PMATH 744 - DIOPHANTINE INEQUALITIES

Content

One of the major mathematical triumphs of of the last century is Schmidt's Subspace Theorem. We plan to put it in context and derive some of the consequences of it.

Let $\alpha \in \mathbb{R}$. **Basic question:** How well can α be approximated by rational numbers? Since the rational numbers are dense we know that they can be approximated as well as we want. So we ask the more interesting question: How well can α be approximated by rationals p/q with $p, q \in \mathbb{Z}$, $q > 0$ in terms of q?

Theorem 1. *(1842 Dirichlet). Let Q be a real number with $Q > 1$. There exists $p, q \in \mathbb{Z}$ with $1 \leq q < Q$ such that*

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{qQ}.$$

Notes:

-This tells us that $\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}$.

-If α is irrational then Dirichlet's Theorem shows that there are infinitely many rationals p/q for which

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}.$$

Clearly this is not true if α is rational.

Algorithmic question: How do we find these good rational approximations to α?

In particular, can it be done efficiently? Yes, because of the continued fraction algorithm.

Given α we produce we produce two sequences $(\alpha_0, \alpha_1, \dots)$ and $(a_0, a_1, \dots)$ with the α_i's in $\mathbb{R}$ and the a_i's in $\mathbb{Z}$ by the following rules:

Put $\alpha = \alpha_0$ and $a_i = [\alpha_i]$ for $i = 0, 1, 2, \dots$ and $\alpha_{i+1} = (\alpha_i - [\alpha_i])^{-1}$ for $i = 0, 1, 2, \dots$ provided that $\alpha_i \neq [\alpha_i]$; here for any $x \in \mathbb{R}$ we denote the greatest integer less than or equal to x by $[x]$. Note that if $\alpha_i = a_i$ for some i we stop the process. In this case $\alpha \in \mathbb{Q}$. We put, for $n = 0, 1, 2, \dots$

$$\frac{p_n}{q_n} = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots}}.$$

here we suppose that $(p_n, q_n) = 1$ and $q_n > 0$.

We then have $\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2}$ for $n = 0, 1, 2 \dots$. Further Legendre showed that if $\left|\alpha - \frac{p}{q}\right| < \frac{1}{2q^2}$ then there exists n such that $\frac{p}{q} = \frac{p_n}{q_n}$.

The $\frac{p_n}{q_n}$ are known as the convergents to α.

Remark:

If the a_i's are all eventually 1 then $\lim_{n\to\infty} q_n^2 \left|\alpha - \frac{p_n}{q_n}\right| = \frac{1}{\sqrt{5}}$.

Given $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. **Interesting question:** How well can we approximate the α_i's by rationals of the same denominator?

Theorem 2. *(Dirichlet) Suppose that $\alpha_1, \dots, \alpha_n$ are real numbers and $Q > 1$ is an integer. Then there exists $q, p_1, \dots, p_n$ with $1 \leq q < Q^n$ for which*

$$\left|\alpha_i - \frac{p_i}{q}\right| \leq \frac{1}{qQ} \qquad \text{for } i = 1, \dots, n.$$

Corollary 1. *Suppose that at least one of the α_i's in Theorem 2 is irrational. Then there exist infinitely many $n+1$-tuples of coprime inegers $(q, p_1, \dots, p_n)$ with $q > 0$ for which*

$$\left|\alpha_i - \frac{p_i}{q}\right| < \frac{1}{q^{1+\frac{1}{n}}} \qquad \text{for } i = 1, \dots, n.$$

Proof:

Note that since one of the α_i's is irrational, Theorem 2 yields infinitely many $n+1$-tuples. Further we may assume that $(q, p_1, \dots, p_n)$ are coprime by factoring out the common factor if necessary. □

Algorithmically there is no "good" way of finding these approximations in sense of the continued fraction algorithm. However there are algorithms which produce some good approximations.

Note: Corollary 1 tells us that the linear forms satisfy $|q\alpha_i - p_i| < \frac{1}{q^{1/n}}$

Theorem 3. *Suppose that $\alpha_i, \dots, \alpha_n$ are real numbers and that Q is an integer with $Q > 1$. Then there exist integers p and $q_1, \dots, q_n$ with $1 \leq \max\limits_{i=1,\dots,n} |q_i| < Q^{1/n}$ for which*

$$|q_1\alpha_1 + \cdots + q_n\alpha_n - p| \leq \frac{1}{Q}.$$

Corollary 2. *Suppose that $\alpha_1, \dots, \alpha_n$ are real numbers with $1, \alpha_1, \dots, \alpha_n$ linearly independent over $\mathbb{Q}$. Then there exist infinitely many coprime $n+1$-tuples $(p, q_1, \dots, q_n)$ such that if we put $q = \max\limits_{i=1,\dots,n} |q_i|$ we have $q \geq 1$ and $|q_1\alpha_1 + \cdots + q_n\alpha_n - p| < \frac{1}{q^n}$.*

Proof:

Since $1, \alpha_1, \dots, \alpha_n$ are linearly independent over $\mathbb{Q}$ we see that $q_1\alpha_1 + \cdots + q_n\alpha_n - p \neq 0$ and so Theorem 3 produces infinitely many coprime $n+1$-tuples of the desired form. □

We can combine Theorems 1,2,3 into a single theorem:

Theorem 4. *(1842) Dirichlet*
Suppose that α_{ij} are real numbers for $1 \leq i \leq n$, $1 \leq j \leq m$ and that Q is an integer with $Q > 1$. Then there exist integers $q_1, \dots, q_m$ and $p_1, \dots, p_n$ with $1 \leq \max\limits_{i=1,\dots,m} |q_i| < Q^{n/m}$ and

$$|\alpha_{i1}q_1 + \cdots + \alpha_{im}q_m - p_i| \leq \frac{1}{Q} \qquad \text{for } i = 1, \dots, n.$$

Corollary 3. *Suppose that $1, \alpha_{i1}, \dots, \alpha_{im}$ are linearly independent over $\mathbb{Q}$ for some i with $1 \leq i \leq n$. Then there exist infinitely many coprime $n+m$-tuples $(p_1, \dots, p_n, q_1, \dots, q_m)$ such that with $1 \leq q = \max\limits_{i=1,\dots,m} |q_i|$ we have*

$$|\alpha_{i1}q_1 + \cdots + \alpha_{im}q_m - p_i| < \frac{1}{q^{m/n}} \qquad \text{for } i = 1, \dots, n.$$

Notation:

Recall that for $x \in \mathbb{R}$ we denote the greatest integer less than or equal to x by $[x]$ and the fractional part of x by $\{x\}$ so $\{x\} = x - [x]$. Also we let $\|x\|$ denote the distance from x to the nearest integer. So then $\|x\| = \min(\{x\}, 1 - \{x\})$.

For $n \in \mathbb{Z}^+$, u^n denotes the unit cube $u^n = \{(t_1, \dots, t_n) \mid 0 \le t_i < 1, \quad \text{for } i = 1, \dots, n\}$ and $\overline{u}^n = \{(t_1, \dots, t_n) \in \mathbb{R}^n \mid 0 \le t_i \le 1, \quad \text{for } i = 1, \dots, n\}$

Proof:

(Theorem 4) Let us divide $\overline{u}^n$ into Q^n subcubes of side length $\frac{1}{Q}$ in such a way that the union of the cubes is $\overline{u}^n$ and so that the intersection of any two subcubes is either a face, edge or point of a subcube or nothing.

We now consider the points in $\overline{u}^n$ of the form $(\{\alpha_{11}x_1 + \ldots + \alpha_{1m}x_m\}, \dots, \{\alpha_{n1}x_1 + \ldots + \alpha_{nm}x_m\})$ where the x_i's are integers with $0 \le x_i < Q^{n/m}$ for $i = 1, \dots, m$. The sequence of such points has Q^n elements. If we include the points $(1, 1, \dots, 1)$ we get $Q^n + 1$ points and so two of them are in the same subcube. These points are say $(\{\alpha_{11}x_1 + \ldots + \alpha_{1m}x_m\}, \dots, \{\alpha_{n1}x_1 + \ldots + \alpha_{nm}x_m\})$ and $(\{\alpha_{11}x'_1 + \ldots + \alpha_{1m}x'_m\}, \dots, \{\alpha_{n1}x'_1 + \ldots + \alpha_{nm}x'_m\})$ or $(\alpha_{11}x_1 + \ldots + \alpha_{1m}x_m - y_1, \dots, \alpha_{n1}x_1 + \ldots + \alpha_{nm}x_m - y_m)$ and $(\alpha_{11}x'_1 + \ldots + \alpha_{1m}x'_m - y'_1, \dots, \alpha_{n1}x'_1 + \ldots + \alpha_{nm}x'_m - y'_m)$. Then

$$|\alpha_{11}(x_1 - x'_1) + \ldots + \alpha_{1m}(x_m - x'_m) - (y_1 - y'_1)| \le \frac{1}{Q}$$

$$\vdots$$

$$|\alpha_{n1}(x_1 - x'_1) + \ldots + \alpha_{nm}(x_m - x'_m) - (y_n - y'_n)| \le \frac{1}{Q}$$

and so the result follows on taking $q_i = x_i - x'_i$ and $p_j = y_j - y'_j$ for $i = 1, \dots, m$ and $j = 1, \dots, n$. Notice that $|q_i| < Q^{n/m}$ for $i = 1, \dots, m$ since $0 \le x_i < Q^{n/m}$. □

Notation:

We denote points in $\mathbb{R}^n$ by $\underline{\underline{x}}$ so $\underline{\underline{x}} = (x_1, \dots, x_n)$ for $x_i \in \mathbb{R}$, $i = 1, \dots, n$.

We put $\overline{|\underline{\underline{x}}|} = \max_{i=1,\dots,n}(|x_i|)$.

If $\underline{\underline{x}} = (x_1, \dots, x_n)$ is such that $x_i \in \mathbb{Z}$ for $i = 1, \dots, n$ then we say that $\underline{\underline{x}}$ is an integer point.

For any set T in $\mathbb{R}^n$ and $\underline{\underline{x}} \in \mathbb{R}^n$ we put $T + \underline{\underline{x}} = \{\underline{\underline{t}} + \underline{\underline{x}} \mid \underline{\underline{t}} \in T\}$.

Further for $\lambda \in \mathbb{R}^+$ we denote λT by $\lambda T = \{\lambda \underline{\underline{t}} \mid \underline{\underline{t}} \in T\}$ here $\lambda \underline{\underline{t}} = (\lambda t_1, \dots, \lambda t_n)$.

Theorem 5. *(Blichfeldt 1914) Let P be a non-empty set of points in $\mathbb{R}^n$ which is invariant under translation by integer points and we suppose that P has precisely N points in u^n. Let A be a subset of $\mathbb{R}^n$ of positive Lebesque measure $\mu(A)$. Then there exists an $x \in u^n$ such that $A + \underline{\underline{x}}$ contains at least $N\mu(A)$ points of P. Further if A is compact then there exists an $\underline{\underline{x}} \in u^n$ such that $A + \underline{\underline{x}}$ contains more than $N\mu(A)$ points of P.*

Proof:

For any set S in $\mathbb{R}^n$ let $\nu(S)$ denote the number of points of P in S. Let $\underline{\underline{P}}_1, \dots, \underline{\underline{P}}_N$ denote the points of P in u^n. Let $P_1, \dots, P_N$ be defined by $P_i = \{\underline{\underline{P}}_i + \underline{\underline{q}} \mid \underline{\underline{q}} \text{ an integer point in } \mathbb{R}^n\}$. Note that $P = P_1 \cup P_2 \cup \ldots \cup P_N$ and $P_i \cap P_j = \phi$ if $i \ne j$ since P is invariant under translation by integer points.

Now let $\nu_i(S)$ denote the number of points of P_i in S for $i = 1, \dots, n$. Observe that

$$\nu(S) = \sum_{i=1}^{N} \nu_i(S).$$

Let χ denote the characteristic function of A. Then for $i = 1, \dots, n$ and $\underline{\underline{x}} \in \mathbb{R}^n$

$$\nu_i(A + \underline{\underline{x}}) = \sum_{\underline{\underline{g}}} \chi(\underline{\underline{P}}_i + \underline{\underline{g}} - \underline{\underline{x}})$$

where the sum is over all integer points $\underline{\underline{g}}$. We have

$$\int_{u^n} \nu_i(A + \underline{\underline{x}})\, d\underline{\underline{x}} = \int_{u^n} \sum_{\underline{\underline{g}}} \chi(\underline{\underline{P}}_i + \underline{\underline{g}} - \underline{\underline{x}})\, d\underline{\underline{x}} = \int_{\mathbb{R}^n} \chi(\underline{\underline{z}})\, d(\underline{\underline{z}}) = \mu(A).$$

Therefore

$$\int_{u^n} \nu(A + \underline{\underline{x}})\, d\underline{\underline{x}} = N\mu(A),$$

and so for some $\underline{\underline{x}} \in u^n$ we have

$$\nu(A + \underline{\underline{x}}) \geq N\mu(A).$$

Suppose now that A is compact. If $N\mu(A)$ is not an integer the result is immediate so we may assume $N\mu(A) = h \in \mathbb{Z}^+$. For $k = 1, 2, \dots$ we define A_k by $A_k = (1 + \frac{1}{k})A$. By what we have just proved there is a sequence $(\underline{\underline{x}}_k)_{k=1}^{\infty}$ of points in u^n for which $\nu(A_k + \underline{\underline{x}}_k) \geq h + 1$, for $k = 1, 2, \dots$. Since the $\underline{\underline{x}}_k$'s are in u^n we may extract a convergent subsequence $(\underline{\underline{x}}_{k_j})_{j=1}^{\infty}$ which converges to $\underline{\underline{x}}'$. All of the sets $A_{k_j} + \underline{\underline{x}}_{k_j}$ are uniformly bounded and so contain only finitely many points of P. Since each set contains $h + 1$ points of P there exist $h + 1$ points $\underline{\underline{\mu}}_1, \dots, \underline{\underline{\mu}}_{h+1}$ which occur in infinitely many of these sets.

Since A is compact so is $A + \underline{\underline{x}}'$ and so either $\underline{\underline{\mu}}_1, \dots, \underline{\underline{\mu}}_{h+1}$ are all in $A + \underline{\underline{x}}'$ or there is one of them, say $\underline{\underline{\mu}}_1$, which is not and then it is a positive distance from $A + \underline{\underline{x}}'$. But the maximum distance from a point of $A_{k_j} + \underline{\underline{x}}_{k_j}$ to $A + \underline{\underline{x}}'$ tends to zero as $j \to \infty$ since $\underline{\underline{x}}_{k_j} \to \underline{\underline{x}}$ and $1 + \frac{1}{k_j} \to 1$. This is a contradiction and so $\underline{\underline{\mu}}_1, \dots, \underline{\underline{\mu}}_{h+1}$ are in $A + \underline{\underline{x}}'$. We now choose $\underline{\underline{g}} \in \mathbb{Z}^n$ so that $\underline{\underline{x}}' - \underline{\underline{g}} \in u^n$ and then $\nu(A + \underline{\underline{x}}' - \underline{\underline{g}}) \geq h + 1$ as required. □

Theorem 4a: Theorem 4 holds for $\mathbb{Q} \in \mathbb{R}$ with $Q > 1$.

