorem 3). If G is a finitely generated dense subgroup of $\mathbb{R}^r$, then every neighborhood of 0 in $\mathbb{R}^r$ contains a basis for G as a free abelian group. ||

8. Reinterpretations. Suppose $A \in {}^{m}\mathbb{R}^{r+m}$, $B \in {}^{r+m}\mathbb{R}^{r}$ are as in Lemma 1, i.e. the rows of A form a basis for the null space of B, equivalently (in view of the dimensions of these matrices) the columns of B form a basis for the null space of A. If we apply our algorithm to the rows of B, then as we noted in our formulation of Corollary 3, the successive changes which this system of vectors undergoes correspond to multiplication of the matrix B on the left by a certain series of matrices in $GL(r+m, \mathbb{Z})$.

If we look at the effect of these operations on the columns of B, we see that there they act, not by adding to one a specified multiple of the other, but by linear transformations from $GL(r+m, \mathbb{Z})$ applied to the ambient vector space ${}^{r+m}\mathbb{R}$. If the algorithm on the rows gives at some stage a system in which the last r-s rows are linearly independent modulo the rest, then the corresponding transformation on ${}^{r+m}\mathbb{R}$ carries the column space of B to a space which contains the last r-s standard basis vectors. Thus as it "discovers" co-Diophantine relations among the rows of B, it also "discovers" Diophantine points in the column space of B.

Note that when we multiply B on the left by $u \in GL(r+m, \mathbb{Z})$, if we also multiply A on the right by $u^{-1} \in GL(r+m, \mathbb{Z})$, our hypotheses relating A and B will be preserved. These operations transform the row space of A by a linear

action on the ambient space $\mathbb{R}^{r+m}$, and the columns of A by operations adding to one column a multiple of another. Note that if the column space of uB contains the last $r-s$ standard basis vectors of ${}^{r+m}\mathbb{R}$, then the row space of Au^{-1}, its annihilator, must consist of vectors with last $r-s$ entries 0, i.e. must lie in a particular canonical system of $r-s$ Diophantine hyperplanes. And finally, the last $r-s$ of the columns of $A\,u^{-1}$ will be zero, which is a trivial system of $r-s$ linear dependence relations on these columns, and yields a nontrivial system of such relations on the original columns of A.

One may ask whether the version of the Euclidean algorithm which we have described in terms of finding co-Diophantine relations on the rows of B can be translated directly to an algorithm on the original subject of our interest, the columns of A. I suspect that it cannot.* The matrix A does not uniquely determine B. When we first select the latter, we must choose a basis for the null space of A. Though we might do so according to some specified rule, once we start applying our transformations to A and B, they will no longer continue to be related by that rule. For instance, starting with $A = (1, x, 1)$ $(x \in \mathbb{R})$, we might take $B = \begin{pmatrix} x & 0 \\ -1 & -1 \\ 0 & x \end{pmatrix}$. If we should apply to A the operations of adding the first column to the middle column and subtracting the last column from the same middle column, then A is left unchanged, but the corresponding operations on B do not leave it fixed. Thus in some sense the choice of B seems to give us something more "solid" to work with than the original system A. (Ferguson and Forcade also remark laconically, "We exploit the nonuniqueness of Q [i.e., B]".)

The additional information that the algorithm gave us, which we stated in terms of the rows of B as Corollary 3, has interesting translations in terms

*But see §9 below.

of the columns of B and the rows and columns of A. For simplicity of presentation I will delay stating the "density" conditions, equivalent to (ii) of Corollary 3, until the others results have been presented, as I find them distracting. The reader is advised, in thinking about Corollaries 5 and 6, to emphasize the case $m = 1$, where the result is the most striking. (For Corollary 5 this means $\operatorname{codim}(W) = 1$, for Corollary 6, $\dim(V) = 1$. The general cases are, in fact, consequences of this one, via going to appropriate overspaces or subspaces.)

Corollary 5. Let W be a vector subspace of ${}^{n}\mathbb{R}$ $(n > 0)$. Then the following conditions are equivalent:

(i) W contains no Diophantine points of ${}^{n}\mathbb{R}$.

(ii) For all $\varepsilon > 0$ there exists $u \in GL(n,\mathbb{Z})$ which, if we regard it by restriction as a linear map $W \to {}^{n}\mathbb{R}$, has operator norm $< \varepsilon$ (say with respect to the standard inner product structure on ${}^{n}\mathbb{R}$).

Proof. Suppose W does contain a Diophantine point p. Elements of $GL(n,\mathbb{Z})$ carry Diophantine points to Diophantine points, and every Diophantine point has norm ≥ 1. Hence no element of $GL(n,\mathbb{Z})$ can act on W with operator norm $< 1/|p|$. Hence (ii) $\Rightarrow$ (i).

To prove the converse, take a basis $b_1, \ldots, b_r$ of W, and let $B \in {}^{n}\mathbb{R}^{r}$ be the matrix having these r columns. If W has no Diophantine points then by Lemma 1, the rows of B satisfy no linear co-Diophantine relation, hence by Corollary 3 there will exist elements $u \in GL(n,\mathbb{Z})$ making the entries of uB arbitrarily small. But these entries are the coordinates of the vectors $u\,b_i$, and by making these images of basis elements sufficiently small we can get the operator norm of u on W as small as we wish. ‖

Example. Note that the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ has eigenvalues $(1\pm\sqrt{5})/2$, with eigenvectors $\begin{pmatrix} 1 \\ \frac{1\pm\sqrt{5}}{2} \end{pmatrix}$ respectively. The eigenvalue $(1-\sqrt{5})/2$ has absolute value < 1, hence the elements $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}^i \in GL(2,\mathbb{Z})$ have norms converging to 0 on the 1-dimensional subspace $W \subseteq \mathbb{R}^2$ spanned by the corresponding eigenvector. The fact that matrices with arbitrarily small norm on this subspace are given by powers of a single matrix, and preserve the subspace, are of course not true in the general situation of Corollary 5; they are results of the fact that the (ordinary) Euclidean algorithm is periodic when applied to this vector. However, the example shows how members of $GL(n,\mathbb{Z})$ can act with small norm on a proper subspace of $\mathbb{R}^n$. The entries of these matrices, incidentally, are Fibonacci numbers, $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}^i = \begin{pmatrix} f_{i-1} & f_i \\ f_i & f_{i+1} \end{pmatrix}$, and a key fact is the equation $\tau^i = f_{i-1}\tau + f_i$, where τ represents either of $\frac{1\pm\sqrt{5}}{2}$.

Corollary 6. Let V be a subspace of $\mathbb{R}^n$ $(n > 0)$. Then the following conditions are equivalent:

(i) V is contained in no Diophantine hyperplane in $\mathbb{R}^n$.

(ii) For all positive real numbers C and ε, there exists $u \in GL(n, \mathbb{Z})$ carrying the ball $\{x \in \mathbb{R}^n \mid |x| \leq C\}$ into the open "tube" of radius ε about V, i.e. the set of points of $\mathbb{R}^n$ at distance $< \varepsilon$ from V.

Proof. Again, one direction may be shown without using the Euclidean algorithm. Say V lies in $\ker f$, where $0 \neq f: \mathbb{R}^n \to \mathbb{R}$ has integer coefficients. Let $|f|$ denote the operator norm of f. Then every Diophantine point $x \in \mathbb{Z}^n$ at which $f(x) \neq 0$ must satisfy $f(x) \geq 1$, hence must lie at a distance $\geq 1/|f|$ from $V \subseteq \ker f$. Now the closed unit ball contains a basis for $\mathbb{Z}^n$, and a

matrix $u \in GL(n,\mathbb{Z})$ cannot carry all members of this basis into the proper subspace $\ker f \subseteq \mathbb{R}^n$. Hence we see that it must take at least one of them to a point at distance $\geq 1/|f|$ from V. Hence (ii) $\Rightarrow$ (i).