Proof:

Let P be the set of points in $\mathbb{R}^n$ of the form

$$(\alpha_{11}x_1 + \cdots + \alpha_{1m}x_m, \dots, \alpha_{n1}x_1 + \cdots + \alpha_{nm}x_m) + \underline{\underline{g}}$$

where $\underline{\underline{g}}$ is an integer point in $\mathbb{R}^n$ and $x_i \in \mathbb{Z}$ with $0 \leq x_i < Q^{n/m}$ for $i = 1, \dots, m$. P is invariant by translation by integer points. Let N be the number of points P in u^n. Then either $N \geq Q^n$ or two points $(\alpha_{11}x_1^{(i)} + \ldots + \alpha_{1m}x_m^{(i)}, \dots, \alpha_{n1}x_1^{(i)} + \ldots + \alpha_{nm}x_m^{(i)})$ for $i = 1, 2$ differ by an integer point. In the latter case we are done.

Let $A = \{t_1, \dots, t_n) | 0 \leq t_i \leq \frac{1}{Q}$ for $i = 1, \dots, n\}$. A is compact and $\mu(A) = \frac{1}{Q^n}$. By Blichfeldt's Theorem there is a point $\underline{\underline{x}} \in u^n$ for which $A + \underline{\underline{x}}$ contains more than $N\frac{1}{Q^n} = 1$ point of P. Thus there are two points of P in $A + \underline{\underline{x}}$ and the result follows on taking the difference of the coordinates of these points. □

Definition:

A set S in $\mathbb{R}^n$ is said to be symmetric about the origin $\underline{\underline{0}}$ if whenever $\underline{\underline{x}} \in S$ then $-\underline{\underline{x}} \in S$.

S is said to be convex if whenever $\underline{\underline{x}}$ and $\underline{\underline{y}}$ are in S then the line segment joining $\underline{\underline{x}}$ and $\underline{\underline{y}}$ is also in S. In particular $\underline{\underline{x}}, \underline{\underline{y}} \in S$ implies $\lambda\underline{\underline{x}} + (1-\lambda)\underline{\underline{y}} \in S$ for $0 \le \lambda \le 1, \lambda \in \mathbb{R}$.

Theorem 6. *(Minkowski's Convex Body Theorem, 1896)*
Let A be a convex set in $\mathbb{R}^n$ which is bounded, symmetric about the origin and has a positive volume $\mu(A)$. If $\mu(A) > 2^n$ or A is compact and $\mu(A) \ge 2^n$ then there is a non-zero integer point in A different from $\underline{\underline{0}}$.

Proof:

Notice that either $\mu(\frac{1}{2}A) > 1$ or $\frac{1}{2}A$ is compact and $\mu(\frac{1}{2}A) \ge 1$. We now apply Blichfeldt's Theorem to the set $\frac{1}{2}A$ where P is the set of integer points in $\mathbb{R}^n$. Thus there is an $\underline{\underline{x}} \in \mathbb{R}^n$ such that $\frac{1}{2}A + \underline{\underline{x}}$ contains two integer points, say $\underline{\underline{g}}_1$ and $\underline{\underline{g}}_2$. Thus there exist $\underline{\underline{x}}_1$ and $\underline{\underline{x}}_2 \in A$ such that $\frac{1}{2}\underline{\underline{x}}_1 + \underline{\underline{x}} = \underline{\underline{g}}_1$ and $\frac{1}{2}\underline{\underline{x}}_2 + \underline{\underline{x}} = \underline{\underline{g}}_2$. By symmetry $-\underline{\underline{x}}_2 \in A$ and by convexity $\frac{1}{2}\underline{\underline{x}}_1 + \frac{1}{2}(-\underline{\underline{x}}_2) = \underline{\underline{g}}_1 - \underline{\underline{g}}_2$ is also in A. But $\underline{\underline{g}}_1 \ne \underline{\underline{g}}_2$ and so $\underline{\underline{g}}_1 - \underline{\underline{g}}_2$ is a non-zero integer point in A. □

Note: The estimate for $\mu(A)$ cannot be weakened since $A = \{(t_1, \dots, t_n) \in \mathbb{R}^n \mid |t_i| < 1 \text{ for } i = 1, \dots, n\}$ is convex, symmetric and bounded with $\mu(A) = 2^n$ and yet the only integer point in A is $\underline{\underline{0}}$.

Theorem 7. *(Minkowski's Linear Forms Theorem)*
Let $B = (\beta_{ij})$ be an $n \times n$ matrix with entries in $\mathbb{R}$ and non-zero determinant. Let $c_1, \dots, c_n$ be positive real numbers with $c_1 \cdots c_n \ge |\det B|$. Then there exists an integer point $\underline{\underline{x}} = (x_1, \dots, x_n)$ with $\underline{\underline{x}} \ne \underline{\underline{0}}$ such that

$$|\beta_{i1}x_1 + \cdots + \beta_{in}x_n| < c_i \quad \text{for } i = 1, \dots, n-1$$

and

$$|\beta_{n1}x_1 + \cdots + \beta_{nn}x_n| \le c_n$$

Proof:

Put $L_i(\underline{\underline{x}}) = \beta_{i1}x_1 + \cdots + \beta_{in}x_n$ for $i = 1, \dots, n$ and $\widetilde{L}_i(\underline{\underline{x}}) = \frac{1}{c_i}L_i(\underline{\underline{x}})$ for $i = 1, \dots, n$. In particular, we wish to solve the system

$$|\widetilde{L}_i(\underline{\underline{x}})| < 1 \quad \text{for } i = 1, \dots, n-1$$

and

$$|\widetilde{L}_n(\underline{\underline{x}})| \le 1 \quad \text{for } \underline{\underline{x}} \in \mathbb{Z}^n - \{\underline{\underline{0}}\}.$$

The absolute value of the determinant associated with the system $(\widetilde{L}_i)_{i=1,\dots,n}$ is at most 1. Therefore we may assume without loss of generality that $c_1 = \cdots = c_n = 1$ and $0 < |\det B| \le 1$.

Let A be the set of $\underline{\underline{x}} \in \mathbb{R}^n$ for which $|L_i(\underline{\underline{x}})| \le 1$ for $i = 1, \dots, n$. Note that A is bounded, symmetric and compact. A is also convex since if λ is a real number with $0 \le \lambda \le 1$ and $\underline{\underline{x}}_1$ and $\underline{\underline{x}}_2$ are in A then

$$\begin{aligned} |L_i(\lambda\underline{\underline{x}}_1 + (1-\lambda)\underline{\underline{x}}_2)| &\le |L_i(\lambda\underline{\underline{x}}_1)| + |L_i((1-\lambda)\underline{\underline{x}}_2)| \\ &\le \lambda|L_i(\underline{\underline{x}}_1)| + (1-\lambda)|L_i(\underline{\underline{x}}_2)| \\ &\le \lambda + 1 - \lambda \le 1 \qquad \text{for } i = 1, \dots, n \end{aligned}$$

Further $\mu(A) = \frac{1}{\det B}\mu(\widetilde{U}^n)$ where $\widetilde{U}^n = \{(t_1, \dots, t_n) \in \mathbb{R}^n | \ |t_i| \leq 1 \text{ for } i = 1, \dots, n\}$. Thus $\mu(A) \geq 2^n$. We now apply Minkowski's Convex Body Theorem to conclude that there is an integer point $\underline{\underline{x}}$ in A with $\underline{\underline{x}} \neq \underline{\underline{0}}$. This gives our result with $|L_i(\underline{\underline{x}})| \leq 1$ for $i = 1, \dots, n$. To get strict inequality for the first $n-1$ forms we need an additional argument.

For each $\epsilon > 0$ we define the set A_ϵ where A_ϵ consits of the $\underline{\underline{x}} \in \mathbb{R}^n$ for which

$$|L_i(\underline{\underline{x}})| < 1 \quad \text{for } i = 1, \dots, n-1$$

and

$$|L_n(\underline{\underline{x}})| < 1 + \epsilon.$$

Note that A_ϵ is bounded, symmetric and convex and $\mu(A_\epsilon) = (1+\epsilon)2^n > 2^n$. By Minkowski's Convex Body Theorem there is a non-zero integer point $\underline{\underline{x}}_\epsilon$ in A_ϵ. Note that $\cup_{0<\epsilon<1} A_\epsilon$ is a bounded set and so contains only finitely many integer points. Thus there is an integer point $\underline{\underline{x}} \in A_\epsilon$ with $\underline{\underline{x}} \neq \underline{\underline{0}}$ for $\epsilon = \frac{1}{m}$ for $m = 1, 2, \dots$ and so for this integer point $\underline{\underline{x}}$ we have

$$|L_i(\underline{\underline{x}})| < 1 \quad \text{for } i = 1, \dots, n-1$$

and

$$|L_n(\underline{\underline{x}})| \leq 1.$$

□

Theorem 7 implies Theorem 4a

Proof:

Put $l = m + n$ and consider the linear forms $\underline{\underline{x}} = (x_1, \dots, x_l)$ given by

$$L_i(\underline{\underline{x}}) = x_i \quad \text{for } i = 1, \dots, m$$

and

$$L_{m+j}(\underline{\underline{x}}) = \alpha_{j1}x_1 + \cdots + \alpha_{jm}x_m - x_{m+j} \quad \text{for } j = 1, \dots n.$$

Notice that the determinant of the system of equations given by the linear forms $L_1, \dots, L_l$ is $(-1)^n$.

Let Q be a real number with $Q > 1$. By Minkowski's Linear Forms Theorem there is a non-zero integer point $\underline{\underline{x}}$ such that

$$|L_i(\underline{\underline{x}})| < Q^{n/m} \quad \text{for } i = 1, \dots, m$$

and

$$|L_{m+j}(\underline{\underline{x}})| \leq \frac{1}{Q} \quad \text{for } j = 1, \dots, n.$$

We now put $q_i = x_i$ for $i = 1, \dots, m$ and $p_j = x_{m+j}$ for $j = 1, \dots, n$. Then

$$q = \max_{i=1,\dots,m} |q_i| < Q^{n/m}$$

and

$$|\alpha_{j1}q_1 + \cdots + \alpha_{jm}q_m - p_j| \leq \frac{1}{Q} \quad \text{for } j = 1, \dots, n.$$

Note that $q \neq 0$ since if it was then $q_1 = \cdots = q_m = 0$ and as a consequence $p_1 = \cdots = p_n = 0$ since $Q > 1$. This is a contradiction since $\underline{\underline{x}} = (q_1, \dots, q_m, p_1, \dots, p_n) \neq \underline{\underline{0}}$. □

Theorem 8. *Let $\alpha_{ij} \in \mathbb{R}$ for $i = 1, \dots, n$, $j = 1, \dots, m$. Put*

$$L_i(\underline{\underline{x}}) = \alpha_{i1}x_1 + \cdots + \alpha_{im}x_m \quad \text{for } i = 1, \dots, n.$$

Put $\mathcal{L}(\underline{\underline{x}}) = (L_1(\underline{\underline{x}}), \dots, L_n(\underline{\underline{x}}))$. Then there is an integer point $(\underline{\underline{x}}, \underline{\underline{y}}) = (x_1, \dots, x_m, y_1, \dots, y_n) \in \mathbb{R}^{m+n}$ with $\underline{\underline{x}} \neq \underline{\underline{0}}$ such that

$$\overline{|\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}|}^n < c_{m,n}\frac{1}{\overline{|\underline{\underline{x}}|}^m} \tag{1}$$

where $c_{m,n} = \frac{m^m n^n}{(m+n)^{m+n}}\frac{(m+n)!}{m!n!}$. Further if whenever $\underline{\underline{x}}$ is a non-zero integer point then $\mathcal{L}(\underline{\underline{x}})$ is not an integer point then there exist infinitely many integer points $(\underline{\underline{x}}, \underline{\underline{y}})$ with $\underline{\underline{x}} \neq \underline{\underline{0}}$ and with coprime components satisfying (1).

Remark:

1. Note that $c_{m,n} < 1$ since it is one of the $m + n + 1$ terms in the binomial expansion

$$1 = 1^{m+n} = \left(\frac{m}{m+n} + \frac{n}{m+n}\right)^{m+n}$$

2. Theorem 4 a) states that for any $Q > 1$, $Q \in \mathbb{R}$, there exists an integer point $(\underline{\underline{x}}, \underline{\underline{y}}) \in \mathbb{R}^{m+n}$ such that $1 \leq \overline{|\underline{\underline{x}}|} < Q^{n/m}$ and $\overline{|\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}|} \leq \frac{1}{Q}$. Thus $\overline{|\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}|}^n < \frac{1}{\overline{|\underline{\underline{x}}|}^m}$.

Lemma 1. *Let m and n be positive integers and let t be a positive real number. Let $K_{m,n}$ be the set of points $(\underline{\underline{x}}, \underline{\underline{y}}) = (x_1, \dots, x_m, y_1, \dots, y_n) \in \mathbb{R}^{m+n}$ satisfying*

$$t^{-n}\overline{|\underline{\underline{x}}|} + t^m\overline{|\underline{\underline{y}}|} \leq 1.$$

Thus $K_{m,n}$ is compact, symmetric about the origin, convex and has volume $2^{m+n}\frac{m!n!}{(m+n)!}$.

Proof:

Plainly $K_{m,n}$ is compact and symmetric about the origin. To see that $K_{m,n}$ is convex let $\lambda \in \mathbb{R}$ with $0 \leq \lambda \leq 1$ and suppose that $(\underline{\underline{x}}^{(1)}, \underline{\underline{y}}^{(1)})$ and $(\underline{\underline{x}}^{(2)}, \underline{\underline{y}}^{(2)})$ are in $K_{m,n}$ then
$\lambda(\underline{\underline{x}}^{(1)}, \underline{\underline{y}}^{(1)}) + (1-\lambda)(\underline{\underline{x}}^{(2)}, \underline{\underline{y}}^{(2)})$ is in $K_{m,n}$ since

$$\begin{aligned} t^{-n}\overline{|\lambda\underline{\underline{x}}^{(1)} + (1-\lambda)\underline{\underline{x}}^{(2)}|} + t^m\overline{|\lambda\underline{\underline{y}}^{(1)} + (1-\lambda)\underline{\underline{y}}^{(2)}|} & \\ \leq t^{-n}\left(\lambda\overline{|\underline{\underline{x}}^{(1)}|} + (1-\lambda)\overline{|\underline{\underline{x}}^{(2)}|}\right) + t^m\left(\lambda\overline{|\underline{\underline{y}}^{(1)}|} + (1-\lambda)\overline{|\underline{\underline{y}}^{(2)}|}\right) & \\ \leq \lambda\left(t^{-n}\overline{|\underline{\underline{x}}^{(1)}|} + t^m\overline{|\underline{\underline{y}}^{(1)}|}\right) + (1-\lambda)\left(t^{-n}\overline{|\underline{\underline{x}}^{(2)}|} + t^m\overline{|\underline{\underline{y}}^{(2)}|}\right) & \\ \leq \lambda + 1 - \lambda = 1 & \end{aligned}$$

We now calculate the volume of $K_{m,n}$. We first note that the linear transformation that sends x_i to $t^n x_i$ for $i = 1, \dots, m$ and y_j to $t^{-m}y_j$ for $j = 1, \dots, n$ has determinant $t^{nm} \cdot (t^{-m})^n = 1$. Thus the volume of $K_{m,n}(t)$ equals the volume of $K_{m,n}(1)$ for $t \in \mathbb{R}^+$. Thus we may suppose $t = 1$. Put $K_{m,n}(1) = K_{m,n}$. Further $vol(K_{m,n}) = 2^{m+n}vol(K^*_{m,n})$ where

$$K^*_{m,n} = \{(\underline{\underline{x}}, \underline{\underline{y}}) \in \mathbb{R}^{m+n} |\ \overline{|\underline{\underline{x}}|} + \overline{|\underline{\underline{y}}|} \leq 1,\ 0 \leq x_i,\ i = 1, \dots, m,\ 0 \leq y_i,\ j = 1, \dots, n\}.$$

Furthermore the volume of $K^*_{m,n}$ is m times the volume of $K^{**}_{m,n}$ where $K^{**}_{m,n} = \{(\underline{\underline{x}}, \underline{\underline{y}}) \in K^*_{m,n} | \ \underline{\underline{x}} = x_1\}$. Notice that if $|\overline{\underline{\underline{x}}}| = x_1$ then $0 \leq x_i \leq x_1$ for $i = 2, \ldots, m$ and $0 \leq y_j \leq 1 - x_1$ for $j = 1, \ldots, n$. Thus

$$vol(K^{**}_{m,n}) = \int_0^1 x_1^m (1 - x_1)^n dx_1 = A(m, n).$$

We claim that the integral is $\frac{(m-1)!n!}{(m+n)!}$. To verify the claim we first observe that by integration by parts

$$\begin{aligned} A(m,n) &= \int_0^1 x_1^{m-1}(1-x_1)^n \, dx_1 = \frac{x_1^m}{m}(1-x_1)^n]_0^1 + \int_0^1 \frac{x_1^m}{m} n(1-x_1)^{n-1} \, dx_1 \\ &= \frac{n}{m} A(m+1, n-1). \end{aligned}$$

We have

$$A(1,n) = \int_0^1 (1-x_1)^n \, dx_1 = \int_0^1 x_1^n \, dx_1 = \frac{1}{n+1}.$$

We now claim that $A(m,n) = \frac{(m-1)!n!}{(m+n)!}$ for m, n positive integers. Prove by induction on m. For $m = 1$ we have $A(1,n) = \frac{0!n!}{(n+1)!}$ as required. Suppose for m the result holds. Then

$$A(m+1,n) = \frac{m}{n+1} A(m, n+1) = \frac{m}{n+1} \frac{(m-1)!(n+1)!}{(m+n+1)!} = \frac{m!n!}{(m+1+n)!}$$

as required. □

Proof: (of Theorem 8)

Let $t \in \mathbb{R}^+$ and let $K_{m,n}(t)$ be as before. Put $C = \frac{2}{(V_{m,n})^{\frac{1}{m+n}}}$ where $V_{m,n}$ is the volume of $K_{m,n}(t)$. Let $T : \mathbb{R}^{m+n} \to \mathbb{R}^{m+n}$ be the linear transformation given by the map that sends $(\underline{\underline{x}}, \underline{\underline{y}})$ to $T(\underline{\underline{x}}, \underline{\underline{y}})$ where x_i is sent to Cx_i for $i = 1, \ldots, m$ and y_j is sent to $C(L_j(\underline{\underline{x}}) - y_j)$ for $j = 1, \ldots, n$.

Put $T(K_{m,n}(t)) = \widetilde{K}_{m,n}(t)$. Note that $\widetilde{K}_{m,n}(t)$ is compact, symmetric about $\underline{\underline{0}}$ and convex since these properties are preserved by linear transformations. The determinant of the matrix associated with T is $(-1)^n C^{m+n} = (-1)^n \frac{2^{m+n}}{V_{m,n}}$. Therefore the volume of $\widetilde{K}_{m,n}(t)$ is 2^{m+n}. Notice that

$$\widetilde{K}_{m,n}(t) = \{T(\underline{\underline{x}}, \underline{\underline{y}}) | \ t^{-n}|\overline{\underline{\underline{x}}}| + t^m |\overline{\underline{y}}| \leq 1\}.$$

T is invertible and so

$$\widetilde{K}_{m,n}(t) = \{(\underline{\underline{x}}, \underline{\underline{y}}) | \ t^{-n}|\overline{\underline{\underline{x}}}| + t^m |\overline{\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}}| \leq C\}.$$

By Minkowski's Convex Body Theorem there is an integer point $(\underline{\underline{x}}, \underline{\underline{y}}) \neq \underline{\underline{0}}$ in $\widetilde{K}_{m,n}(t)$. In particular,

$$t^{-n}|\overline{\underline{\underline{x}}}| + t^m |\overline{\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}}| \leq C. \tag{2}$$

Notice that for each integer point $(\underline{\underline{x}}, \underline{\underline{y}})$ there are only finitely many real numbers t for which

$$t^{-n}|\overline{\underline{\underline{x}}}| + t^m |\overline{\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}}| = C. \tag{3}$$

Thus these exist only countably many real numbers for which (3) has a solution with $(\underline{\underline{x}}, \underline{\underline{y}})$ an integer point. We shall suppose that t is not one of these reals and that

$$t^m > C \tag{4}$$

Then we may replace (2) by

$$t^{-n}|\underline{\underline{x}}| + t^m|\underline{\underline{\mathcal{L}(\underline{x})}} - \underline{\underline{y}}| < C \tag{5}$$

By the arithmetic-geometric mean inequality for real numbers $z_1, \dots, z_l$ with $z_i \geq 0$ for $i = 1, \dots, l$ we have $(z_1 \cdots z_l)^{1/l} \leq \frac{z_1+\cdots+z_l}{l}$. We take $l = m + n$ and $z_1 = \cdots = z_m = \frac{t^{-n}|\underline{\underline{x}}|}{m}$ and $z_{m+1} = \cdots = z_{m+n} = \frac{t^m|\underline{\underline{\mathcal{L}(\underline{x})}} - \underline{\underline{y}}|}{n}$

$$\left(\frac{t^{-n}|\underline{\underline{x}}|}{m}\right)^m \left(\frac{t^m|\underline{\underline{\mathcal{L}(\underline{x})}} - \underline{\underline{y}}|}{n}\right)^n \leq \frac{\left(t^{-n}|\underline{\underline{x}}| + t^m|\underline{\underline{\mathcal{L}(\underline{x})}} - \underline{\underline{y}}|\right)^{m+n}}{(m+n)^{m+n}}.$$

Thus by (5)

$$|\underline{\underline{\mathcal{L}(\underline{x})}} - \underline{\underline{y}}|^n < \frac{m^m n^n}{(m+n)^{m+n}} C^{m+n} \frac{1}{|\underline{\underline{x}}|^m} < C_{m,n} \frac{1}{|\underline{\underline{x}}|^m}.$$

Note that $\underline{\underline{x}} \neq \underline{\underline{0}}$ since if $\underline{\underline{x}} = \underline{\underline{0}}$ then by (4) and (5) we have $\underline{\underline{y}} = \underline{\underline{0}}$ also which is a contradiction since $(\underline{\underline{x}}, \underline{\underline{y}}) \neq (\underline{\underline{0}}, \underline{\underline{0}})$. This completes the proof of the first assertion.

To prove the second assertion note that if $(\underline{\underline{x}}, \underline{\underline{y}})$ satisfies (1) with $\underline{\underline{x}} \neq \underline{\underline{0}}$ and $\mathcal{L}(\underline{\underline{x}})$ is not an integer point for $\underline{\underline{x}}$ an integer point then $|\underline{\underline{\mathcal{L}(\underline{x})}} - \underline{\underline{y}}| > 0$. Thus, for t sufficiently large (5) does not hold. Accordingly, we may apply our argument again to get a new integer point $(\underline{\underline{x}}_1, \underline{\underline{y}}_1)$ with $\underline{\underline{x}}_1 \neq \underline{\underline{0}}$ for which $|\underline{\underline{\mathcal{L}(\underline{x}_1)}} - \underline{\underline{y}}_1|^n < C_{m,n} \frac{1}{|\underline{\underline{x}}_1|^m}$. Continuing in this way we produce infinitely many such integer points. □

If $n = m = 1$ then $C_{1,1} = \frac{1}{2}$. Thus if $\alpha \in \mathbb{R}$ with $\alpha \notin \mathbb{Q}$ then there exits infinitely many pairs of coprime integers p, q with $q \neq 0$ and

$$|q\alpha - p| < \frac{1}{2|q|}.$$

or equivalently

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{2q^2}. \tag{6}$$

Since in this case we cannot replace 2 by a number larger than $\sqrt{5}$ in (6) when α is a real number whose continued fraction has partial quotients which are eventually all 1 we might suspect that for other pairs (m, n), $C_{m,n}$ can't be replaced by an arbitrarily small number. In fact this is the case.

Definition:

Let α_{ij} be real numbers for $i = 1, \dots, n$ and $j = 1, \dots, m$ and put

$$L_i(\underline{\underline{x}}) = \alpha_{i1}x_1 + \cdots + \alpha_{im}x_m \quad \text{for } i = 1, \dots, n.$$

Put $\mathcal{L}(\underline{\underline{x}}) = (L_1(\underline{\underline{x}}), \dots, L_n(\underline{\underline{x}}))$. $L_1, \dots, L_n$ is said to be a badly approximable system of linear forms if there is a positive real number $\gamma = \gamma(L_1, \dots, L_n) = \gamma(\alpha_{11}, \dots, \alpha_{nm})$ such that for all integer points $(\underline{\underline{x}}, \underline{\underline{y}})$ with $\underline{\underline{x}} \neq \underline{\underline{0}}$ we have

$$|\underline{\underline{\mathcal{L}(\underline{x})}} - \underline{\underline{y}}|^n > \gamma \frac{1}{|\underline{\underline{x}}|^m}.$$

Lemma 2. *For every positive integer l there exists a real algebraic number θ of degree l over $\mathbb{Q}$ for which all of the conjugates $\theta = \theta_1, \ldots, \theta_l$ of θ over $\mathbb{Q}$ are real numbers.*

Proof:

Let $l \in \mathbb{Z}^+$. Put $f_l(x) = (x-4)(x-8)\cdots(x-4l) - 2$. Note that f_l is irreducible over $\mathbb{Q}$ by Eisenstein's theorem with $p = 2$. It remains to show that f_l has distinct real roots since we then take θ to be a root of f_l. Notice that for $l \geq 2$,

$$\begin{aligned} f_l(4l+2) &= (2)(6)\cdots(4l-2) - 2 > 0 \\ f_l(4l-2) &= (-2)(2)(6)\cdots(4(l-1)-2) - 2 < 0 \\ &\vdots \\ f_l(2) &= (-2)(-6)\cdots(2-4l) - 2 \begin{cases} < 0 & \text{if } l \text{ is odd} \\ > 0 & \text{if } l \text{ is even} \end{cases} \end{aligned}$$

Note between 2 and 6, 6 and 10, ..., $4l-2$ and $4l+2$ f changes sign. Therefore f_l has l distinct real zeros as required. □

Theorem 9. *Let $1, \alpha_1, \ldots, \alpha_m$ be a basis for a real algebraic number field of degree $m+1$ over $\mathbb{Q}$. Then the linear form $L(\underline{\underline{x}}) = \alpha_1 x_1 + \cdots + \alpha_m x_m$ is badly approximable.*

Proof:

Let $q_1, \ldots, q_m$ and p be integers with $q_1, \ldots, q_m$ not all zero and for which

$$|\alpha_1 q_1 + \cdots + \alpha_m q_m - p| < 1. \tag{7}$$

Let $c_1, c_2, \ldots$ denote positive numbers which can be determined in terms of $\alpha_1, \ldots, \alpha_m$.
Put $q = \max_{i=1,\ldots,m} |q_i|$. Then, by (7), $|p| < c_1|q|$.

Let $\alpha_j^{(i)}$ for $i = 1, \ldots, m+1$ denote the conjugates of over $\mathbb{Q}$ of α_j for $j = 1, \ldots, m$. Then $\alpha_1^{(i)} q_1 + \cdots + \alpha_m^{(i)} q_m - p$ is a conjugate of $\alpha_1 q_1 + \cdots + \alpha_m q_m - p = \alpha_1^{(1)} q_1 + \cdots + \alpha_m^{(1)} q_m - p$ for $i = 1, \ldots, m+1$. Further, for $i = 1, \ldots, m+1$,

$$|\alpha_1^{(i)} q_1 + \cdots + \alpha_m^{(i)} q_m - p| < c_2 q.$$

Observe that the norm of $(\alpha_1 q_1 + \cdots + \alpha_m q_m - p)$ is a rational number which is non-zero since since $\alpha_1, \ldots, \alpha_m$ is a basis and since $(q_1, \ldots, q_m) \neq \underline{\underline{0}}$. Thus $|N(\alpha_1 q_1 + \cdots + \alpha_m q_m - p)| > 0$.