Conversely, assuming (i), let $a_1, \ldots, a_m$ be a basis for V, let A be the matrix with these rows, and let $B \in {}^{n}\mathbb{R}^{n-m}$ be a matrix whose columns form a basis for the null space of A. By Lemma 1 (ii) $\Rightarrow$(iv), our hypotheses imply that the rows of B satisfy no linear co-Diophantine relation, hence by Corollary 3 there exist matrices $u \in GL(n,\mathbb{Z})$ making the entries of uB arbitrarily small. Now the columns of B, $b_1, \ldots, b_{n-m}$ represent a basis for the linear functionals on $\mathbb{R}^n$ annihilating V, and it is not hard to see that given C, $\varepsilon > 0$, we can find $\varepsilon' > 0$ such that

$$\forall\, x \in \mathbb{R}^n,\ (|x\, b_i| < \varepsilon' \ (i=1,\ldots,n-r)) \Rightarrow (\mathrm{dist}(x,V) < \varepsilon\ ).$$

Now choose $u \in GL(n,\mathbb{Z})$ such that all columns of uB have length $< \varepsilon'/C$. Then for any $x \in \mathbb{R}^n$ we see that

$$(|x| \leq C) \Rightarrow (|x\, u\, b_{(i)}| < C(\varepsilon'/C) = \varepsilon' \ (i=1,\ldots,n-r)) \Rightarrow (\mathrm{dist}(xu, V) < \varepsilon).$$

as required. ‖

Let us consider the above Corollary in the case where V is 1-dimensional. Since an element $u \in GL(n,\mathbb{Z})$ has determinant ± 1, it is volume-preserving, hence if it moves all vectors close to the line V, it must also squeeze them far out <u>along</u> that line. Hence taking a random vector x in the closed unit ball, it will tend to give a very long vector xu, which is also very close to V. Thus the ratios of the coordinates of xu will in general be close to the coordinate-ratio characterizing points of the line V. Of course, since the image under u of any ball contains a neighborhood of 0, the statement in

this form applies only to "most points". But suppose we look only at the points $x = (1,0,\ldots,0), \ldots, (0,\ldots,0,1)$. Then the xu will all be Diophantine points. (In fact, they are the rows of u, and form a basis of $\mathbb{Z}^n$.) As we force them to be close to V, they must all become long, since in any fixed ball there are not Diophantine points arbitrarily close to V. Thus the algorithm gives bases of $\mathbb{Z}^n$ whose points have large norm, but lie arbitrarily close to V.

(An algorithm which generates, for a given 1-dimensional V, matrices $u \in GL(n,\mathbb{Z})$ the distance of whose rows from V tends to 0 is said by Ferguson and Forcade to "approximate" for V. If on the other hand the algorithm eventually produces a u such that some coordinate of all points of Vu is 0, they say the algorithm "terminates" for V. An algorithm, like the one they describe or the version we describe above, which either approximates or terminates for every V, they say "splits". They indicate that they have examples showing that the Jacobi-Perron algorithm, cf. p.2 above, does not split.*)

Again, the matrices $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}^i = \begin{pmatrix} f_{i-1} & f_i \\ f_i & f_{i+1} \end{pmatrix}$ give a simple example: they move points in any bounded subset of $\mathbb{R}^2$ into narrow strips about the line $(1, (1+\sqrt{5})/2)\mathbb{R}$.

The next, and final "translation" of Corollary 3, in terms of the columns of A, is not as simple to state as those that precede. Given vectors $\alpha_1, \ldots, \alpha_{r+m} \in {}^m\mathbb{R}$, and vectors $\alpha'_1, \ldots, \alpha'_{r+m}$ in any normed vector space X, and a real number $\varepsilon > 0$, let us say that the latter system is <u>linearly ε-approximable</u> by the former if there exists a linear map $f: \mathbb{R}^r \to X$ such that $|f(\alpha_i) - \alpha'_i| = 0$. The next Corollary follows easily from Corollary 6, so we omit the proof.

*Actually, they say it "can terminate without a relation", but this makes no sense that I can see, so I assume they mean it can be applied to a linearly dependent family without yielding a relation.

Corollary 7. Let $m \leq n$ be nonnegative integers, and $\alpha_1, \ldots, \alpha_n$ a system of vectors spanning ${}^m\mathbb{R}$. Then the following conditions are equivalent:

(i) $\alpha_1, \ldots, \alpha_n$ are $\mathbb{Z}$-linear independent.

(ii) For all positive real numbers C, ε, there exists $u \in GL(n, \mathbb{Z})$ which, when applied to any n-tuple of vectors in any real normed vector space X, whose lengths are all $\leq C$, yields an n-tuple which is linearly ε-approximable by $\alpha_1, \ldots, \alpha_n$. ||

(To see how this follows from Corollary 6, the reader might look at the case $r = m = 1$; $\alpha_1 = 1$, $\alpha_2 = (1+\sqrt{5})/2$; $u = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right)^i$.)

Let us now state a third equivalent condition for each of the last three Corollaries, corresponding to condition (ii) of Corollary 3, (density of the subgroup). To enable us to state the version for Corollary 7, let us define a map from a real vector space into a torus ${}^h\mathbb{R}/{}^h\mathbb{Z}$ to be "linear" if it is induced by a linear map into ${}^h\mathbb{R}$. (These "linear" maps are just the continuous group homomorphisms.) Then we can define "ε-linear approximability" of a system of elements of a torus by a system of elements of a vector space. Let us just say a system of elements of a torus is "linearly approximable" by a system in a vector space if it is ε-linearly approximable for all ε. (Of course, the linear maps used will depend on ε.) The equivalence we will state using this definition is very likely known to ergodic theorists.

Corollary 8. The equivalent pairs of conditions of Corollaries 5-7 are also respectively equivalent to:

(5,iii) The set of linear functionals on W induced by linear functionals on ${}^{n}\mathbb{R}$ with integer coefficients is dense in all linear functionals on W.

(6,iii) The image of $\mathbb{Z}^n$ in $\mathbb{R}^n/V$ is dense.

(7,iii) Every n-tuple of vectors in any torus group ${}^{h}\mathbb{R}/{}^{h}\mathbb{Z}$ is linearly approximable by $\alpha_1, \ldots, \alpha_n \in {}^{m}\mathbb{R}$. ‖

9. Some further remarks. Though neither of the two matrices A and B in Lemma 1 determines the other in general, in the case $m = 1$ we see that given B, the 1-by-(r+1) matrix A will be unique up to a scalar factor. In fact, up to such a factor and appropriate signs, the terms of A are the $r \times r$ minors of B; and the constant factor does not change under the action of $GL(r+1, \mathbb{Z})$, so we may fix it at 1. Thus, if one pictures the algorithm on $\beta_1, \ldots, \beta_{r+1}$ geometrically, one can say that $\alpha_1, \ldots, \alpha_{r+1}$ are the volumes of the parallelopipeds spanned by r-tuples of the β's, and the question of whether we can set up some version of the algorithm that works only with the α's is equivalent to asking whether a strategy for reducing the lengths of the β's by subtracting multiples of one from another can be formulated in which the operations to be performed at any stage are determined wholly by the ratio of these volumes.

The ratio of the volumes determines $\beta_1, \ldots, \beta_{r+1}$ up to a linear automorphism of $\mathbb{R}^r$. (Indeed A clearly always determines B up to right multiplication by a member of $GL(r, \mathbb{R})$.) So still another way to put the problem is: Can an algorithm be described based only on the linear structure of the system $\beta_1, \ldots, \beta_{r+1} \in \mathbb{R}^r$, or must one introduce some other structure, such as the metric structure

I have used or a coordinate system as used by Forcade and Ferguson? The goal that all lengths of β_i's "eventually" approach 0 can be expressed in terms of the vector space structure, because the topology of a real finite-dimensional vector space is uniquely determined; but it is not clear whether there is a systematic way of achieving this without using more structure.

Afterthought: Indeed, there must be such a form of the algorithm! For given $\beta_1, \ldots, \beta_{r+1}$ spanning $\mathbb{R}^r$, let $C(\beta) \subseteq \mathbb{R}^r$ denote the convex hull of $\{\pm\beta_i\}$. Using the algorithm as presented here, one can associate to every such (r+1)-tuple a finite sequence of operations of the sort we are using, which will either reveal a co-Diophantine relation, or transform the given (r+1)-tuple to one whose vectors all lie in $\frac{1}{2}C(\beta)$. By straightforward arguments this can be done in a way that is invariant under linear automorphisms of $\mathbb{R}^r$, and reasonably nice — i.e. such that the set of all (r+1)-tuples for which a given sequence of operations is used forms is described by a fairly simple system of inequalities. Iteration of this process will be the desired algorithm. So the real questions are—Can this be done in a way that is really convenient to work with? In a way such that the corresponding algorithm on the α's can be conceptually explained? In a way such that the number of operations of adding to one vector a combination of the others which constitutes "one step" of this algorithm is bounded? A candidate algorithm, which is nice, but which may or may not work, will be mentioned in §12.

10. Periodic continued fractions. I shall prove here a classical result, using the "ordinary" Euclidean algorithm. But there is some variability as to what the "ordinary" Euclidean algorithm can mean, which we should settle first. One matter is simply notational: we shall find it most convenient to use an algorithm without "reindexing". Thus, starting with two real numbers α and β, we get at the next stage a pair α', β' of the form α, $\beta - n\alpha$, and at the next stage we operate in the opposite way, getting α'', β'' in the form $\alpha' - n'\beta'$, β', etc.. The coefficients n, n', ..., $n^{(i)}$,... are the terms in the continued fraction expansion:

$$\beta/\alpha = n + \cfrac{1}{n' + \cfrac{1}{n'' + \ddots + \cfrac{1}{n^{(i)} + \ddots}}}.$$

Secondly, we recall that one may perform the Euclidean algorithm either choosing the $n^{(i)}$ so as to give the least positive remainders, or so as to give the remainders of smallest absolute value. This leads to two interpretations of "the" continued fraction expansion of a real number. We shall find that with trivial readjustments, the same proof applies to both.