On the other hand

$$|N(\alpha_1 q_1 + \cdots + \alpha_m q_m - p)| \leq |\alpha_1 q_1 + \cdots + \alpha_m q_m - p| \cdot (c_2 q)^m. \tag{8}$$

Further there exists a positive integer h such that $h\alpha_1, \ldots, h\alpha_m$ are all algebraic integers. Then $|N(h\alpha_1 q_1 + \cdots + h\alpha_m q_m - hp)| \geq 1$ so

$$|N(\alpha_1 q_1 + \cdots + \alpha_m q_m - p)| \geq \frac{1}{h^{m+1}}| \tag{9}$$

By (8) and (9) we see that

$$|\alpha_1 q_1 + \cdots + \alpha_m q_m - p| \geq \frac{1}{h^{m+1} c_2^m} \frac{1}{q^m}.$$

□

Theorem 10. *(Perron 1921) For each pair of positive integers (m,n) there exist algebraic numbers α_{ij}, $1 \le i \le n$ and $1 \le j \le m$ for which the associated system of linear forms $L_i(\underline{\underline{x}}) = \alpha_{i1}x_1 + \cdots + \alpha_{im}x_m$, for $i = 1, \dots, n$ is badly approximable.*

Proof:

Put $l = m + n$ and let θ_1 be a real algebraic integer of degree l with the property that all of the conjugates $\theta_1, \dots, \theta_l$ say of θ_1 are real numbers. As usual, we put $\underline{\underline{x}} = (x_1, \dots, x_m)$ and $\underline{\underline{y}} = (y_1, \dots, y_n)$. Next we put

$$M_k(\underline{\underline{x}}, \underline{\underline{y}}) = \sum_{i=1}^{n} \theta_k^{i-1} y_i + \sum_{j=1}^{m} \theta_k^{n+j-1} x_j, \quad \text{for } k = 1, \dots, l.$$

Observe that if $(\underline{\underline{x}}, \underline{\underline{y}})$ is an integer point with $(\underline{\underline{x}}, \underline{\underline{y}}) \neq (\underline{\underline{0}}, \underline{\underline{0}})$ then $M_k(\underline{\underline{x}}, \underline{\underline{y}}) \neq 0$ since otherwise θ_k would be the root of a polynomial with integer coefficients of degree less than l which is a contradiction. Further, by construction $M_1(\underline{\underline{x}}, \underline{\underline{y}}), \dots, M_l(\underline{\underline{x}}, \underline{\underline{y}})$ are conjugate algebraic integers. Therefore since the norm of a non-zero algebraic integer is a non-zero integer,

$$|M_1(\underline{\underline{x}}, \underline{\underline{y}})| \cdots |M_l(\underline{\underline{x}}, \underline{\underline{y}})| \ge 1 \tag{10}$$

for all integer points $(\underline{\underline{x}}, \underline{\underline{y}}) \neq (\underline{\underline{0}}, \underline{\underline{0}})$. We now define the linear forms $L_1, \dots, L_n$ by the rule

$$M_k(\underline{\underline{x}}, \underline{\underline{y}}) = \sum_{i=1}^{n} \theta_k^{i-1}(y_i - L_i(\underline{\underline{x}})), \quad \text{for } k = 1, \dots, n.$$

In other words we require that $-\sum_{i=1}^{n} \theta_k^{i-1} L_i(\underline{\underline{x}}) = \sum_{j=1}^{m} \theta_k^{n+j-1} x_j$, for $k = 1, \dots, n$. Hence

$$-\begin{pmatrix} 1 & \theta_1 & \theta_1^2 & \cdots & \theta_1^{n-1} \\ 1 & \theta_2 & \theta_2^2 & \cdots & \theta_2^{n-1} \\ & \vdots & & & \vdots \\ 1 & \theta_n & \theta_n^2 & \cdots & \theta_n^{n-1} \end{pmatrix} \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1m} \\ \vdots & & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} \theta_1^n & \cdots & \theta_1^{n+m-1} \\ \vdots & & \vdots \\ \theta_n^n & \cdots & \theta_n^{n+m-1} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

We can solve this system for the α_{ij}'s since the matrix $\begin{pmatrix} 1 & \theta_1 & \cdots & \theta_1^{n-1} \\ & \vdots & & \vdots \\ 1 & \theta_n & \cdots & \theta_n^{n-1} \end{pmatrix}$ is a vanderMonde matrix with determinant $(-1)^{\frac{n(n-1)}{2}} \prod_{i<j}(\theta_i - \theta_j)$. Since $\theta_i \neq \theta_j$ for $i \neq j$ the determinant is non-zero and so the matrix is invertible. Further each α_{ij} is an algebraic number since it is a rational function of the θ_t's. Next observe that for $k = n+1, \dots, n+m$

$$M_k(\underline{\underline{x}}, \underline{\underline{y}}) = \sum_{i=1}^{n} \theta_k^{i-1}(y_i - L_i(\underline{\underline{x}})) + \sum_{j=1}^{m} \lambda_{kj} x_j,$$

where λ_{kj} is an algebraic number determined by the θ_i's.

Let $c_1, c_2, \dots,$ denote positive numbers which can be determined in terms of $\theta_1, \dots, \theta_l$, n and m. Let $(\underline{\underline{x}}, \underline{\underline{y}}) \neq (\underline{\underline{0}}, \underline{\underline{0}})$ be an integer point and suppose that $\overline{|\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}|} < 1$. Then

$$|M_k(\underline{\underline{x}}, \underline{\underline{y}})| \le c_1 \overline{|\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}|}, \quad \text{for } k = 1, \dots, n,$$

and

$$|M_k(\underline{\underline{x}}, \underline{\underline{y}})| \le c_2 \overline{|\underline{\underline{x}}|} \quad \text{for } k = n+1, \dots, n+m.$$

Thus by (10), $1 \le c_1^n \overline{|\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}|}^n \cdot c_2^m \overline{|\underline{\underline{x}}|}^m$, for integer points $(\underline{\underline{x}}, \underline{\underline{y}}) \neq (\underline{\underline{0}}, \underline{\underline{0}})$. Since $\overline{|\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}|} < 1$ we have $\underline{\underline{x}} \neq \underline{\underline{0}}$ and thus $\overline{|\mathcal{L}(\underline{\underline{x}}) - \underline{\underline{y}}|}^n > c_3 \frac{1}{\overline{|\underline{\underline{x}}|}^m}$, as required. □

Observe that $C_{m,n}$ is Theorem 8 cannot be replaced by an arbitrarily small real number for any pair of positive integers (m,n). Recall $C_{1,1} = \frac{1}{2}$ and the best possible constant is $\frac{1}{\sqrt{5}}$. If $m = 1$ we have $C_{1,n} = \left(\frac{n}{n+1}\right)^n$. In 1914, Blichfeldt improved this to

$$\left(\frac{n}{n+1}\right)^n \left(1 + \left(\frac{n-1}{n+1}\right)^{n+3}\right)^{-1}.$$

Thus given real numbers $\alpha_1, \ldots, \alpha_n$ with at least one of them irrational there exist, by Theorem 8, infinitely many integer points $(x, y_1, \ldots, y_n)$ with $x > 0$ for which

$$\max(|\alpha_1 x - y_1|, \ldots, |\alpha_n x - y_n|) < \left(\frac{n}{n+1}\right) \frac{1}{x^{1/n}}.$$

By Blichfeldt $\frac{n}{n+1} \to \left(\frac{n}{n+1}\right)\left(1 + \left(\frac{n-1}{n+1}\right)^{n+3}\right)^{-1/n}$. Thus if $n = 2$, Theorem 8 gives $\frac{2}{3}$ and Blichfeldt gives $.66323\ldots$. In fact, for $n = 2$ the best possible constant is between $\sqrt{2/7} = .534\ldots$ and $.615\ldots$.

A badly approximable system of linear forms $L_i(x) = \alpha_i x$ for $i = 1, \ldots, n$ satisfies, for x a non-zero integer,

$$|x| \left[\max(\|\alpha_1 x\|, \ldots, \|\alpha_n x\|)\right]^n > \gamma$$

for some positive real number γ.

Littlewood conjectured that if $\alpha_1, \ldots, \alpha_n$ are real numbers with $n \geq 2$ then

$$\liminf_{x \to \infty} |x| \cdot \|\alpha_1 x\|, \ldots, \|\alpha_n x\| = 0,$$

where the liminf is taken over positive integers x. The conjecture is still open.

In 1926, Khintchine proved that the set of badly approximable n-tuples $(\alpha_1, \ldots, \alpha_n)$ in $\mathbb{R}^n$ is of Lebesgue measure 0.

Let α_{ij} with $1 \leq i \leq n$ and $1 \leq j \leq m$, be real numbers and put

$$L_i(\underline{\underline{x}}) = \alpha_{i1} x_1 + \cdots + \alpha_{im} x_m \quad \text{for } i = 1, \ldots, n.$$

Associated to the system $L_i(\underline{\underline{x}}), \ldots, L_n(\underline{\underline{x}})$ there is a dual system of linear forms $M_j(\underline{\underline{u}})$ with

$$M_j(\underline{\underline{u}}) = \alpha_{1j} u_1 + \cdots + \alpha_{nj} u_n, \quad \text{for } j = 1, \ldots, m.$$

By a Transference Theorem, Khintchine proved that if $L_1, \ldots, L_n$ is a badly approximable system of linear forms then so is $M_1, \ldots, M_m$.

Let $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ be linearly independent vectors in $\mathbb{R}^n$. Consider the set of points

$$\Lambda = \{g_1 \underline{\underline{a}}_1 + \cdots + g_n \underline{\underline{a}}_n \mid g_i \in \mathbb{Z}, i = 1, \ldots, n\}.$$

This set is known as a lattice in $\mathbb{R}^n$ with basis $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$. Note that if $\underline{\underline{x}} \in \Lambda$ then, since $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ are linearly independent, there is a unique representation for $\underline{\underline{x}}$ of the form $\underline{\underline{x}} = g_1 \underline{\underline{a}}_1 + \cdots + g_n \underline{\underline{a}}_n$ with $g_i \in \mathbb{Z}$ for $i = 1, \ldots, n$. The basis $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ for Λ is not uniquely determined in general since if we put $\underline{\underline{a}}'_i = \sum_{j=1}^n b_{ij} \underline{\underline{a}}_j$, for $i = 1, \ldots, n$ where the b_{ij}'s are integers and $\det(b_{ij}) = \pm 1$ then $\underline{\underline{a}}'_1, \ldots, \underline{\underline{a}}'_n$ is also a basis for Λ. To see this note that $(b_{ij})^{-1} = (c_{ij})$ with $c_{ij} \in \mathbb{Z}$ since $\det(b_{ij}) = \pm 1$. We then have $\underline{\underline{a}}_i = \sum_{j=1}^n c_{ij} \underline{\underline{a}}'_j$ for $i = 1, \ldots, n$.

Suppose that $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ is a basis for a lattice Λ in $\mathbb{R}^n$. Suppose that $\underline{\underline{y}}_1, \ldots, \underline{\underline{y}}_n$ is another basis for Λ. Thus

$$\underline{\underline{a}}_i = \sum_{j=1}^{n} d_{ij} \underline{\underline{y}}_j \quad \text{with } d_{ij} \in \mathbb{Z},$$

and

$$\underline{\underline{y}}_i = \sum_{j=1}^{n} e_{ij} \underline{\underline{a}}_j \quad \text{with } e_{ij} \in \mathbb{Z},$$

Thus we have

$$(d_{ij})(e_{ij}) = I_n$$

so $\det(d_{ij}) \det(e_{ij}) = 1$. Since the d_{ij}'s and $e'_{ij}s$ are integers we see that

$$\det(d_{ij}) = \det(e_{ij}) = \pm 1.$$

Therefore if $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ and $\underline{\underline{a}}'_1, \ldots, \underline{\underline{a}}'_n$ are two bases for Λ then

$$\det(\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n) = \det(c_{ij}) \det(\underline{\underline{a}}'_1, \ldots, \underline{\underline{a}}'_n)$$

where (c_{ij}) is obtained by expressing $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ in terms of $\underline{\underline{a}}'_1, \ldots, \underline{\underline{a}}'_n$. We then define $d(\Lambda)$ by

$$d(\Lambda) = |\det(\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n)|$$

where $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ is any basis for Λ.

Note: that $d(\Lambda) > 0$ since $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ are linearly independent.

Minkowski's Convex Body Theorem, II

Let Λ be a lattice in $\mathbb{R}^n$. Let A be a convex set in $\mathbb{R}^n$ which is symmetric about the origin with positive measure $\mu(A)$. If $\mu(A) > 2^n d(\Lambda)$ or A is compact and $\mu(A) \geq 2^n d(\Lambda)$ then A contains a point of Λ different from $\underline{\underline{0}}$.

Proof:

Suppose that $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ is a basis for Λ. Then we have

$$\underline{\underline{a}}_i = (\alpha_{i1}, \ldots, \alpha_{in}) \quad \text{for } i = 1, \ldots, n$$

Let T be the linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$ associated with the matrix (α_{ij}). Then $\Lambda = T\Lambda_0$ where Λ_0 is the lattice of integer points in $\mathbb{R}^n$. Notice that

$$\mu(T^{-1}A) = d(\Lambda)^{-1} \mu(A).$$

Further, since T is a linear transformation and A is convex and symmetric about $\underline{\underline{0}}$ then so is $T^{-1}A$. We now apply Minkowski's Theorem I, to conclude the proof. $\square$

Theorem 11. *A subset Λ of $\mathbb{R}^n$ is a lattice if and only if*
i) $\underline{\underline{a}} + \underline{\underline{b}} \in \Lambda$ and $-\underline{\underline{a}} \in \Lambda$ for all $\underline{\underline{a}}, \underline{\underline{b}} \in \Lambda$
ii) Λ contains n linearly independent points.
iii) Λ is discrete.

Proof:

$\Rightarrow$ Immediate from definition of a lattice.

$\Leftarrow$ We'll prove this by induction on n.
First, consider the case $n = 1$. By *ii*) there is a non-zero point a in Λ. By *i*) we may suppose that $a > 0$. Further, we may suppose that a is the smallest positive real number in Λ which we know exists by *iii*). Then by *i*) $\{ga \mid g \in \mathbb{Z}\}$ is contained in Λ. Further there are no other points in Λ since otherwise we could find a smaller positive real in Λ.

Assume the result holds for $n-1$ with $n \geq 2$. We may choose our coordinate system so that Λ has $n-1$ linearly independent points on the set $S = \{(x_1, \ldots, x_{n-1}, 0) \mid (x_1, \ldots, x_{n-1}) \in \mathbb{R}^{n-1}\}$. Let Λ' be the intersection of Λ with the set S. By the inductive hypothesis Λ' is a lattice in S. Let $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_{n-1}$ be a basis for Λ'. Then by *ii*) there is a point $\underline{\underline{b}}_n$ in Λ which is linearly independent of $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_{n-1}$. We may choose $\underline{\underline{b}}_n$ so that the n-th coordinate of $\underline{\underline{b}}_n$ is positive. Further we may choose $\underline{\underline{b}}_n = (b_{n1}, \ldots, b_{nn})$ so that b_{nn} is minimal since otherwise we obtain an infinite sequence of distinct points in a compact subset of $\mathbb{R}^n$ from which we can extract a convergent subsequence which contradicts *iii*).