Theorem (Lagrange). Let c be an irrational real number. Then the following conditions are equivalent.

(i) The continued fraction expansion of c (understood in either of the above ways) is eventually periodic.

(ii) $[\mathbb{Q}(c) : \mathbb{Q}] = 2$.

Proof. One way is straightforward. If we assume (i), then c may be expressed as a fractional linear transform with integer coefficients of the element c' represented by the periodic part of the continued fraction, and c' by periodicity becomes a fractional linear transform of itself: $c' = (pc' + q)/(rc' + s)$ $(p, q, r, s \in \mathbb{Z})$. The latter equation yields a quadratic equation satisfied by c', and this makes c quadratic as well.

Now assume (ii). Let us write $c = \beta/\alpha$, where α, β are algebraic integers in $Q(c)$, and apply the Euclidean algorithm to this pair of real numbers. We shall call the pair obtained at the i^{th} step $\alpha^{(i)}$, $\beta^{(i)}$. These will clearly again be algebraic integers.

Consider now the determinant

$$(29) \qquad \begin{vmatrix} \alpha^{(i)} & \beta^{(i)} \\ \bar{\alpha}^{(i)} & \bar{\beta}^{(i)} \end{vmatrix} = \alpha^{(i)}\bar{\beta}^{(i)} - \bar{\alpha}^{(i)}\beta^{(i)},$$

where $x \mapsto \bar{x}$ denotes the automorphism of the quadratic field $Q(c)$, which we shall call conjugation. Note that the transition from the i^{th} to the $(i+1)^{st}$ step affects the matrix in (29) by a column operation, which has no effect on the determinant. Thus the value of (29) is an invariant of our sequence of pairs. Further, conjugation affects (29) by interchanging the rows of the matrix, hence changing the sign of the determinant, so the common value of the elements (29) must be of the form $\sqrt{D}$ (where $D \in \mathbb{Z}$, and $\sqrt{D}$ represents a fixed one of its square roots, not necessarily the positive one.)

We now consider for each i the polynomial satisfied by the ratio of the terms at the i^{th} stage, $\beta^{(i)}/\alpha^{(i)}$,

$$(30) \qquad (\alpha^{(i)}t - \beta^{(i)})(\bar{\alpha}^{(i)}t - \bar{\beta}^{(i)})$$

$$= (\alpha^{(i)}\bar{\alpha}^{(i)})\,t^2 - (\alpha^{(i)}\bar{\beta}^{(i)} + \bar{\alpha}^{(i)}\beta^{(i)})\,t + (\beta^{(i)}\bar{\beta}^{(i)})$$

$$= a^{(i)}t^2 - b^{(i)}t + c^{(i)}.$$

The coefficients $a^{(i)}$, $b^{(i)}$, $c^{(i)}$ will be integers because they are algebraic integers in $Q(c)$ invariant under conjugation. The discriminant of this polynomial, $b^{(i)\,2} - 4\,a^{(i)}c^{(i)}$, is easily seen to be D, the square of (29).

(Either by computation, or by noting that (29) equals the numerator of the difference of the roots of (30), written with denominator $a^{(i)}$.) Thus (30) gives a sequence of polynomials over $\mathbb{Z}$ all having the same discriminant.

We next note that the coefficient $b^{(i)} = \alpha^{(i)} \bar{\beta}^{(i)} + \bar{\alpha}^{(i)} \beta^{(i)}$ may be written in two ways:

$$(31) \qquad b^{(i)} = 2\,\alpha^{(i)} \bar{\beta}^{(i)} - \sqrt{D} = 2\,\bar{\alpha}^{(i)} \beta^{(i)} + \sqrt{D}\,.$$

Now we claim that for $i > 0$,

(32) If $b^{(i)} \notin [-3|\sqrt{D}|,\ 3|\sqrt{D}|]$, then $b^{(i+1)}$ must be smaller in absolute value than $b^{(i)}$, while if $b^{(i)}$ lies within this interval, so will $b^{(i+1)}$.

Indeed, consider first the case where the pair $\alpha^{(i+1)}$, $\beta^{(i+1)}$ is obtained from $\alpha^{(i)}$, $\beta^{(i)}$ by modifying the first term and preserving the second. We then use the first expression for $b^{(i)}$ in (31). In this expression only the α term changes at this step, and the new term $\alpha^{(i+1)}$ will be (depending on which form of the Euclidean algorithm we use) either of the same sign as $\alpha^{(i)}$ but smaller in absolute value, or of arbitrary sign and smaller in absolute value by at least half. In either case, (32) is easily checked* If, conversely, the $i+1^{st}$ pair is obtained from the i^{th} by modifying the β term, we use the second formula in (31) to get the same result.

Now (32) says that as we perform our algorithm, $b^{(i)}$ eventually becomes limited to a finite set of values. Since $b^{(i)\,2} - 4\,a^{(i)}\,c^{(i)}$ is constant, $4\,a^{(i)}\,c^{(i)}$ also ranges over only a finite set of values, and by factorization properties of $\mathbb{Z}$, we get only finitely many possibilities for $a^{(i)}$ and $c^{(i)}$, hence only finitely many possibilities for the quadratic equation (30). Each

*If one uses the version of the algorithm in which the sign of α does not change, then (32) is even true using the smaller interval $[-|\sqrt{D}|,\ |\sqrt{D}|]$.

Such equation has only two roots, so we eventually get a repetition

$$\beta^{(i+r)}/\alpha^{(i+r)} = \beta^{(i)}/\alpha^{(i)}. \tag{33}$$

Since the steps applied in the Euclidean algorithm depend only on this ratio, the algorithm must repeat with period r. ‖

Remark: The above proof is in essence the same as all proofs of Lagrange's result that I have seen, but most presentations involve less transparent computations because the authors work with the field-elements $c^{(i)} = \beta^{(i)}/\alpha^{(i)}$ rather than the pairs of algebraic integers $\alpha^{(i)}$, $\beta^{(i)}$. I.e., they work strictly in terms of continued fractions, and not in terms of the Euclidean algorithm.

Note that in the situation (33), we can write

$$\alpha^{(i+r)} = a\ \alpha^{(i)}, \qquad \beta^{(i+r)} = a\ \beta^{(i)} \qquad (a \in Q(c)). \tag{34}$$

From the fact that $a\alpha^{(i)}$ and $a\beta^{(i)}$ lie in the additive group spanned by $\alpha^{(i)}$ and $\beta^{(i)}$ it follows that a is an algebraic integer. (Indeed, if we take a matrix u over $\mathbb{Z}$ such that $(\alpha^{(i+r)}, \beta^{(i+r)}) = (\alpha^{(i)}, \beta^{(i)})\ u$, then a will satisfy the characteristic polynomial of u.) But $\alpha^{(i)}$ and $\beta^{(i)}$ are similarly expressible in terms of $\alpha^{(i+r)}$ and $\beta^{(i+r)}$ (i.e. $u \in GL(2, \mathbb{Z})$) so a^{-1} is also an algebraic integer. Thus a is a <u>unit</u> in the ring of algebraic integers in $Q(c)$. Since $|\alpha^{(i+r)}| < |\alpha^{(i)}|$, $|a| < 1$, so $a \notin \mathbb{Z}$, i.e. it is a 'nontrivial' unit of this ring of algebraic integers.

11. Ideas on finding units in higher number fields using the generalized algorithm. Suppose the algorithm described in §4, applied to an $(r+1)$-tuple of real numbers $\alpha_1, \ldots, \alpha_{r+1}$, equivalently, to an $(r+1)$-tuple of vectors $\beta_1, \ldots, \beta_{r+1} \in \mathbb{R}^r$, becomes periodic. Then the matrix $u \in GL(r+1, \mathbb{Z})$ representing the operations of the algorithm over the course of one period can easily be seen to have as an eigenvector the value of $(\alpha_1, \ldots, \alpha_{r+1})$ occuring at the beginning of a period. The eigenvalue on that vector will be an invertible algebraic integer a. Thus $[\mathbb{Q}(a): \mathbb{Q}] \leq r+1$, and if one has equality, one finds that the ratios α_i/α_1 must form a basis of $\mathbb{Q}(a)$ over $\mathbb{Q}$.

Hence to find units a in a real algebraic number field K, it seems reasonable to start with a basis $\alpha_1, \ldots, \alpha_{r+1}$ of K over $\mathbb{Q}$, and perform our algorithm on this system of numbers.

Can we duplicate the above proof of Lagrange's theorem, or some important parts of it, for this generalized situation? Two key points in the proof were the invariance of the determinant (29), and the fact that every step decreased one of our two real numbers, leaving the other fixed.

On the other hand, the important idea of Ferguson and Forcade on which the generalized algorithm is based is that one needs to consider the effect of one's transformations not on the numbers $\alpha_1, \ldots, \alpha_{r+1}$, but on the vectors $\beta_1, \ldots, \beta_{r+1}$; equivalently, on the matrix B. Now this leads to a great number of analogs of (29). Let $\sigma_1, \ldots, \sigma_{r+1}$ be the distinct embeddings of K in the complex numbers. Then for any column b_i of B, the determinant $\det(\sigma_1 b_i, \ldots, \sigma_{r+1} b_i)$ will be constant under the operations of our construction. But we can form still more determinants with this property! E.g. $\det(\sigma_{i_1} b_{j_1}, \ldots, \sigma_{i_{r+1}} b_{j_{r+1}})$ for fairly arbitrary sequences $i_1, \ldots, i_{r+1}$ and $j_1, \ldots, j_{r+1}$, which may even involve repetitions, as long as both do not repeat simultaneously.

This is too much data to know what to do with. But recall that our original data $\alpha_1, \ldots, \alpha_{r+1}$ do not uniquely determine B. Might it be possible to choose B so that some of these invariants have easily described relationships to others? For instance, by making the various columns of B images of one another under the σ_i ?

Following this idea through, I finally came to the conclusion that it was perhaps most natural not to start with the α's at all. Consider instead the following approach:

Let K be an abstract algebraic number field, of degree $r+1$, and $\sigma_1, \ldots, \sigma_{r+1}$ its distinct embeddings in C. Let us temporarily assume that these are in fact real-valued, i.e. that K is a totally real field. Let us take a Q-basis of K consisting of algebraic integers, which we shall denote $\beta_1^*, \ldots, \beta_{r+1}^*$. We now define vectors $\beta_1, \ldots, \beta_{r+1} \in \mathbb{R}^r$ by $\beta_i = (\sigma_1(\beta_i^*), \ldots, \sigma_r(\beta_i^*))$, and perform our algorithm on this system.

In fact, one may look at what I have described as the embedding of K in $\mathbb{R}^r$, using the r real embeddings $\sigma_1, \ldots, \sigma_r$, and performing our algorithm in K itself, using the structure of normed space it acquires under this embedding. If the algorithm becomes periodic, then the period will correspond to multiplication by a unit which has length < 1 under all of these r embeddings (and hence, of course, length > 1 under σ_{r+1}.)

We remark that the vector $\alpha_1, \ldots, \alpha_{r+1}$ corresponding to this system, formed from the minors of B (cf. §9), can be shown to have the ratios α_i/α_1 in $\sigma_{r+1}(K)$; in fact they will form a basis thereof.

In the situation described above, there is a unique natural analog of (29), namely

$$\det\left(\sigma_i(\beta_j^*)\right)_{1 \leq i,\, j \leq r+1} \tag{35}$$

This will be the square root of an integer, and constant under the algorithm.

The analog of (30) would seem to be a homogeneous polynomial in $r+1$ indeterminates:

$$(36) \qquad \prod_i \left(\sum_j \sigma_i(\beta_j^*)\, t_j \right) \in \mathbb{Z}[t_1, \ldots, t_{r+1}].$$

But whether one can prove that under (some version of) our algorithm, (36) assumes only finitely many values — which would yield the analog of Lagrange's theorem! — I can't say. Perhaps one should first think about the simplest case $r+1 = 3$.

Putting this key question aside, let us ask how we should vary the above considerations if K is not totally real. We can, of course, simply use the version of our algorithm with $\mathbb{C}$ and $\mathbb{Z}[i]$ in place of $\mathbb{R}$ and $\mathbb{Z}$. Unfortunately, this would correspond to looking for units in $K(i)$, which might not lie in K. However, suppose that σ_{r+1} is real; equivalently that those of $\sigma_1, \ldots, \sigma_r$ that are not real occur in complex conjugate pairs. Say σ_1 and σ_2 are such a pair. Note that in $\mathbb{C}^r$, we can make a unitary transformation so that the r-tuple β_j, instead of beginning $(\sigma_1(\beta_j^*), \sigma_2(\beta_j^*), \ldots)$ begins

$$\left(\frac{\sigma_1(\beta_j^*) + \sigma_2(\beta_j^*)}{\sqrt{2}}, \frac{\sigma_1(\beta_j^*) - \sigma_2(\beta_j^*)}{i\sqrt{2}}, \ldots \right)$$

$$= \left(\sqrt{2}\, \mathrm{Re}(\sigma_1(\beta_j^*)), \sqrt{2}\, \mathrm{Im}(\sigma_1(\beta_j^*)), \ldots \right).$$

These two functions are always real. Doing the same with each conjugate pair, we get a system of $r+1$ vectors in $\mathbb{C}^r$ which are in fact $\mathbb{R}^r$-valued, and will remain so under application of our algorithm. So the complex algorithm, applied to these vectors, reduces to a case of the real algorithm! (As noted, the factor $\sqrt{2}$ keeps the new system unitarily equivalent to the old one. Whether there would be any loss in dropping it, I don't know.)

Of course, when σ_{r+1} is itself complex, I suppose one must deal with every pair of the complex algorithm.

When the embedding one wishes to distinguish is itself complex, I suspect that the best approach would be to call the degree of K $r+2$, let $\sigma_1, \ldots, \sigma_r$ be all embeddings except the distinguished one and its complex conjugate (for one cannot allow a unit to have absolute value > 1 under one of these embeddings but not the other), and apply the algorithm to the $r+2$ vectors $(\sigma_1(b_j^*), \ldots, \sigma_r(b_j^*)) \in \mathbb{C}^r$. As before, a change of coordinates reduces this in turn to an application of the algorithm in $\mathbb{R}^r$.

Woops! see p.53, note 4

Of course, the complex algorithm per se would be appropriate when looking for units in extensions of the Euclidean fields $Q(i)$, $Q(\omega)$, etc..

12. A simple candidate algorithm. We mentioned on p.1 that the one property of the Euclidean algorithm which it was clear could be generalized without difficulty of systems of n real numbers was that of detecting commensurability.

Now a closely related property of the ordinary Euclidean algorithm is that of detecting approximation to ratios of small integers. Example: My 13 year old stepson Jeff Watson told me that his batting average, for the weeks that his class had been playing softball in Physical Education, was .846. Wondering what score he had computed this from, I applied the Euclidean algorithm to 1000 and 846, getting the continued fraction expansion $.846 = \frac{1}{1+}\ \frac{1}{5+}\ \frac{1}{2+}\ \frac{1}{38}$. Clearly, this is close to $\frac{1}{1+}\ \frac{1}{5+}\ \frac{1}{2} = \frac{11}{13}$. I asked whether he had gotten 11 hits in 13 times at bat, and he said yes (but pointed out that I had know way of knowing it wasn't 22 out of 26.)

Suppose, now, that we are given three or more numbers whose ratio is an approximation of a ratio of small integers, which we wish to reconstruct. We can try a Euclidean algorithm type program of reducing various terms by integer multiples of others. But whether we are led to the correct ratio, despite the approximate nature of our data, may depend upon our choices of which terms to

subtract from which at each stage. Let us, then, consider each number as containing a small error term, and try to use the procedure which will cause the least build-up of these errors. Thus, if $|a| < |b| < |c|$, and we have a choice of reducing c by subtracting a large integer times a or a smaller integer times b, the latter would be preferable, since it would add to c a smaller multiple of an error term. This leads to the algorithm: Always subtract from the largest member, x, of the current system of real numbers, the least integer multiple of the next-to-largest member, y, which leaves a remainder smaller than y in absolute value. (This may be contrasted with the Jacobi-Perron algorithm where, after the first cycle, one always reduces the largest term by an integral multiple of the smallest. This cuts down the size of one's numbers very quickly, but presumably also allows errors to build up fast if one's numbers are not exact.)

See p. 64, note 5 re earlier discovery of above algorithm

The above reasoning is, of course, pragmatic rather than rigorous. I have not thought about how to formalize or prove a statement to the effect that this is a particularly good algorithm. However, in the same speculative vein we may inquire further. As pointed out at the end of §9, there should exist some algorithm which will detect linear dependence among a family of real numbers $\alpha_1, \ldots, \alpha_{r+1}$, without recourse to a family of vectors $\beta_1, \ldots, \beta_{r+1}$ such as we have used. Might the above be such an algorithm?

Still more speculatively, might it become periodic when applied to a basis of an algebraic number field?

As an experiment, I applied it to 1, $\sqrt[3]{2}$, $\sqrt[3]{4}$, using for the last two the six-place values in CRC Standard Mathematical Tables: 1.259921 and 1.587401. To record the results, let "n" denote "n times the second largest value was subtracted from the largest value, "." denote "the remainder then became the

second largest value", and ":" denote "the remainder then became the smallest value." The result of the calculation was

(37) 1:1:3:1·3·1:3:1·1:1:1:1:1:1:3:1·3·1·2:1:1·19:1 .