We claim that $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_n$ is a basis for Λ. For $\underline{\underline{d}} \in \Lambda$ say $\underline{\underline{d}} = (d_1, \ldots, d_n)$. Notice that d_n is an integral multiple of b_{nn} and then $\underline{\underline{d}} - \left(\frac{d_n}{b_{nn}}\right) \underline{\underline{b}}_n \in S$ and so, by induction, $\underline{\underline{d}} - \left(\frac{d_n}{b_{nn}}\right) \underline{\underline{b}}_n$ is an integer linear combination of $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_{n-1}$ as required. $\square$

Let $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ be points in a lattice Λ in $\mathbb{R}^n$ with a basis $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_n$. Then

$$(*) \qquad \underline{\underline{a}}_i = \sum_{j=1}^{n} t_{ij} \underline{\underline{b}}_j \quad \text{with } t_{ij} \in \mathbb{Z}.$$

The integer I give by

$$I = |\det(t_{ij})| = \frac{|\det(\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n)|}{|\det(\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_n)|} = \frac{|\det(\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n)|}{d(\Lambda)}$$

is called the index of $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ in Λ. $I = 0$ if and only if $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ are linearly dependent. If $\underline{\underline{a}}_1, \ldots, \underline{\underline{a}}_n$ are linearly independent they generate a lattice Λ' which is a sublattice of Λ. Then

$$I = \frac{d(\Lambda')}{d(\Lambda)}.$$

Recall $(*)$. Suppose $D = |\det(t_{ij})| \neq 0$. Then $(t_{ij})^{-1} = (w_{ij})$ with w_{ij} rational numbers for which $Dw_{ij} \in \mathbb{Z}$ for i, j.

Theorem 12. *Let Λ be a sublattice of a lattice M in $\mathbb{R}^n$. Let $\underline{\underline{b}}_1, \dots, \underline{\underline{b}}_n$ be a basis of M. Then there exists a basis $\underline{\underline{a}}_1, \dots, \underline{\underline{a}}_n$ of Λ such that*

$$\begin{aligned}\underline{\underline{a}}_1 &= t_{11}\underline{\underline{b}}_1\\ \underline{\underline{a}}_2 &= t_{21}\underline{\underline{b}}_1 + t_{22}\underline{\underline{b}}_2\\ &\vdots\\ \underline{\underline{a}}_n &= t_{n1}\underline{\underline{b}}_1 + \cdots + t_{nn}\underline{\underline{b}}_n\end{aligned}$$

where the t_{ij}'s are integers with $t_{ii} > 0$ for $i = 1, \dots, n$ and with $t_{ij} < t_{ii}$ for $j = 1, \dots, i-1$. (Alternatively, we can have an upper triangular arrangement i.e. $t_{ji} < t_{ii}$).

Proof:

Let $D = |\det(t_{ij})|$. Then $D\underline{\underline{b}}_i$ is an integral linear combination of $\underline{\underline{a}}_1, \dots, \underline{\underline{a}}_n$ for $i = 1, \dots, n$. In particular, $D\underline{\underline{b}}_i \in \Lambda$ for $i = 1, \dots, n$. Then for each i with $1 \le i \le n$ there exist $\underline{\underline{x}}_i$ in Λ of the form

$$\underline{\underline{x}}_i = v_{i1}\underline{\underline{b}}_1 + \cdots + v_{ii}\underline{\underline{b}}_i$$

with $v_{ij} \in \mathbb{Z}$ and $v_{ii} \neq 0$. We now choose $\underline{\underline{x}}_i$ so that $|v_{ii}|$ is non-zero and minimal. We claim that $\underline{\underline{x}}_1, \dots, \underline{\underline{x}}_n$ is a basis for Λ.

To see this suppose that $\underline{\underline{c}} \in \Lambda$ and $\underline{\underline{c}}$ is not a linear combination of $\underline{\underline{x}}_1, \dots, \underline{\underline{x}}_n$. Then $\underline{\underline{c}} \neq 0$ and $\underline{\underline{c}} \in M$ and so there exists integers $l_1, \dots, l_n$ so that $\underline{\underline{c}} = l_1\underline{\underline{b}}_1 + \cdots + l_n\underline{\underline{b}}_n$, and in fact we may write $\underline{\underline{c}} = l_1\underline{\underline{b}}_1 + \cdots + l_k\underline{\underline{b}}_k$ with $k \le n$ and $l_k \neq 0$.

Suppose that $\underline{\underline{c}}$ is chosen with k minimal. Since $v_{kk} \neq 0$ we can find an integer s such that $|l - sv_{kk}| < |v_{kk}|$. Then

$$\underline{\underline{c}} - s\underline{\underline{x}}_k = (l_1 - sv_{k1})\underline{\underline{b}}_1 + \cdots + (l - sv_{kk})\underline{\underline{b}}_k.$$

By the minimality of $|v_{kk}|$ we see that $l - sv_{kk} = 0$ and this contradicts the minimality of k for $\underline{\underline{c}}$. Thus $\underline{\underline{x}}_1, \dots, \underline{\underline{x}}_n$ is a basis for Λ.

We now observe that by replacing $\underline{\underline{x}}_k$ by $-\underline{\underline{x}}_k$ if necessary we may suppose that $v_{kk} > 0$.

To complete our proof we put

$$\underline{\underline{a}}_i = h_{i1}\underline{\underline{x}}_1 + \cdots + h_{ii-1}\underline{\underline{x}}_{i-1} + \underline{\underline{x}}_i$$

where the h_{ij} are integers to be determined. Then

$$\underline{\underline{a}}_i = t_{i1}\underline{\underline{b}}_1 + \cdots + t_{ii}\underline{\underline{b}}_i$$

Since

$$\begin{aligned}\underline{\underline{a}}_i =& h_{i1}(v_{11}\underline{\underline{b}}_1) + h_{i2}(v_{21}\underline{\underline{b}}_1 + v_{22}\underline{\underline{b}}_2) + \cdots + h_{ij}(v_{j1}\underline{\underline{b}}_1 + \cdots + v_{jj}\underline{\underline{b}}_j) + \cdots +\\ & h_{ii-1}(v_{i-11}\underline{\underline{b}}_1 + \cdots + v_{i-1i-1}\underline{\underline{b}}_{i-1}) + 1 \cdot (v_{i1}\underline{\underline{b}}_1 + \cdots + v_{ii}\underline{\underline{b}}_i)\end{aligned}$$

Thus $t_{ii} = v_{ii} > 0$. Further

$$t_{ij} = h_{ij}v_{ji} + h_{ij+1}v_{j+1j} + \cdots + h_{ii-1}v_{i-1j} + v_{ij}.$$

We can now choose in order $h_{ii-1}, h_{ii-2}, \dots, h_{i1}$ so that $0 \le t_{ij} < t_{ii}$ for $j < i$. □

Recall the following elementary unimodular column operations:

i) exchanging two columns.
ii) multiplying a column by -1.
iii) adding an integer multiple of one column to another column.

Notice that if a matrix represents a sublattice within a lattice with respect to a given basis then the unimodular column operations on the matrix give a new matrix which represents the same sublattice. Thus we can recast out theorem in this setting. It suffices then to show that a non-singular matrix with integer entries can be put in Hermite normal form with integer entries by a sequence of unimodular column operations.

Proof B of Theorem 12:

Suppose we start with a matrix A. By unimodular column operations we may suppose that the first row $(a_{11}, \dots, a_{1n})$ of A is such that $a_{11} \geq a_{12} \geq \cdots \geq a_{1n} \geq 0$, $a_{11} > 0$, and $a_{11} + \cdots + a_{1n}$ minimal.

Note that $a_{12} = 0$ since if $a_{12} > 0$ then we could get a smaller sum $a_{11} + \cdots + a_{1n}$ by subtracting the second column from the first column which is a contradiction. Similarly $a_{13}, \dots, a_{1n}$ are zero. We repeat the argument with the last $n-1$ coordinates of the second row to get $a_{23} = \cdots = a_{2n} = 0$. Continuing in this way we obtain a lower-triangular matrix with integer entries and positive integer entries along the main diagonal.

For $i = 2, \dots, n$ and $j = 1, \dots, i-1$ we add an integer multiple of the i-th column of A to the j-th column of A so that the ij-th entry of A is non-negative and less than a_{ii}. Then A is in integer Hermite normal form as required. □

Theorem 13. *Let A and A' be $n \times n$ non-singular matrices with integer entries and (row) Hermite normal forms B and B' respectively. Let $\underline{\underline{b}}_1, \dots, \underline{\underline{b}}_n$ be a basis for $\mathbb{R}^n$. Then A generates the same lattice as A' with respect to $\underline{\underline{b}}_1, \dots, \underline{\underline{b}}_n$ if and only if $B = B'$.*

Proof:

$\Leftarrow$ We have $A = UB$ and $A' = U'B'$ where U and U' are unimodular matrices, so matrices with integer entries and determinant ± 1. Hence they generate the same lattice.

$\Rightarrow$ Let $B = (B_{ij})$ and $B' = (B'_{ij})$ and suppose that $B \neq B'$. Let ij be the entry for which $B_{ij} \neq B'_{ij}$ with j minimal. Without loss of generality we may assume $B_{jj} \geq B'_{jj}$. Let $\underline{\underline{r}}_i = B_{i1}\underline{\underline{b}}_1 + \cdots + B_{i1}\underline{\underline{b}}_n$ and $\underline{\underline{r}}'_i = B'_{i1}\underline{\underline{b}}_1 + \cdots + B'_{in}\underline{\underline{b}}_n$. Then $\underline{\underline{r}}_i \in \Lambda$ and $\underline{\underline{r}}'_i \in \Lambda$ so $\underline{\underline{r}}_i - \underline{\underline{r}}'_i \in \Lambda$. Thus $\underline{\underline{r}}_i - \underline{\underline{r}}'_i$ is an integer linear combination of $\underline{\underline{b}}_1, \dots, \underline{\underline{b}}_n$. By our choice of ij we have

$$\underline{\underline{r}}_i - \underline{\underline{r}}'_i = (B_{ij} - B'_{ij})\underline{\underline{b}}_j + \cdots + (B_{in} - B'_{in})\underline{\underline{b}}_n$$

Note that $\underline{\underline{r}}_i - \underline{\underline{r}}'_i$ is the span of $\underline{\underline{b}}_j, \dots, \underline{\underline{b}}_n$ and so $(B_{ij} - B'_{ij})$ is an integer multiple of B_{jj}. But $0 < |B_{ij} - B'_{ij}| < B_{jj}$ since if $i = j$ then $0 < B_{jj} - B'_{jj} < B_{jj}$, while if $i < j$ then $0 \leq B_{ij} < B_{jj}$ and $0 \leq B'_{ij} < B'_{jj} \leq B_{jj}$ which is a contradiction. □

Remark: Since the row Hermite normal form of the matrix associated with a sublattice of a lattice is uniquely determined the representation given in Theorem 12 is uniquely determined.

Let Λ be a sublattice of a lattice M. We split the elements of M into equivalence classes under the equivalence relation $\sim$. We say that $c \sim d$ if and only if $c - d \in \Lambda$.

Lemma 1. *Let Λ be a sublattice of M. The index of Λ in M is the number of equivalence classes under $\sim$.*

Proof:

Let $\underline{\underline{a}}_1, \dots, \underline{\underline{a}}_n$ and $\underline{\underline{b}}_1, \dots, \underline{\underline{b}}_n$ be bases for Λ and M respectively with $\underline{\underline{a}}_1, \dots, \underline{\underline{a}}_n$ chosen in the form described in Theorem 12. Plainly the index of Λ in M is $\prod_{i=1}^{n} t_{ii}$. It suffices then to show that every element in M is equivalent to precisely one term of the form

$$q_1\underline{\underline{b}}_1 + \cdots + q_n\underline{\underline{b}}_n \quad \text{where } 0 \le q_i < t_{ii} \text{ for } i = 1, \dots, n. \tag{11}$$

Let $\underline{\underline{c}} = c_1\underline{\underline{b}}_1 + \cdots + c_n\underline{\underline{b}}_n$ be in M. We first show that $\underline{\underline{c}}$ is equivalent to an element of M of the form (11). To see this note that $\underline{\underline{c}}$ is equivalent to $\underline{\underline{c}} - q\underline{\underline{a}}_n$ for any $q \in \mathbb{Z}$. Thus we may subtract a multiple of $\underline{\underline{a}}_n$ to ensure that c_n is replaced by an integer q_n with $0 \le q_n < t_{nn}$. Next we subtract a multiple of $\underline{\underline{a}}_{n-1}$ and so replace c_{n-1} by q_{n-1} with $0 \le q_{n-1} < t_{n-1n-1}$. Continuing in this way we obtain an element of M equivalent to $\underline{\underline{c}}$ and of the form (11).

Finally we show that any two lattice elements of M of the form (11) are distinct. If the difference of two such lattice elements is in Λ then we have $\underline{\underline{r}} = r_1\underline{\underline{b}}_1 + \cdots + r_n\underline{\underline{b}}_n \in \Lambda$ with $|r_i| < t_{ii}$ for $i = 1, \dots, n$, and not all of the r_i's zero. Suppose that j is the largest integer for which $r_j \neq 0$. Then in that case $\underline{\underline{r}}$ is an integer linear combination of $\underline{\underline{a}}_1, \cdots, \underline{\underline{a}}_j$ say

$$\underline{\underline{r}} = d_1\ \underline{\underline{a}}_1 + \cdots + dj\ \underline{\underline{a}}_j.$$

But then by Theorem 12

$$\underline{\underline{r}} = m_1\ \underline{\underline{b}}_1 + \cdots + m_j\ \underline{\underline{b}}_j$$

with $m_1, \cdots, m_j$ integers and $m_j = d_j\ t_{jj}$. On the other hand $m_j = r_j$ and $0 < |r_j| < t_{jj}$ which is a contradiction.

Thus all the elements of the form (11) are in distinct equivalence classes as required. □

Lemma 2. *Let n, m and $k_1, \dots, k_m$ be positive integers. Let a_{ij} for $1 \le i \le m, 1 \le j \le n$ be integers. The set Λ of integers points $\underline{\underline{u}}$ in $\mathbb{R}^n$ satisfying*

$$\sum_{j=1}^{n} a_{ij}u_j \equiv 0 \pmod{k_i}, \quad i = 1, \dots, m$$

is a lattice with $d(\Lambda) \le k_1 \cdots k_m$.

Proof: Λ is discrete since it is a subset of Λ_0. Next since

$$(k_1 \cdots k_m, 0, \dots, 0), (0, k_1 \cdots k_m, 0, \dots, 0), \dots, (0, \dots, 0, k_1 \cdots k_m) \in \Lambda$$

we see that Λ contains n linearly independent vectors. Finally if $\underline{\underline{u}} = (u_1, \dots, u_n)$ and $\underline{\underline{v}} = (v_1, \dots, v_n)$ are in Λ then $\underline{\underline{u}} \pm \underline{\underline{v}} \in \Lambda$ since

$$\sum_{j=1}^{n} a_{ij}(u_j \pm v_j) \equiv 0 \pmod{k_i} \quad i = 1, \dots, m$$

Therefore, by Theorem 11, Λ is a lattice and so is a sublattice of Λ_0.