Notice that the pattern "1:3:1·3·1" occurs both right near the beginning (starting with the second "1"), and toward the end. Conjecturing that this represented the beginning of the repetition of a period obscured by the inexactness of the given approximations, I wrote down the presumed period, 1:3:1·3·1:3:1·1:1:1:1:1·, and computed the corresponding matrix $u \in GL(3,\mathbb{Z})$, and its inverse:

$$u = \begin{pmatrix} 24 & -8 & -5 \\ 6 & -19 & -16 \\ -11 & -13 & 14 \end{pmatrix}, \qquad u^{-1} = \begin{pmatrix} 58 & 177 & 223 \\ 92 & 281 & 354 \\ 131 & 400 & 504 \end{pmatrix}.$$

If my conjecture was correct, I should get $(\sqrt[3]{4} - \sqrt[3]{2},\ 1,\ \sqrt[3]{2})\ u = a(\sqrt[3]{4} - \sqrt[3]{2},\ 1,\ \sqrt[3]{2})$, since $(\sqrt[3]{4} - \sqrt[3]{2},\ 1,\ \sqrt[3]{2})$ is the image of the original 3-tuple under the initial step "1:", which precedes the conjectured period, for some $a \in Q(\sqrt[3]{2})$. A necessary and sufficient condition for u to behave in this way is that it commute with the matrix v representing multiplication by $\sqrt[3]{2}$ in terms of the basis $(\sqrt[3]{4} - \sqrt[3]{2},\ 1,\ \sqrt[3]{2})$. This is

$$v = \begin{pmatrix} -1 & 0 & 1 \\ 2 & 0 & 0 \\ -1 & 1 & 1 \end{pmatrix}$$

One finds that, indeed, $uv = vu$, and concludes that u will indeed act on the indicated 3-tuple as multiplication by an element a. One easily obtains a by looking at the middle term of the product 3-tuple, and one similarly gets a^{-1} from the middle term of the image of the 3-tuple under u^{-1}:

$$a = 19 - 5\sqrt[3]{2} - 8\sqrt[3]{4}, \qquad a^{-1} = 281 + 223\sqrt[3]{2} + 177\sqrt[3]{4}.$$

Note that the period discovered is "palindromic", in the sense that it has centers of symmetry at the ".3." and at the middle of the series of ":1:"s. The continued fraction expansions of square roots of rational numbers are also known to have such a property (cf. [2], p.825 item (f)).

On the other hand, a similar experiment with $1, \sqrt[3]{3}, \sqrt[3]{9}$ gave:

1:1:1:1:1:1·1:2:2·1·3:1:4·1:2:3:1:3:1·1:1·3:1·2,

with no evidence of periodicity or palindromicity. Perhaps 6-digit approximations are not good enough for this case, or perhaps my algorithm, like the Jacobi-Perron algorithm, simply gives results sometimes but not always. See p.55, note 6.

I would be interested in knowing whether this algorithm is one that has been examined before.

13. Appendix. A result on dense subgroups. In §7 we generalized the algorithm studied from the case of $r+1$ vectors to $r+m$ vectors in $\mathbb{R}^r$, and translated the results into statements about dense subgroups. This suggests the following question: Suppose G is a dense finitely generated subgroup of $\mathbb{R}^r$, say free abelian of rank $r + m$. Will G in general contain a subgroup of rank $r + 1$ (i.e. free abelian on $r+1$ generators) which is still dense in $\mathbb{R}^r$?

It seems that for "most" G this is so, but definitely not for all. In this section we shall characterize precisely those G for which it fails when $r = 2$.

The result, and/or many of the observations used to prove it, may be known. I would be grateful to anyone who could give me references.

For any subset $X \subseteq \mathbb{R}^r$, let $V(X)$ denote the subspace of $\mathbb{R}^r$ spanned by X, $A(X) \subseteq V(X)$ the additive subgroup generated by X, and $U(X) \subseteq V(X)$ the connected component of 0 in the closure $\mathrm{cl}(A(X))$. It is clear from the description of the structure of closed subgroups mentioned in §7 that $U(X)$ is open in $\mathrm{cl}(A(X))$, hence that

(38) $\qquad A(X) \cap U(X)$ is dense in $U(X)$.

For X finite, it is convenient to study $A(X)$ by taking in X a vector-space basis Y of $V(X)$, which will satisfy $U(Y) = \{0\}$, $V(Y) = V(X)$, and seeing how $U(Y \cup Z)$ grows as additional families of elements $Z \subseteq X - Y$ are added. Once Y is fixed, the possible vector-spaces occurring as $U(Y \cup Z)$ $(Z \subseteq V(Y))$ turn out to be quite limited. In the following four assertions, Y will be any $\mathbb{R}$-linearly independent subset of $\mathbb{R}^r$. I could supply proofs, but since the results are not deep and may be well-known, and since I am short of time, I shall not do so here. (41) and (42) are deduced from (39) and (40).

(39) If U is a vector subspace of $V(Y)$, then $A(Y) + U \subseteq V(Y)$ is closed if and only if U has a basis contained in $A(Y)$. When this holds, the connected component of 0 in the closed group $A(Y) + U$ is just U.

(40) If P is any closed subgroup of $V(Y)$ which contains $A(Y)$ then $U(P)$ satisfies the above conditions, and $A(Y) + U(P)$ has finite index in P.

(41) If Z_1, Z_2 are subsets of $V(Y)$, then $U(Y \cup Z_1 \cup Z_2) = U(Y \cup Z_1) + U(Y \cup Z_2)$.

(42) Suppose $Y = \{y_1, \ldots, y_t\}$, and let $z \in V(Y)$. Let us write $z = c_1y_1 + \ldots + c_ty_t$ ($c_i \in \mathbb{R}$). Let $\{1 = p_0, p_1, \ldots, p_u\}$ be a $\mathbb{Q}$-basis for the $\mathbb{Q}$-vector-subspace of $\mathbb{R}$ spanned by $1, c_1, \ldots, c_t$. Thus we can write $z = \sum p_i(r_{i1}y_1 + \ldots + r_{it}y_t)$, where the r_{ij} are rational numbers, of which those with $i \neq 0$ may, by appropriate modification of $p_1, \ldots, p_u$, be taken to be integers. Then an $\mathbb{R}$-basis for $U(Y \cup \{z\})$ contained in $A(Y)$ is given by the vector cofactors of $p_1, \ldots, p_u$ in the above expression.

From (41) it is easy to deduce:

(43) Any finitely generated dense subgroup $G \subseteq \mathbb{R}^r$ contains a subgroup generated by $\leq 2r$ elements which is also dense.

Indeed, let X be a finite generating set for G, and let us choose from X a basis Y of $\mathbb{R}^r$. Then from (41) we see that $U(X) = \mathbb{R}^r$ will be the sum over all $z \in X - Y$ of the vector subspaces $U(Y \cup \{z\})$. Clearly a subsum of $\leq r$ terms will give all of $\mathbb{R}^r$; hence the r elements of Y together with $\leq r$ additional elements generate a dense subgroup.

We ask for which groups G there are no dense subgroups of rank smaller than $2r$. For $r = 2$ this is the question we started with, since $2r-1$ is the same as $r + 1$. To state the answer, we need an invariant of groups $G \subseteq \mathbb{R}^r$:

(44) If G is any subgroup of $\mathbb{R}^r$, take an $\mathbb{R}$-basis Y for $V(G)$; thus every element of G may be written as a linear combination of elements of Y with coefficients in $\mathbb{R}$. Let $\mathbb{Q}(G) \subseteq \mathbb{R}$ denote the <u>field</u> generated over the rationals by all coefficients occurring in the expressions for elements of G to this basis. Then the field $\mathbb{Q}(G)$ is independent of the choice of the basis Y.

This is easily deduced by looking at change-of-basis matrices. We can now state:

<u>Lemma 9.</u> Let G be a finitely generated dense subgroup of $\mathbb{R}^r$, where $r \geq 2$. Then the following two conditions are equivalent.

(i') Every subgroup of G dense in $\mathbb{R}^r$ has rank $\geq 2r$.

(ii) $[\mathbb{Q}(G): \mathbb{Q}] = 2$.

<u>Proof.</u> Assume (i). Let $\hat{Y} \subseteq G$ be a vector space basis for $\mathbb{R}^r$. Then for any generating set X for G, $\sum_X U(Y \cup \{x\}) = U(Y \cup X) = \mathbb{R}^r$. Now if any of these spaces $U(Y \cup \{x\})$ had dimension greater than 1, we could add to it fewer than $r-1$ other spaces $U(Y \cup \{x'\})$ and get all of $\mathbb{R}^r$. The result would be a subset $Y \cup Z'$ of cardinality $< 2r$ such that $U(Y \cup Z') = \mathbb{R}^r$, i.e. such that $A(Y \cup Z')$ is dense. Thus

(45) For all $x \in G$, $\dim U(Y \cup \{x\}) \leq 1$.

Now let us take r elements $x_1, \ldots, x_r$ such that the spaces $U(Y \cup \{x_i\})$ are 1-dimensional and sum to $\mathbb{R}^r$. By (38), elements of G are dense in each of these spaces, so by (43) we can find in each $U(Y \cup \{x_i\})$ two elements which generate a dense subgroup therein; call these z_i and $c_i z_i$ $(c_i \in \mathbb{R} - \mathbb{Q})$.

Now for any i, j, let us apply (45) with $\{z_1, \ldots, z_r\}$ in place of Y, and $c_i z_i + c_j z_j$ for x. We conclude that $\dim U(\{z_1, \ldots, z_r\} \cup \{c_i z_i + c_j z_j\}) = 1$.

By (42) this tells us that the Q-vector-space spanned by 1, c_i and c_j has a basis of the form $\{1, c\}$. Since none of the c_i's lie in Q this means

(46) For all i and j, c_i is a Q-linear combination of 1 and c_j.

But we can perform the same calculation with the roles of z_i and $c_i z_i$ reversed, getting the conclusion that c_i^{-1} is also a Q-linear combination of 1 and c_j. This leads to a Q-linear relation among 1, c_i and c_i^{-1}, so c_i satisfies a quadratic equation over Q. Hence by (46), all c_j's lie in a common quadratic number field, say $Q(c)$.

To complete the proof of (ii), pick any $x \in G$. If $\dim U(\{z_1,\dots,z_r\} \cup \{x\}) = 0$, then the coefficients expressing x in terms of the z_i's are all rational (cf.(42)), and hence trivially in $Q(c)$. If not, then this dimension is 1, so by (42), the Q-vector-space spanned by 1 and the coefficients expressing x in terms of $z_1,\dots,z_r$ is spanned by 1 and one other element, which we can take to be the coefficient of z_i. Now pick a $j \neq i$, and apply the same considerations to $x + c_j z_j$. For this result to remain true, the space spanned by the original coefficients must have been $Q(c)$. So $Q(G) = Q(c)$, establishing (ii).

Conversely, assuming (ii) we can see from (42) that for any $\mathbb{R}$-basis $Y \subseteq G$ of $\mathbb{R}^r$, (45) will hold, from which (i) follows immediately. ||

We remark that the proof of (ii) $\Rightarrow$ (i) generalizes to show that for any finitely generated dense subgroup $G \subseteq \mathbb{R}^r$, a dense subgroup of G must have rank $\geq \left(1 + \frac{1}{[Q(G):Q] - 1}\right) r$.

The reverse implication (i) $\Rightarrow$ (ii) does not have the obvious sort of generalization. If all dense subgroups of a given G have ranks $\geq r+m$ for some m, this does not yield an upper bound on $Q(G)$ except when $m = r$. For

example, let d be an irrational number, without restriction on degree or algebraicity, and let $Q(c)$ be a quadratic extension of Q. If $X = \{x_1,\dots, x_r\}$ is a basis of $\mathbb{R}^r$, then the subgroup G generated by $x_1,\dots, x_r$, $cx_1,\dots, cx_{r-1}$, dx_r has $Q(G) = Q(c,d)$, which can have arbitrary degree, but has no dense subgroups of rank $< 2r-1$. To see this, project G onto $V(\{z_1,\dots,z_{r-1}\})$ along $V(\{z_r\})$. A dense subgroup $H \subseteq G$ must project onto a dense subgroup of the image, which must have rank $\geq 2(r-1)$ by Lemma 9. One can deduce from (39) that H must also have nonzero intersection with $V(\{z_r\})$, hence its rank is at least one more than that of its projection.

However one can prove

(47) Exercise: $G \subseteq \mathbb{R}^r$ is a finitely generated dense subgroup, and every dense subgroup of G has rank at least $r + m$, then $Q(G)$ must have a subfield F such that $1 < [F:Q] < \frac{r}{m-1} + 1$.

(Hint: if $[Q(c_i):Q] = n$, then $(c_i+1)^{-1},\dots, (c_i+n-1)^{-1}$ and 1 form a Q-basis of $Q(c_i)$.)

Finally, we note that in a different sense, all finitely generated dense subgroups do arise from finitely generated dense subgroups with $m = 1$:

(48) Exercise: Suppose $r \leq s$, and $G \subseteq \mathbb{R}^r$ is a dense subgroup free of rank $s+1$ as an abelian group. Then there is a dense subgroup $G' \subseteq \mathbb{R}^s$ free of rank $s+1$, such that the projection of $\mathbb{R}^s$ onto the first r coordinates takes G' onto G.

[Continue from p.55, point 8.]

R E F E R E N C E S

1. H. R. P. Ferguson and R. W. Forcade, Generalization of the Euclidean algorithm for real numbers to all dimensions higher than two, Bull.A. M. S. (new series) 1 (1979) no. 6, 912-914.

2. D. H. Fowler, Ratio in early Greek mathematics, Bull. A. M. S. (new series) 1 (1979) no. 6, 807-846.

Department of Mathematics
University of California
Berkeley, CA 94720

corrections and addenda to my "NOTES ON FERGUSON AND FORCADE'S GENERALIZED EUCLIDEAN ALGORITHM"

George M. Bergman

Since writing the above Notes (here denoted |0|), I have corresponded with Ferguson and Forcade. Most of the comments below are the products of this correspondence. They may be read after the whole of |0|, or each item can be read after the point of |0| referred to. Point 8, occupying the last 2/3 of these pages, may be considered an additional section of |0|, "§12$\frac{1}{2}$".

1. p.2, lines 7-11. My description of the Jacobi-Perron algorithm is incorrect. Rather, it can be described as acting on a tuple of real numbers by alternately (i) subtracting from every term after the first that integer multiple of the first term which gives the smallest positive remainder, and then (ii) shifting the first term to the last position. (Thus the parenthetical comment on this algorithm on p.43 lines 8-12 should also be deleted.)

Unfortunately, most authors seek to "take advantage" of the fact that the steps of the algorithm depend only on the ratio of the given numbers, by normalizing their n-tuple after each iteration so that its first term is 1, and then suppressing this redundant term. The cost of this simplification is a less transparent, less elegant, and computationally more difficult algorithm; though certainly for some purposes it is appropriate to consider normalized elements; e.g. the ergodic properties of the normalized algorithm, which would be meaningless for the non-normalized form, are studied in |6|. To further complicate things, some authors consider a single iteration of the algorithm to consist of the sequence of steps "(i),(ii)" while others use the sequence "(ii),(i)", yielding the different normalized forms: $(\alpha_1,\ldots,\alpha_n) \mapsto \left(\frac{\alpha_2 - [\alpha_2]}{\alpha_1 - [\alpha_1]},\ldots, \frac{\alpha_n - [\alpha_n]}{\alpha_1 - [\alpha_1]}, \frac{1}{\alpha_1 - [\alpha_1]}\right)$

respectively: $(\alpha_1,\ldots,\alpha_n) \mapsto \left(\frac{\alpha_2}{\alpha_1} - \left[\frac{\alpha_2}{\alpha_1}\right], \ldots, \frac{\alpha_n}{\alpha_1} - \left[\frac{\alpha_n}{\alpha_1}\right], \frac{1}{\alpha_1} - \left[\frac{1}{\alpha_1}\right]\right)$. See |5|, |6|.

An example showing that the Jacobi-Perron algorithm can fail to detect a relation of $\mathbb{Z}$-linear dependence (cf. p.31 lines 13-14) is given toward the end of point 8 below.

2. pp.7-9, §5. The algorithm I develop in this section (which works in dimensions 2, 3 and with some modification, 4) turns out to have been introduced by Barkley Rosser in |7|, though with some differences in point of view. Rosser did not note that finding what I have called a co-Diophantine relation among m+1 points of $\mathbb{R}^m$ was equivalent to finding a $\mathbb{Z}$-linear dependence relation among m+1 real numbers; in fact, the question of interest to him was not whether m+1 points of $\mathbb{R}^m$ satisfied one co-Diophantine relation, but whether it satisfied m such relations, i.e. whether it generated a <u>discrete</u> subgroup of $\mathbb{R}^m$, and if so, how to find the "smallest" $\mathbb{Z}$-basis. He considered such applications as minimizing values of quadratic forms. In |8|, |9| he also applies this algorithm to problems of approximating a ratio of several irrational numbers by a ratio of integers (cf. |0| beginning of §12 and also Cor.6 p.31 et seq.)

3. p.9, first three paragraphs of §4. These paragraphs are not a complete description of Ferguson and Forcade's algorithm, though the sentence which follows them may seem to imply that it is. What is true is that their algorithm is inductive, the algorithm for n-1 being used as a subroutine in the algorithm for n, and that it makes use of the max norm in selecting which vector to distinguish. For a more extensive presentation of their construction and its consequences than |1|, see |3|, |4|.