We have $d(\Lambda) = \frac{d(\Lambda)}{d(\Lambda_0)}$ is the index of Λ in Λ_0. By Lemma 1 this is the number of equivalence classes of points in Λ_0 under $\sim$. But $\underline{\underline{u}} \sim \underline{\underline{v}}$ if and only if

$$\sum_{j=1}^{n} a_{ij}(u_j - v_j) \equiv 0 \pmod{k_i} \quad i = 1, \dots, m.$$

Therefore $d(\Lambda) \leq k_1 \cdots k_m$. □

Theorem 14. *(Lagrange's Theorem)*
Every positive integer is the sum of four squares.

Proof:

We may assume, without loss of generality, that $m > 1$ and that m is square free. So suppose $m = p_1 \cdots p_r$ with $p_1, \ldots, p_r$ distinct primes. We first observe that for each prime p there exist integers a_p and b_p for which

$$a_p^2 + b_p^2 + 1 \equiv 0 \pmod{p}.$$

Note that if $p = 2$ then $a_p = 1$ and $b_p = 0$. Suppose p is odd. Then consider a^2 with $0 \leq a < \frac{p}{2}$ and $-1 - b^2$ with $0 \leq b < \frac{p}{2}$. We have $\left[\frac{p}{2}\right] + 1$ terms in each grouping and so two must be the same modulo p. In particular there exists a_p and b_p with

$$a_p^2 = -1 - b_p^2 \pmod{p}$$

as required.

We now consider the set Λ of integer points (u_1, u_2, u_3, u_4) which satisfy the congruences

$$u_1 \equiv a_{p_i} u_3 + b_{p_i} u_4 \pmod{p_i}$$
$$u_2 \equiv b_{p_i} u_3 - a_{p_i} u_4 \pmod{p_i}$$

for $i = 1, \ldots, r$. By Lemma 2, Λ is a lattice and $d(\Lambda) \leq p_1^2 \cdots p_r^2 = m^2$. Let A be the set of points

$$A = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 \mid x_1^2 + x_2^2 + x_3^2 + x_4^2 < 2m\}.$$

Notice that A is symmetric about $\underline{\underline{0}}$, convex and $\mu(A) = \frac{\pi^2}{2}(2m)^2 = 2\pi^2 m^2$, since A is the sphere in $\mathbb{R}^4$ of radius $\sqrt{2m}$. By Minkowski's Convex Body Theorem II, since $2^4 d(\Lambda) \leq 2^4 m^2 < 2\pi^2 m^2 = \mu(A)$, there is a non-zero lattice point (u_1, u_2, u_3, u_4) in Λ which is in A. In particular

$$0 < u_1^2 + u_2^2 + u_3^2 + u_4^2 < 2m. \tag{12}$$

But

$$u_1^2 + u_2^2 + u_3^2 + u_4^2 \equiv (a_{p_i}^2 + b_{p_i}^2 + 1)(u_3^2 + u_4^2) \equiv 0 \pmod{p_i} \quad \text{for } i = 1, \ldots, r.$$

Thus

$$u_1^2 + u_2^2 + u_3^2 + u_4^2 \equiv 0 \pmod{m}$$

by the Chinese Remainder Theorem. Therefore by (12),

$$u_1^2 + u_2^2 + u_3^2 + u_4^2 = m$$

as required. □

Proposition 1. *Let R be a positive real number and let n be a positive integer. The volume of the sphere of radius R in $\mathbb{R}^n$ is $\omega_n R^n$ where $\omega_n = \frac{\pi^{n/2}}{\Gamma\left(1+\frac{n}{2}\right)}$.*

Proof:

It suffices to prove that ω_n is the volume of the unit sphere $\{(x_1, \dots, x_n) \mid x_1^2 + \cdots + x_n^2 \leq 1\}$. We have $\omega_1 = 2$ and $\omega_2 = \pi$. We then compute ω_n inductively for $n = 3, 4, \dots$ by the formula

$$\omega_n = \int_{x_1^2+\cdots+x_n^2 \leq 1} d_{x_1} \cdots d_{x_n} = \int_{-1}^{1} \int_{-1}^{1} \left(\int_{\mathbb{R}^{n-2}} g(x_1, \dots, x_n) d_{x_1} \cdots d_{x_{n-2}} \right) d_{x_{n-1}} d_{x_n}$$

where g is the characteristic function of the unit sphere in $\mathbb{R}^n$. Thus by our inductive hypothesis

$$\begin{aligned} \omega_n &= \int_{x_n^2+x_{n-1}^2 \leq 1} \omega_{n-2}(1 - x_{n-1}^2 - x_n^2)^{\frac{n-2}{2}} \, dx_{n-1} dx_n \\ &= \omega_{n-2} \int_{x_n^2+x_{n-1}^2 \leq 1} (1 - x_{n-1}^2 - x_n^2)^{\frac{n-2}{2}} \, dx_{n-1} dx_n. \end{aligned}$$

Change to polar variables r and θ. Thus

$$\begin{aligned} \omega_n &= \omega_{n-2} \int_0^{2\pi} \left(\int_0^1 (1 - r^2)^{\frac{n-2}{2}} \, r dr \right) d\theta \\ &= 2\pi\omega_{n-2} \left[-\frac{(1 - r^2)^{n/2}}{n} \right]_0^1 = \frac{2\pi\omega_{n-2}}{n} \end{aligned}$$

Therefore

$$\omega_{2n} = \frac{2\pi}{2n} \cdot \frac{2\pi}{2(n-1)} \cdots \frac{2\pi}{4} \cdot \pi = \frac{\pi^n}{n!}$$

while

$$\begin{aligned} \omega_{2n+1} &= \frac{2\pi}{2n+1} \cdot \frac{2\pi}{2n-1} \cdots \frac{2\pi}{3} \cdot 2 \\ &= \frac{\pi}{n+\frac{1}{2}} \cdot \frac{\pi}{n-\frac{1}{2}} \cdots \frac{\pi}{\frac{3}{2}} \cdot 2 \\ &= \frac{\pi^n}{(n+\frac{1}{2})(n-\frac{1}{2}) \cdots \frac{3}{2} \cdot \frac{1}{2}} \end{aligned}$$

Recall that Γ function satisfies the relation $\Gamma(x+1) = x\Gamma(x)$ for $x \in \mathbb{R}^+$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. The result now follows. □

Proposition 2. *Let Λ be a lattice in $\mathbb{R}^n$. There is a non-zero point $\underline{\underline{x}}$ in Λ for which*

$$0 < \underline{\underline{x}} \cdot \underline{\underline{x}} = x_1^2 + \cdots + x_n^2 \leq 4 \left(\omega_n^{-1} d(\Lambda) \right)^{2/n}.$$

Proof:

We apply Minkowski's Convex Body Theorem II to the set A where

$$A = \{(y_1, \dots, y_n) \in \mathbb{R}^n \mid y_1^2 + y_2^2 + \cdots + y_n^2 \leq t\}$$

with $t = 4\left(\omega_n^{-1} d(\Lambda)\right)^{2/n}$. Since A is convex, symmetric about $\underline{\underline{0}}$ and

$$\mu(A) = \omega_n t^{n/2} = \omega_n 2^n \omega_n^{-1} d(\Lambda) = 2^n d(\Lambda).$$

Further A is compact and our result follows. □

The natural question to ask is how good this result is.

Minkowski proved that for each $n \in \mathbb{Z}^+$ there exists a lattice Λ in R^n for which

$$\min_{\underline{\underline{x}} \in \Lambda, \underline{\underline{x}} \neq \underline{\underline{0}}} \underline{\underline{x}} \cdot \underline{\underline{x}} \geq \left(\omega_n^{-1} d(\Lambda)\right)^{2/n}.$$

Thus Proposition 2 is best possible up to a factor of 4. For dimensions $n < 8$ the best possible version of Proposition 2 is known.

Rogers proved that one can replace $4\omega_n^{-2/n}$ by $4\left(\frac{\sigma_n}{\omega_n}\right)^{2/n}$ where $\sigma_n = \frac{vol(B)}{vol(\Delta n)}$ where Δn is an equilateral n-simplex in $\mathbb{R}^n$ with edge length 2 and B is the set of all points in Δn with distance ≤ 1 from a vertex of Δn. One can prove that

$$\sigma_n \sim \frac{n}{2^{n/2}} \text{ as } n \to \infty.$$

We have

$$\omega_n^{-2/n} \sim \frac{n}{2\pi e}, \quad 4\left(\frac{\sigma_n}{\omega_n}\right)^{2/n} \sim \frac{n}{\pi e}, \quad 4\omega_n^{-2/n} \sim \frac{2n}{\pi e}$$

where the former is probably the truth.

We will now address the question of algorithms for producing good approximations given by Theorem 4.

For Dirichlet's theorem (Theorem 1) we have the continued fraction algorithm. In general we shall appeal to an algorithm based on the L^3- algorithm, named after Lenstra, Lenstra, and Lovasz, which gives us an efficient way to find small vectors in a lattice.

Let $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_n$ be a basis for a lattice Λ in $\mathbb{R}^n$. The Gram-Schmidt orthogonalization process produces vectors $\underline{\underline{b}}_i^*$ for $i = 1, \ldots, n$ and real numbers μ_{ij} for $1 \leq j < i \leq n$ inductively by

$$\underline{\underline{b}}_i^* = \underline{\underline{b}}_i - \sum_{j=1}^{i-1} \mu_{ij} \underline{\underline{b}}_j^* \quad \text{with } \mu_{ij} = \frac{(\underline{\underline{b}}_i, \underline{\underline{b}}_j^*)}{(\underline{\underline{b}}_j^*, \underline{\underline{b}}_j^*)}.$$

where $(\ ,\ )$ denotes the standard inner product on $\mathbb{R}^n$.

By construction $\underline{\underline{b}}_i^*$ is the projection of $\underline{\underline{b}}_i$ on the orthogonal complement of the Span of $\underline{\underline{b}}_1^*, \ldots, \underline{\underline{b}}_{i-1}^*$. Further $\mathrm{Sp}\{\underline{\underline{b}}_1^*, \ldots, \underline{\underline{b}}_i^*\} = \mathrm{Sp}\{\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_i\}$ for $i = 1, \ldots, n$.

Definition:

A basis $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_n$ for a lattice Λ in $\mathbb{R}^n$ is said to be reduced if:

i) $|\mu_{ij}| \leq \frac{1}{2}$ for $1 \leq j < i \leq n$.

ii) $|\underline{\underline{b}}_i^* + \mu_{ii-1}\underline{\underline{b}}_{i-1}^*|^2 \geq \frac{3}{4}|\underline{\underline{b}}_{i-1}^*|^2$ for $i = 2, \ldots, n$.

(Here $|\underline{\underline{x}}| = (\underline{\underline{x}} \cdot \underline{\underline{x}})^{1/2} = (x_1^2 + \cdots + x_n^2)^{1/2}$, is the Euclidean length of the vector $\underline{\underline{x}}$.)

Remark:

1) The vectors $\underline{\underline{b}}_i^* + \mu_{ii-1}\underline{\underline{b}}_{i-1}^*$ and $\underline{\underline{b}}_{i-1}^*$ are the orthogonal projections of $\underline{\underline{b}}_i$ and $\underline{\underline{b}}_{i-1}$ respectively on the complement of $\mathrm{Sp}\{\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_{i-2}\}$.

2) The notion of a reduced basis is not canonical in the sense that the constant $\frac{3}{4}$ could be replaced by any real number y with $\frac{1}{4} < y < 1$.

We'll now deduce some properties of reduced bases for a lattice. Then we'll give the algorithm to transform a given basis to a reduced basis.

Proposition 3. *Let $\underline{b}_1, \dots, \underline{b}_n$ be a reduced basis for a lattice Λ in $\mathbb{R}^n$ and let $\underline{b}_1^*, \dots, \underline{b}_n^*$ be obtained from the Gram-Schmidt orthogonalization process. Then*

i) $|\underline{b}_j|^2 \le 2^{i-1}|\underline{b}_i^*|^2 \quad$ *for* $1 \le j \le i \le n$.

ii) $d(\Lambda) \le \prod_{i=1}^{n} |\underline{b}_i| \le 2^{\frac{n(n-1)}{4}} d(\Lambda)$.

iii) $|\underline{b}_1| \le 2^{\frac{n-1}{4}} d(\Lambda)^{\frac{1}{n}}$.

Proof:

By the definition of a reduced basis

$$|\underline{b}_i^* + \mu_{ii-1}\underline{b}_{i-1}^*|^2 \ge \frac{3}{4}|\underline{b}_{i-1}^*|^2 \quad \text{with } |\mu_{ii-1}| \le \frac{1}{2}.$$

Note that

$$|\underline{b}_i^* + \mu_{ii-1}\underline{b}_{i-1}^*|^2 = (\underline{b}_i^* + \mu_{ii-1}\underline{b}_{i-1}^*, \underline{b}_i^* + \mu_{ii-1}\underline{b}_{i-1}^*) = |\underline{b}_i^*|^2 + \mu_{ii-1}^2|\underline{b}_{i-1}^*|^2.$$

Therefore

$$|\underline{b}_i^*| \ge (\frac{3}{4} - \frac{1}{4})|\underline{b}_{i-1}^*|^2 = \frac{1}{2}|\underline{b}_{i-1}^*|.$$

Thus, by induction,

$$|\underline{b}_j^*|^2 \le 2^{i-j}|\underline{b}_i^*|^2 \quad \text{for } 1 \le j \le i \le n. \tag{13}$$

Next observe that

$$\begin{aligned}
|\underline{b}_i|^2 &= |\underline{b}_i^*|^2 + \sum_{j=1}^{i-1} \mu_{ij}^2 |\underline{b}_j^*|^2 \\
&\le |\underline{b}_i^*|^2 + \sum_{j=1}^{i-1} \frac{1}{4} 2^{i-j} |\underline{b}_i^*|^2 \quad \text{by (13)} \\
&\le |\underline{b}_i^*|^2 \left(1 + \frac{1}{4}(2^i - 2)\right) \\
&\le 2^{i-1}|\underline{b}_i^*|^2
\end{aligned}$$

Thus we have

$$|\underline{b}_j|^2 \le 2^{j-1} \cdot 2^{i-j}|\underline{b}_i^*|^2 = 2^{i-1}|\underline{b}_i^*|^2 \qquad \text{for } 1 \le j \le i \le n.$$

This proves i).

We have $d(\Lambda) = |\det(\underline{b}_1, \dots, \underline{b}_n)|$ and so by Hadamard's inequality

$$d(\Lambda) \le |\underline{b}_1| \cdots |\underline{b}_n|.$$

By constuction $|\det(\underline{b}_1, \dots, \underline{b}_n)| = |\det(\underline{b}_1^*, \dots, \underline{b}_n^*)|$ But the $\underline{b}_i^*$ are orthogonal and so by Hadamard's inequality

$$d(\Lambda) = |\underline{b}_1^*| \cdots |\underline{b}_n^*|.$$

By i)

$$|\underline{b}_j| \le 2^{\frac{i-1}{2}}|\underline{b}_i^*| \qquad \text{for } 1 \le j \le i \le n.$$

Thus

$$\begin{aligned}
|\underline{b}_1| \cdots |\underline{b}_n| &\le 2^{\frac{0}{2}} \cdot 2^{\frac{1}{2}} \cdots 2^{\frac{n-1}{2}} |\underline{b}_1^*| \cdots |\underline{b}_n^*| \\
&\le 2^{\frac{1}{2}\left(\frac{n(n-1)}{2}\right)} d(\Lambda)
\end{aligned}$$

which proves ii).