4. p.42, top paragraph. The approach I suggest there is nonsense, since the extended algorithm as I describe it, when applied to r+2 vectors in C^r, will ignore the last one until a co-Diophantine relation is found among the first

r+1 vectors, which cannot happen in this situation; and application of the algorithm to these r+1 vectors cannot give periodic result since K has degree > r+1. The question is somewhat academic at this point, since we don't know whether the algorithm becomes periodic even when the distinguished embedding is real. Nonetheless, here are two possible ways out: (i) Modify the algorithm so that it does not "ignore" weak indices. (For instance in (10) p.11 change the hypothesis "If $|\beta_i^{(i-1)}| > |\beta_{i+1}^{(i-1)}|$" to "If $|\beta_i^{(j)}| > |\beta_{i+1}^{(j)}|$, where j is the largest index $< i$ such that $s^{(j)} \neq s^{(i)}$". But other modifications might be superior.) (ii) Find units in K(i) (cf. preceding page) and apply the norm map $K(i) \to K$ to get units in K.

(5) pp.42-43, §12. The "candidate algorithm" discussed in this section, which I motivate by the task of detecting approximation to small-integer ratio, but suggest might also detect linear dependence and/or give units of algebraic number fields, turns out to have been introduced by Viggo Brun |10|. Brun proved that his algorithm would detect linear dependence in a 3-tuple $(\alpha_1,\alpha_2,\alpha_3)$. Forcade and Ferguson came up with the same algorithm independently and proved the same result, but also found examples showing that it may fail to detect linear dependence relations in 4-tuples and that when applied to a linearly independent 4-tuple x, it may not "approximate" in their sense. I.e., the rows of the successive matrices it produces will not in general approach arbitrarily close to the 1-dimensional subspace $x\mathbb{R}$. (Like the Jacobi-Perron algorithm, it has the property that the distances from these rows to $x\mathbb{R}$, divided by the lengths of the rows, do approach zero. This is a much weaker condition, which Ferguson and Forcade express by saying that such algorithms give "angular approximations".) I will present the counterexamples in point (8) below.

(6) pp.44-45. Warren Dicks (Bedford College, London) has tried out this algorithm on $(1, n^{1/3}, n^{2/3})$ for $n=2, 3, 4$ by computer, using 115 decimal place accuracy. For $n = 3$, there is still no trace of periodicity! For $n = 4$ (which, of course, corresponds to the same field extension as $n = 2$), he does get a periodic pattern:

$1:(1:1:1.1.1.2:1:1:1:1.6.1:1:1:1:2.1.1.1:1:1.1:2.1.1:7:1.1.2:1.)^{\omega}$

Note that this is again "palindromic", with centers of symmetry at the "6" and the "7".

(7) p.48, Lemma 9. The proof should begin, "Assume (i)".

(8) Ferguson and Forcade give in [3], p.7 an example of a $\mathbb{Z}$-linearly dependent 4-tuple for which they state without proof that Brun's algorithm (see point 5 above) fails to terminate, and a $\mathbb{Z}$-linearly independent 4-tuple for which they state that it fails to approximate. I shall sketch here some results which motivate and explain these examples, and give an example similar to the first of these for the Jacobi-Perron algorithm.

In examining Brun's algorithm, let us fix an integer n, and restrict our attention to n-tuples of real numbers $(\alpha_1, \ldots, \alpha_n)$ satisfying

$$0 \leq \alpha_1 \leq \ldots \leq \alpha_n, \quad 0 < \alpha_n. \tag{49}$$

which we shall call Brun n-tuples. The set of all Brun n-tuples will be denoted B, and the set of Brun n-tuples modulo scalar multiplication, $\bar{B}$. The latter set can be identified with the set of Brun n-tuples having $\alpha_n = 1$ ("normalized" n-tuples; cf. point 1), which we see from the defining condition (49) is compact. We shall call a Brun n-tuple nondegenerate if $\alpha_{n-1} \neq 0$.

Next, given positive integers $i \leq n-1$, $j < \omega$, we define the Brun matrix $b(i,j)$ to be the matrix which acts on row vectors $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ by

subtracting $j\alpha_{n-1}$ from the term α_n, and then moving this decreased term to the i^{th} place, giving $(\alpha_1, \ldots, \alpha_{i-1}, \alpha_n - j\alpha_{n-1}, \alpha_i, \ldots, \alpha_{n-1})$. We can see that

(50) for any nondegenerate Brun n-tuple α there exists a Brun matrix $b(i,j)$ such that $\alpha\, b(i,j)$ is again a Brun n-tuple.

A Brun matrix $b(i,j)$ with the above property will be called "admissible" for α, and a sequence of Brun matrices $b(i_1, j_1), b(i_2, j_2), \ldots$ (finite or infinite) will be called admissible for α if each $b(i_r, j_r)$ is admissible for $\alpha\, b(i_1, j_1) \ldots b(i_{r-1}, j_{r-1})$. Then Brun's algorithm consists of applying to a Brun n-tuple successive admissible Brun matrices, continuing indefinitely unless one reaches a degenerate n-tuple (if one is testing for commensurability) or an n-tuple with $\alpha_1 = 0$ (if one is testing for $\mathbb{Z}$-linear dependence).

(It is clear that the Brun matrix in (50) is "usually" unique. In fact, j is unique unless $\alpha_n = m\,\alpha_{n-1}$ $(m \geq 2)$ in which case there are the two possibilities $j = m, m-1$; i is unique unless $\alpha_n - j\alpha_{n-1}$ is equal to some α_h $(h \leq n-2)$. It follows that in any case of nonunique admissible $b(i,j)$, we get for $\alpha\, b(i,j)$ an n-tuple for which one of the "$\leq$" signs in (49) becomes "$=$"; and one can show that such an n-tuple must, after a finite number of further admissible Brun operations yield an n-tuple with $\alpha_1 = 0$. From this it will follow that in the class of examples to be considered, one has uniqueness at every step; but I shall not belabor the details in this sketch.)

We now ask, given an arbitrary finite or infinite sequence of Brun matrices, what can be said about the class of $\alpha \in B$ admitting this sequence. To study this question we must look at inverses of Brun matrices. These are even more nicely behaved than the matrices themselves. We note that $b(i,j)^{-1}$ acts on an

n-tuple by increasing the i^{th} term of the n-tuple by j times the last term, and then relocating this increased term in the last position. Clearly

$$B\ b(i,j)^{-1} \subseteq B. \tag{51}$$

(Thus (50) says that the sets $B\ b(i,j)^{-1}$ cover B except for the degenerate points $(0,\ldots,0,\alpha_n)$, and our remarks on uniqueness say that points are covered uniquely except at the boundaries of the images $B\ b(i,j)^{-1}$.)

The actions of the $b(i,j)^{-1}$ clearly extend to the factor-set $\bar{B}$. Given an infinite sequence of Brun matrices $b(i_h, j_h)$ $(h=1,2,\ldots)$ we see by the compactness of $\bar{B}$ that the intersection of the chain

$$\bar{B} \supseteq \bar{B}\ b(i_1, j_1)^{-1} \supseteq \bar{B}\ b(i_2, j_2)^{-1}\ b(i_1, j_1)^{-1} \supseteq \ldots \tag{52}$$

is nonempty. By metric considerations one finds that it reduces to a single point $\alpha\mathbb{R} \in \bar{B}$, i.e. there is a point $\alpha \in B$ unique up to scalars which admits the given sequence. Note that if none of the i_h equals 1, then the algorithm is essentially being performed on the (n-1)-tuple $(\alpha_2,\ldots,\alpha_n)$, and in fact one finds that in this case $\alpha_1 = 0$. On the other hand, if $i_h = 1$ for at least n-1 values of h, one can show that $\alpha_1 \neq 0$. In particular, if $i_h = 1$ for infinitely many values of h, then neither α nor any of the iterates $\alpha b(i_1, j_1)$, $\alpha b(i_1,j_1)\ b(i_2,j_2)$ etc. has first term 0. (And one can deduce that every step of the Brun algorithm applied to this α is unique.)

We now consider a finite product of Brun matrices, $b = b(i_r,j_r)\ldots b(i_1,j_1)$, where $r \geq 1$ and at least one i_h equals 1. Applying the above considerations to the infinite sequence

$$b(i_1, j_1), \ldots, b(i_r,j_r), b(i_1,j_1),\ldots, b(i_r,j_r), \ldots, \tag{53}$$

we conclude that there will be a unique point $\tau\mathbb{R} \in \bar{B}$ invariant under b,

i.e. a unique (up to scalars) eigenvector τ for b in B. The Brun algorithm will be periodic on τ, with sequence of operations (53).