To prove *iii)* we apply *i)* with $j = 1$ to get

$$|\underline{\underline{b}}_1| \leq 2^{\frac{i-1}{2}} |\underline{\underline{b}}_i^*| \quad \text{for } i = 1, \ldots, n.$$

Thus

$$\begin{aligned} |\underline{\underline{b}}_1|^n &\leq 2^{\frac{0}{2}} \cdot 2^{\frac{1}{2}} \cdots 2^{\frac{n-1}{2}} |\underline{\underline{b}}_1^*| \cdots |\underline{\underline{b}}_n^*| \\ &\leq 2^{\frac{1}{2}\left(\frac{n(n-1)}{2}\right)} d(\Lambda) \\ |\underline{\underline{b}}_1| &\leq 2^{\frac{n-1}{4}} d(\Lambda)^{\frac{1}{n}} \end{aligned}$$

as required. □

Proposition 4. *Let Λ be a lattice in $\mathbb{R}^n$ with reduced basis $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_n$. Then for any non-zero vector $\underline{\underline{x}}$ in Λ we have*

$$|\underline{\underline{b}}_1|^2 \leq 2^{n-1} |\underline{\underline{x}}|^2.$$

Proof:

Write $\underline{\underline{x}}$ out in terms of the basis $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_n$, so

$$\underline{\underline{x}} = g_1 \underline{\underline{b}}_1 + \cdots + g_n \underline{\underline{b}}_n \quad \text{with } g_i \in \mathbb{Z}.$$

Further we have

$$\underline{\underline{x}} = \lambda_1 \underline{\underline{b}}_1^* + \cdots + \lambda_n \underline{\underline{b}}_n^* \quad \text{with } \lambda_i \in \mathbb{R}.$$

Let i be the largest index for which $g_i \neq 0$. Then on examing the Gram-Schmidt process we see that $\lambda_i = g_i$. Then

$$|\underline{\underline{x}}|^2 \geq \lambda_i^2 |\underline{\underline{b}}_i^*|^2 = g_i^2 |\underline{\underline{b}}_i^*|^2 \geq |\underline{\underline{b}}_i^2|^2.$$

By Prop 3,

$$2^{i-1} |\underline{\underline{x}}|^2 \geq 2^{i-1} |\underline{\underline{b}}_i^*|^2 \geq |\underline{\underline{b}}_1|^2$$

and the result follows since $i \leq n$. □

Proposition 5. *Let Λ be a lattice in $\mathbb{R}^n$ with reduced basis $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_n$. Let $\underline{\underline{x}}_1, \ldots \underline{\underline{x}}_t$ be linearly independent points of Λ. Then*

$$|\underline{\underline{b}}_j|^2 \leq 2^{n-1} \max\{|\underline{\underline{x}}_1|^2, \ldots, |\underline{\underline{x}}_t|^2\} \quad \textit{for } j = 1, \ldots, t.$$

Proof:

Write

$$\underline{\underline{x}}_j = \sum_{i=1}^{n} g_{ij} \underline{\underline{b}}_i \quad \text{with } g_{ij} \in \mathbb{Z},$$

for $j = 1, \ldots, t$. For each j let $i(j)$ be the largest index for which $g_{i(j)j} \neq 0$. As in the proof of Prop 4,

$$|\underline{\underline{x}}_j|^2 \geq |\underline{\underline{b}}_{i(j)}^*|^2 \quad \text{for } j = 1, \ldots, t.$$

Let us renumber the $\underline{\underline{x}}_j$'s so that $i(1) \leq i(2) \leq \cdots \leq i(j)$. Notice that $j \leq i(j)$ for $j = 1, \ldots, t$ since if $i(j) < j$ then $\underline{\underline{x}}_1, \ldots, \underline{\underline{x}}_j$ would be linearly dependent. Therefore by Prop 3 *i)*

$$|\underline{\underline{b}}_j|^2 \leq 2^{i(j)-1} |\underline{\underline{b}}_{i(j)}^*|^2 \leq 2^{i(j)-1} |\underline{\underline{x}}_j|^2 \leq 2^{n-1} |\underline{\underline{x}}_j|^2 \quad \text{for } j = 1, \ldots, t.$$

□

The L^3-algorithm gives an efficient way of transforming a basis for a lattice to a reduced basis.

L^3-Algorithm

Let $\underline{\underline{b}}_1, \dots, \underline{\underline{b}}_n$ be a basis for Λ. Use Gram-Schmidt to compute $\underline{\underline{b}}_1^*, \dots, \underline{\underline{b}}_n^*$ and μ_{ij}. Throughout the algorithm we'll change our basis to a new basis for Λ many times. Each time we do so, we recalculate the $\underline{\underline{b}}_i^*$'s and the μ_{ij}'s.

At each step of the algorithm there is a current subscript k from $\{1, 2, \dots, n+1\}$. We start with $k = 2$. We shall now perform a sequence of steps that starts from and returns to a situation where

1) $|\mu_{ij}| \leq \frac{1}{2}$ for $1 \leq j < i < k$

and

2) $|\underline{\underline{b}}_i^* + \mu_{ii-1}\underline{\underline{b}}_{i-1}^*|^2 \geq \frac{3}{4}|\underline{\underline{b}}_{i-1}^*|^2$ for $1 < i < k$.

Note that 1) and 2) are trivially satisfied when $k = 2$. When $k = n+1$ the basis is reduced and the algorithm terminates. So, suppose that $k \leq n$. In this case we first achieve

3)$|\mu_{kk-1}| \leq \frac{1}{2}$ for $k > 1$.

If 3) does not hold let r be the integer closest to μ_{kk-1}. Then replace $\underline{\underline{b}}_k$ by $\underline{\underline{b}}_k - r\underline{\underline{b}}_{k-1}$. Plainly this does not change Λ. Further, the numbers μ_{kj} are replaced by $\mu_{kj} - r\mu_{k-1j}$ for $j < k-1$ and μ_{kk-1} is replaced by $\mu_{kk-1} - r$. The other μ_{ij}'s are unchanged as are the other $\underline{\underline{b}}_i^*$'s. We now distinguish two cases:

Case 1) If $k \geq 2$ and $|\underline{\underline{b}}_k^* + \mu_{kk-1}\underline{\underline{b}}_{k-1}^*|^2 < \frac{3}{4}|\underline{\underline{b}}_{k-1}^*|^2$.

Case 2) $k = 1$ or $|\underline{\underline{b}}_k^* + \mu_{kk-1}\underline{\underline{b}}_{k-1}^*|^2 \geq \frac{3}{4}|\underline{\underline{b}}_{k-1}^*|^2$.

If we are in Case 1, we interchange $\underline{\underline{b}}_k$ and $\underline{\underline{b}}_{k-1}$ and leave all the other $\underline{\underline{b}}_i$'s unchanged. Notice that $\underline{\underline{b}}_k^*$, $\underline{\underline{b}}_{k-1}^*$ are recalculated as are the numbers μ_{kk-1}, μ_{k-1j}, μ_{kj} for $j < k-1$ and the numbers μ_{ik-1}, μ_{ik} for $i > k$. Let us call our new basis $\underline{\underline{c}}_1, \dots, \underline{\underline{c}}_n$. So, $\underline{\underline{c}}_i = \underline{\underline{b}}_i$ for $i \neq k, k-1$, $\underline{\underline{c}}_k = \underline{\underline{b}}_{k-1}$, and $\underline{\underline{c}}_{k-1} = \underline{\underline{b}}_k$. Then $\underline{\underline{c}}_{k-1}^* = \underline{\underline{b}}_k^* + \mu_{kk-1}\underline{\underline{b}}_{k-1}^*$ is the projection of $\underline{\underline{b}}_k$ on the orthogonal compliment of the $\mathrm{Sp}\{\underline{\underline{b}}_1^*, \dots, \underline{\underline{b}}_{k-2}^*\}$. Thus $|\underline{\underline{c}}_{k-1}^*|^2 < \frac{3}{4}|\underline{\underline{b}}_{k-1}^*|^2$. Therefore the new $\underline{\underline{b}}_{k-1}^* (= \underline{\underline{c}}_{k-1}^*)$ is such that $|\underline{\underline{b}}_{k-1}^*|^2$ is less than $\frac{3}{4}$ of what it was before. We now replace k by $k-1$ and return to the start of the alogirthm.

If we are in Case 2 then we want to achieve $|\mu_{kj}| \leq \frac{1}{2}$ for $1 \leq j \leq k-1$. To accomplish this we first consider the largest integer l for which $|\mu_{kl}| > \frac{1}{2}$. Notice that $l < k-1$. Let r be the nearest integer to μ_{kl} and replace $\underline{\underline{b}}_k$ by $\underline{\underline{b}}_k - r\underline{\underline{b}}_l$. The numbers μ_{kj} with $j < l$ are replaced by $\mu_{kj} - r\mu_{lj}$ and μ_{kl} is replaced by $\mu_{kl} - r$. Note that the other μ_{ij}'s are unchanged and all of the $\underline{\underline{b}}_i^*$'s are unchanged. We now repeat this procedure until $|\mu_{kj}| \leq \frac{1}{2}$ for $1 \leq j \leq k-1$. We then replace k by $k+1$ and return to the start of the algorithm.

Question: Does the algorithm terminate? Yes! To show this we need to introduce the following quantities:

$$
\begin{aligned}
d_i &= \det\Big((\underline{\underline{b}}_k, \underline{\underline{b}}_l)\Big)_{\substack{k=1,\dots,i \\ l=1,\dots,i}} \qquad \text{for i=1, } \dots \text{,n} \\
&= \det\Big((\underline{\underline{b}}_1, \dots, \underline{\underline{b}}_i)(\underline{\underline{b}}_1, \dots, \underline{\underline{b}}_i)^{tr}\Big) \\
&= \det\Big((\underline{\underline{b}}_1^*, \dots, \underline{\underline{b}}_i^*)(\underline{\underline{b}}_1^*, \dots, \underline{\underline{b}}_i^*)^{tr}\Big)
\end{aligned}
$$

since the determinant is unchanged when we add a multiple of one row to another. In addition we put

$$D = \prod_{i=1}^{n} d_i.$$

Observe that $d_n = d(\Lambda)^2$. Let Λ_i be the lattice generated by $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_i$ in the i-dimensional space spanned by these vectors. Then $d_i = |\underline{\underline{b}}_1^*|^2 \cdots |\underline{\underline{b}}_i^*|^2 = d(\Lambda_i)^2$.

As we proceed with the L^3 algorithm, D does not change in Case 2 since the $\underline{\underline{b}}_i^*$'s don't change. In Case 1, we interchange $\underline{\underline{b}}_k$ and $\underline{\underline{b}}_{k-1}$ hence change $\underline{\underline{b}}_k^*$ and $\underline{\underline{b}}_{k-1}^*$. So, $d_1, \ldots, d_{k-2}, d_k, \ldots, d_n$ are unchanged. This leaves d_{k-1} to be considered. The new value of $|\underline{\underline{b}}_{k-1}^*|^2$ is less than $\frac{3}{4}$ of the old value, so the new value of d_{k-1} is less than $\frac{3}{4}$ of the old value. Further the new value of D is less than $\frac{3}{4}$ of the old value of D.

To show the algorithm terminates it is enough to show that D is bounded from below by a positive number which depends on Λ. Put $m(\Lambda) = \min\{\underline{\underline{x}} \cdot \underline{\underline{x}} \,|\, x \in \Lambda, \underline{\underline{x}} \neq \underline{\underline{0}}\}$. By Proposition 2,

$$m(\Lambda_i) \leq 4\left(\omega_i^{-1} d(\Lambda_i)\right)^{2/i}$$

hence

$$d_i = d(\Lambda_i)^2 \geq m(\Lambda_i)^i 4^{-i} \omega_i^2.$$

Since $m(\Lambda_i) \geq m(\Lambda)$ for $i = 1, \ldots, n$ and $\Lambda_n = \Lambda$,

$$d_i \geq (m(\Lambda))^i 4^{-i} \omega_i^2 \quad \text{for } i = 1, \ldots, n.$$

Thus

$$D = d_1 \cdots d_n \geq \left(\frac{m(\Lambda)}{4}\right)^{\frac{n(n+1)}{2}} (\omega_1 \cdots \omega_n)^2,$$

as required. Thus we can only pass through case 1 finitely many times. We pass through at most $n-1$ times more than we pass through case 1. Thus the algorithm terminates.

In fact, the algorithm is very efficient. Lentstra, Lenstra and Lovasz proved that if Λ is a sublattice of $\mathbb{Z}^n$ with basis $\underline{\underline{b}}_1, \ldots, \underline{\underline{b}}_n$ and B is a real number with $|\underline{\underline{b}}_i|^2 \leq B$ for $i = 1, \ldots, n$ then the number of arithmetical operations required for the algorithm is $O(n^4 \log B)$ and the integers on which these operations are performed have binary length $O(n \log B)$. Thus the algorithm runs in polynomial time in terms of the input.

Note: An arithmetical operation is an addition, subtraction, multiplication or division of two integers.

Let $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Given ϵ with $0 < \epsilon < 1$, how do we efficiently find integers $p_1, \ldots, p_n$ and q for which $1 \leq q \leq 2^{\frac{n(n+1)}{4}} \varepsilon^{-n}$ and

$$|q\alpha_i - p_i| < \epsilon \quad \text{for } i = 1, \ldots, n.$$

If $\alpha_1, \ldots, \alpha_n$ are in $\mathbb{Q}$ and ϵ is in $\mathbb{Q}$ then we can use the L^3 algorithm to solve this problem in polynomial time in terms of the input.

To do so we consider the lattice Λ generated by the rows of the matrix

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & 1 & 0 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n & \delta \end{pmatrix}$$

where $\delta = 2^{-\frac{n(n+1)}{4}}\varepsilon^{n+1}$. We have $d(\Lambda) = \delta$. By L^3 we can find a short non-zero vector $\underline{\underline{b}}_1 = \underline{\underline{b}}$ in Λ. It has the form $(q\alpha_1 - p_1, q\alpha_2 - p_2, \dots, q\alpha_n - p_n, q\delta)$ with q and $p_1, \dots, p_n$ integers not all zero. Note that we can suppose that $q \geq 0$ by replacing $\underline{\underline{b}}$ by $-\underline{\underline{b}}$ if necessary. Further by Proposition 3, iii)

$$|\underline{\underline{b}}| \leq 2^{\frac{n}{4}} d(\Lambda)^{\frac{1}{n+1}} = 2^{\frac{n}{4}} 2^{-\frac{n}{4}} \cdot \epsilon = \epsilon < 1.$$

Notice that $q \neq 0$ since the $\underline{\underline{b}} = (p_1, \dots, p_n, 0)$ and since $|\underline{\underline{b}}| < 1$ this would mean $p_1 = \cdots = p_n = q = 0$ which is a contradition. Thus

$$|q\alpha_i - p_i| < \epsilon \quad \text{for } i = 1, \dots, n$$

and

$$|q\delta| < \epsilon.$$

Therefore we have $1 \leq q \leq \epsilon\delta^{-1} = 2^{\frac{n(n+1)}{4}}\epsilon^{-n}$.