Let us now suppose for simplicity that the characteristic polynomial of b^{-1} has distinct roots, $\lambda_1, \ldots, \lambda_n$, with eigenvectors $\theta^{(1)}, \ldots, \theta^{(n)}$. Note that

$$\lambda_1 \cdots \lambda_n = \det b^{-1} = \pm 1. \tag{54}$$

Now almost all points of $\mathbb{R}^n$, and hence almost all points of B, will have the form

$$\theta = c_1\theta^{(1)} + \cdots + c_n\theta^{(n)}, \text{ with } c_1, \ldots, c_n \neq 0. \tag{55}$$

If we iterate b^{-1} on θ, we see that the eigenvector(s) corresponding to the eigenvalue(s) of largest absolute value will "dominate". On the other hand, from our considerations concerning (52) we can see that the images in $\bar{B}$ of the iterates $\theta\, b^{-n}$ must converge to the unique fixed point $\tau\mathbb{R}$. It is easily deduced that there must be a unique eigenvalue of maximal absolute value, say λ_1, that $\theta^{(1)} = \tau$ (up to scalars), and that $\lambda_1 > 1$. This is already a nontrivial condition on the characteristic polynomials of products of inverses of Brun matrices.

Now it is not hard to see that if Brun's algorithm is to approximate τ in the strong sense of Ferguson and Forcade (i.e. iterates $\theta\, b^{-n}$ actually approach the line $\tau\mathbb{R}$ in B, not merely in $\bar{B}$), then all the eigenvalues λ_i other than λ_1 must have absolute value < 1. If one examines the 4-tuple they give as an example of non-approximation (rearranged as in (49)) one finds that it is a fixed point of $b = b(3,2)\, b(1,1)\, b(2,1)\, b(3,1)$. The characteristic polynomial of b^{-1}, $t^4 - 4t^3 - 6t^2 - t - 1$, has eigenvalues $\lambda_1 \approx 5.198$, $\lambda_2 \approx -1.142$, $\lambda_3, \lambda_4 = -0.028 \pm 0.405\, i$. Since λ_2 is in fact strictly greater than 1, "approximation" fails badly: the iterates of (55) actually move away from the line $\tau\mathbb{R}$. (For the record, their example is, up to a scalar: $(3\lambda_1^2,\ 4\lambda_1^2+1,\ \lambda_1^3+\lambda_1,\ 3\lambda_1^3)$. But they write the terms in reverse order, and express them in terms of λ_1^{-1}, the least eigenvalue of b.)

Consider next what can happen if the characteristic polynomial of b^{-1} is not irreducible. Since the eigenvector τ is a solution of $\tau(b^{-1} - \lambda_1 I) = 0$, it can be taken to have entries in the field $Q(\lambda_1)$. But in this case λ_1 has degree $< n$, so the entries of τ are linearly dependent over $\mathbb{Z}$. But by construction, Brun's algorithm applied to τ is periodic and never gives an n-tuple with $\alpha_1 = 0$, hence it never "discovers" the linear dependence relation among the entries of τ.

There is in fact a rather straightforward approach to finding examples of this sort. Start with an n-tuple ξ of positive and negative integers, e.g. $(1,1,0,-1)$. (Of course, ξ is not a Brun n-tuple.) Try to find a product b^{-1} of inverse Brun matrices carrying ξ to $-\xi$; in the above example, $b(3,1)^{-1}\ b(2,1)^{-1}\ b(1,1)^{-1}$ does so. Then b^{-1} will have an eigenvalue -1, and hence characteristic polynomial divisible by $t+1$, and in particular, not irreducible. Indeed, the example they give has b^{-1} as above, with characteristic polynomial $(t+1)(t^3 - 3t + 1)$, and eigenvector $\tau = (1, \lambda_1 - 1, \lambda_1^2 - 2\lambda_1, \lambda_1) \in B$, where λ_1 is the largest root of the cubic factor.

The Jacobi-Perron algorithm can be examined in a similar way. Let us consider its basic step to consist of subtracting from the terms $\alpha_2, \ldots, \alpha_n$ of an n-tuple α the largest integer multiples of α_1 giving nonnegative remainders (say $\alpha_2 - j_2\alpha_1, \ldots, \alpha_n - j_n\alpha_1$), then moving α_1 to the last position. Then after the second step it will always give n-tuples of non-negative real numbers with α_n greater than or equal to all of $\alpha_1, \ldots, \alpha_{n-1}$. Let us call the set of such n-tuples J. The steps from this point on will also satisfy

$$0 \leq j_h \leq j_n \quad (h=2,\ldots,n-1), \quad 1 \leq j_n. \tag{56}$$

Let us call the matrix which takes $(\alpha_1, \ldots, \alpha_n)$ to $(\alpha_2 - j_2\alpha_1, \ldots, \alpha_n - j_n\alpha_1, \alpha_1)$ (subject to (56)) $p(j_2, \ldots, j_n)$. It is again true that every $\alpha \in J$ admits a matrix $p(j_2, \ldots, j_n)$, unique except in marginal cases. But it is not true

that $J\,p(j_2,\ldots,j_n)^{-1}$ is always contained in J; this is so if and only if $1 < j_n$ and $j_h < j_n$ $(h = 2,\ldots,n-1$; cf. (56)). Nevertheless, with a little experimentation one finds that for $\xi=(1,-1,2,-1)$ one has $\xi\,p(0,1,1) = -\xi$; and $p(0,1,1)$ has an eigenvector in J, namely $(\lambda_1^2, \lambda_1, \lambda_1^2+1, 2\lambda_1^2-\lambda_1+1)$, where $\lambda_1 \approx 1.755$ is the real root of $t^3 - 2t^2 + t - 1$, this times $t+1$ being the characteristic polynomial of $p(0,1,1)^{-1}$. One concludes that the Jacobi-Perron algorithm fails to detect the $\mathbb{Z}$-linear dependence of the terms of this 4-tuple.

One might ask whether versions of these same algorithms which choose their coefficients to give remainders with smallest <u>absolute values</u> at each step, rather than least <u>positive</u> remainders, might do better. Here both algorithms come to have the difficulty of not satisfying the analog of (51). For instance, the natural class of matrices to look at for the Jacobi-Perron case are those $p(j_2,\ldots, j_n)$ (defined formally as above) with <u>arbitrary</u> integers $j_2,\ldots, j_n$, subject to $2|j_h| \leq |j_n|$, $2 \leq |j_n|$, but the analog of (51) holds only if all these inequalities are strict. This means that if one uses operations not satisfying these stronger inequalities, then one must check that the string of operations in question is in fact admissible on the eigenvector τ that one ends up with. I have not investigated these questions in depth, but here, at least, is an example for which this version of the Jacobi-Perron algorithm fails to find linear dependence: Taking $\xi=(-1, 2, -3, -1)$ we find that $\xi\,p(-1,1,2) = -\xi$. The characteristic polynomial of $p(-1,1,2)^{-1}$ is $(t+1)(t^3-2t^2+3t-1)$. Letting $\lambda_1 \approx 7/3$ denote the largest root of the latter factor, we find that $(\lambda_1^2, -\lambda_1^2+\lambda_1, \lambda_1^2-\lambda_1+1, 3\lambda_1^2 - 2\lambda_1 + 1) = \tau$ is carried by the algorithm to $\tau\,p(-1,1,2) = \lambda_1^{-1}\tau$, so no zero-term ever appears, though the entries are $\mathbb{Z}$-linearly dependent.

Note that in the above examples, by creating periodicity we were able to know the complete behavior of these algorithms on certain elements, and thus

prove that certain other phenomena failed to occur. It would seem that to prove that one of these algorithms, when applied to some basis of some algebraic number field, did not show periodicity, should require much more subtle methods.

REFERENCES

|0| G. M. Bergman, Notes on Ferguson and Forcade's generalized Euclidean algorithm (51 pp.)

|3| H. R. P. Ferguson and R. W. Forcade, Generalization of the Euclidean algorithm to all dimensions higher than 2 (preprint, 21 pp.)

|4| ——, New $\mathbb{Z}$-linear dependence algorithms (preprint, 30 pp.)

|5| L. Bernstein, The Jacobi-Perron Algorithm, its theory and applications. SLN 207 (1971) 161 pp.

|6| F. Schweiger, The metrical theory of Jacobi-Perron algorithm, SLN 334 (1973) 111 pp.

|7| (J.) Barkley Rosser, A generalization of the Euclidean algorithm to several dimensions, Duke Math. J. 9(1942)59-95.

|8| ——, A note on the linear diophantine equation, Am. Math. Monthly 48(1941)662-666.

|9| ——, Generalized ternary continued fractions, Am. Math. Monthly 57(1950)528-535.

|10| V. Brun, En generalisation av kjedebrøken, Videnskapsselskapets Skrifter, Oslo, 1919 og 1920.

|11| ——, Music and ternary continued fractions, Norske Vid. Selsk. Forh., Trondheim, 23(1950) 38-40. (MR 12, p.675).