Question: Given $\alpha_1, \dots, \alpha_n$ in $\mathbb{R}$, how do we find a small linear form in the α_i's?

Let ϵ be a real number with $0 < \epsilon < 1$. We want to find in an efficient manner integers $q_1, \dots, q_n$ and p not all zero for which $|q_1\alpha_1 + \cdots + q_n\alpha_n| < \epsilon$ and $|q_i| \leq 2^{\frac{n+1}{4}}\varepsilon^{-\frac{1}{n}}$. Let Λ be the lattice generated by the rows of the following matrix

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \alpha_1 & \delta & 0 & \cdots & 0 \\ \alpha_2 & 0 & \delta & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \alpha_n & 0 & \cdots & 0 & \delta \end{pmatrix} \quad \text{where } \delta = \left(\frac{\varepsilon^{1/n}}{2^{1/4}}\right)^{n+1}.$$

Then the L^3 algorithm yields a vector $\underline{\underline{b}}$ in Λ with

$$|\underline{\underline{b}}| \leq 2^{n/4} d(\Lambda)^{\frac{1}{n+1}} = 2^{n/4}\delta^{\frac{n}{n+1}} = 2^{\frac{n}{4}}\epsilon 2^{-\frac{n}{4}} = \epsilon.$$

We have

$$\underline{\underline{b}} = (q_1\alpha_1 + \cdots + q_n\alpha_n - p, q_1\delta, q_2\delta, \dots, q_n\delta).$$

Thus

$$|q_i\delta| \leq 2^{n/4}\delta^{\frac{n}{n+1}} = \epsilon \quad \text{for } i = 1, \dots, n.$$

and

$$|q_1\alpha_1 + \cdots + q_n\alpha_n - p| < \epsilon.$$

Note that

$$|q_i| \leq \frac{2^{\frac{n}{4}}}{\delta^{\frac{1}{n+1}}} = \frac{2^{\frac{n}{4}}2^{\frac{1}{4}}}{\epsilon^{\frac{1}{n}}} = 2^{\frac{n+1}{4}}\epsilon^{-\frac{1}{n}} \quad \text{for } i = 1, \dots, n.$$

Take $\epsilon = \frac{1}{Q}$ to compare with Theorem 3.

Suppose now that α_{ij} for $i = 1, \dots, n$, $j = 1, \dots, m$ are real numbers. Consider the lattice Λ associated with the matrix

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots & \vdots & & \vdots \\ \vdots & & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ \alpha_{11} & \alpha_{21} & \cdots & \alpha_{n1} & \delta & & 0 \\ \vdots & \vdots & & \vdots & & \ddots & \\ \alpha_{1m} & \alpha_{2m} & \cdots & \alpha_{nm} & 0 & & \delta \end{pmatrix}$$

where

$$\delta = \left(2^{-\left(\frac{n+m-1}{4}\right)}\epsilon\right)^{\frac{n}{m}+1}.$$

By L^3 we find a vector $\underline{\underline{b}} \neq \underline{\underline{0}}$ in Λ with $|\underline{\underline{b}}| \leq \delta^{\frac{m}{n+m}} 2^{\frac{n+m-1}{4}}$. So, $|b| \leq 2^{-\left(\frac{n+m-1}{4}\right)} \cdot 2^{\frac{n+m-1}{4}}\epsilon = \epsilon$. We have

$$\underline{\underline{b}} = (q_1\alpha_{11} + q_2\alpha_{12} + \cdots + q_m\alpha_{1m} - p_1, \dots, q_1\alpha_{n1} + q_2\alpha_{n2} + \cdots + q_m\alpha_{nm} - p_n, q_1\delta, \dots, q_m\delta)$$

for some integers $q_1, q_2, \dots, q_m, p_1, \dots, p_n$ not all zero. Thus

$$|\alpha_{i1}q_1 + \cdots + \alpha_{im}q_m - p_i| \leq \epsilon \quad \text{for } i = 1, \dots, n$$

and

$$|q_i\delta| \leq \epsilon \quad \text{for } i = 1, \dots, m.$$

Thus

$$|q_i| \leq \frac{\epsilon}{\delta} = 2^{\frac{n+m-1}{4}\frac{n+m}{m}}\epsilon^{-\frac{n}{m}}.$$

Further, since $|\underline{\underline{b}}| \leq \epsilon$ and $\epsilon < 1$ not all of the q_i's are zero. Taking $\epsilon = \frac{1}{Q}$ we obtain the analogue of Theorem 4.

The special case when all of the α_{ij}'s are algebraic.

Let us first consider a single linear form. Let $\alpha_1, \dots, \alpha_n$ be real algebraic numbers with $1, \alpha_1, \dots, \alpha_n$ linearly independent over the rationals. Let d be the degree of $\mathbb{Q}(\alpha_1, \dots, \alpha_n)$ over $\mathbb{Q}$. Extend $1, \alpha_1, \dots, \alpha_n$ to a basis $1, \alpha_1, \dots, \alpha_n, \dots, \alpha_{d-1}$ for $\mathbb{Q}(\alpha_1, \dots, \alpha_n)$ over $\mathbb{Q}$. By Theorem 9, $\alpha_1, \dots, \alpha_{d-1}$ are badly approximable. Thus

$$|\alpha_1 q_1 + \cdots + \alpha_{d-1}q_{d-1} - p| > c_1 q^{-d+1},$$

for a positive number c_1 and for all d-tuples of integers $(q_1, \dots, q_{d-1}, p)$ where $q = \max\{|q_1|, \dots, |q_{d-1}|\}$ and $q > 0$. Thus, on taking $q_{n+1} = \cdots = q_{d-1} = 0$, we see that:

Theorem 15. *Suppose that $\alpha_1, \dots, \alpha_n$ are real algebraic numbers and $1, \alpha_1, \dots, \alpha_n$ are linearly independent over $\mathbb{Q}$. Put $d = [\mathbb{Q}(\alpha_1, \dots, \alpha_n) : \mathbb{Q}]$ then*

$$|\alpha_1 q_1 + \cdots + \alpha_n q_n - p| > c_1 q^{-d+1},$$

for all $n+1$-tuples of integers $(q_1, \dots, q_n, p)$ with $q = \max\{|q_1|, \dots, |q_n|\}$ and $q > 0$.

If $n = 1$ then Theorem 15 is Liouville's Theorem. For example $\alpha = \sum_{i=1}^{\infty} \frac{1}{10^{n!}}$ is transcendental.

The following result is a consequence of Schmidt's Subspace Theorem:

Theorem 16. *Let $\alpha_1, \dots, \alpha_n$ be real algebraic numbers with $1, \alpha_1, \dots, \alpha_n$ linearly independent over $\mathbb{Q}$. Let $\delta > 0$. There are only finitely many n-tuples of non-zero integers $(q_1, \dots, q_n)$ with*

$$|q_1 q_2 \cdots q_n|^{1+\delta} \cdot \|q_1\alpha_1 + \cdots + q_n\alpha_n\| < 1.$$

Here for any real number x, $\|x\|$ denotes the distance from x to the nearest integer.

Corollary 16: Let $\alpha_1, \dots, \alpha_n$ be real algebraic numbers with $1, \alpha_1, \dots, \alpha_n$ linearly independent over $\mathbb{Q}$. Let $\delta > 0$. Then there are only finitely many $n+1$-tuples of integers $(q_1, \dots, q_n, p)$ with $q = \max\{|q_1|, \dots, |q_n|\} > 0$ for which

$$|q_1\alpha_1 + \cdots + q_n\alpha_n - p| > \frac{1}{q^{n+\delta}}.$$

Proof: Apply Theorem 16 to all of the non-empty subsets of $\{\alpha_1, \dots, \alpha_n\}$. □

Schmidt also deduced:

Theorem 17. *Let $\alpha_1, \dots, \alpha_n$ be real algebraic numbers with $1, \alpha_1, \dots, \alpha_n$ linearly independent over $\mathbb{Q}$. Let $\delta > 0$. Then there are only finitely many positive integers q such that*

$$q^{1+\delta}\|q\alpha_1\| \cdots \|q\alpha_n\| < 1.$$

This implies

Corollary 17: Let $\alpha_1, \dots, \alpha_n$ be real algebraic numbers with $1, \alpha_1, \dots, \alpha_n$ linearly independent over $\mathbb{Q}$. Let $\delta > 0$. Then there are only finitely many rational n-tuples $(\frac{p_1}{q}, \dots, \frac{p_n}{q})$ with $q > 0$ and

$$\left|\alpha_i - \frac{p}{q}\right| < \frac{1}{q^{1+\frac{1}{n}+\delta}} \quad \text{for } i = 1, \dots, n.$$

Note that if you take $n = 1$ in Corollary 17 or Corollary 16 we get:

Roth's Theorem: Let α be a real irrational algebraic number. Let $\delta > 0$. There exist only finitely many rationals $\frac{p}{q}$ with $q > 0$ for which $\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^{2+\delta}}$.

Note that Roth's theorem is ineffective. It doesn't tell us how to find the approximations. In 1972, W. Schmidt established his Subspace Theorem from which Theorem 16 and 17 follow. It is a profound generalization of Roth's Theorem.

Theorem 18. *(Schmidt's Subspace Theorem)*
Suppose that $L_1(\underline{\underline{x}}), \dots, L_n(\underline{\underline{x}})$ are linearly independent linear forms in $\underline{\underline{x}}$ with (real or complex) algebraic coefficients. Let $\delta > 0$. There are finitely many proper subspaces $T_1, \dots, T_w$ of $\mathbb{R}^n$ such that every integer point $\underline{\underline{x}} = (x_1, \dots, x_n) \neq \underline{\underline{0}}$ for which $|L_1(\underline{\underline{x}}) \cdots L_n(\underline{\underline{x}})| < \frac{1}{|\underline{\underline{x}}|^\delta}$ lies in T_i for some i with $1 \leq i \leq w$.

Notes:

1) The Subspace Theorem is ineffective just as Roth's Theorem in the following sense. Given the linear forms and δ the proof does not yield a method for determining the proper subspaces $T_1, \dots, T_w$.

2) The integer points in a subspace T span a rational linear subspace, so a subspace determined by a linear equation with rational coefficients. Thus $T_1, \dots, T_w$ are rational linear subspaces.

3) We won't give the proof of the Subspace Theorem.

Now we'll deduce Theorem 17 from the Subspace Theorem.

Proof: (of theorem 17) Let q be a positive integer for which $q^{1+\delta}\|\alpha_1 q\|\cdots\|\alpha_n q\| < 1$. Let p_i be an integer for which $\|\alpha_i q\| = \alpha_i q - p_i$ for $i = 1, \dots, n$. Put $\underline{\underline{x}} = (x_1, \dots, x_{n+1}) = (p_1, \dots, p_n, q)$ and let $C_1, C_2, \dots$ be positive numbers which depend on n and $\alpha_1, \dots, \alpha_n$ only. Plainly

$$\overline{|\underline{\underline{x}}|} < C_1 q.$$

Consider the linear forms

$$L_i(\underline{\underline{x}}) = \alpha_i x_{n+1} - x_i \quad \text{for } i = 1, \dots, n$$

and

$$L_{n+1}(\underline{\underline{x}}) = x_{n+1}.$$

Notice that $L_1(\underline{\underline{x}}), \dots, L_{n+1}(\underline{\underline{x}})$ are $n+1$ linearly independent linear forms in $x_1, \dots, x_{n+1}$ with algebraic coefficients. Then

$$|L_1(\underline{\underline{x}})\cdots L_{n+1}(\underline{\underline{x}})| = \|\alpha_1 q\|\cdots\|\alpha_n q\| \cdot q$$

hence

$$|L_1(\underline{\underline{x}})\cdots L_{n+1}(\underline{\underline{x}})| < \frac{1}{q^\delta} < \frac{1}{\overline{|\underline{\underline{x}}|}^{\delta/2}}$$

provided that q is sufficiently large. Thus by the Subspace Theorem $\underline{\underline{x}}$ lies in one of a finite collection of proper subspaces $T_1, \dots, T_w$ of $\mathbb{R}^{n+1}$. Say $\underline{\underline{x}}$ lies in T_1. Since T_1 is a rational subspace of $\mathbb{R}^{n+1}$ and so there exists rational numbers $c_1, \dots, c_{n+1}$ not all zero such that

$$c_1 x_1 + \cdots + c_{n+1} x_{n+1} = 0$$

hence that

$$c_1 p_1 + \cdots + c_n p_n + c_{n+1} q = 0.$$

Note that we have

$$\begin{aligned} |c_1(\alpha_1 q - p_1) + c_2(\alpha_2 q - p_2) + \cdots + c_n(\alpha_n q - p_n)| &= |c_1\alpha_1 q + \cdots + c_n\alpha_n q - c_1 p_1 - \cdots - c_n p_n| \\ &= |c_1\alpha_1 q + \cdots + c_n\alpha_n q + c_{n+1} q| \\ &= |c_1\alpha_1 + \cdots + c_n\alpha_n + c_{n+1}| q. \end{aligned}$$

Since $1, \alpha_1, \dots, \alpha_n$ are linearly independent over $\mathbb{Q}$, $|c_1\alpha_1 + \cdots + c_n\alpha_n + c_{n+1}| = C_2 > 0$. Therefore,

$$|c_1|\,|\alpha_1 q - p_1| + \cdots + |c_n|\,|\alpha_n q - p_n| \geq C_2 q$$

hence $|c_1| + \cdots + |c_n| \geq C_2 q$. Thus q is bounded. □

We now deduce Theorem 16 from the Subspace Theorem.

Proof: (of theorem 16). We prove the result by induction on n. For $n = 1$ the result follows from Theorem 17. Suppose that $q_1, \dots, q_n$ satisfy the hypotheses of Theorem 16 and choose p so that

$$\|\alpha_1 q_1 + \cdots + \alpha_n q\| = \alpha_1 q_1 + \cdots + \alpha_n q_n - p.$$

Write $\underline{\underline{x}} = (q_1, \dots, q_n, p)$. Then $\overline{|\underline{\underline{x}}|} < C_3 q$ where $q = \max\{|q_1|, \dots, |q_n|\} > 0$. Put

$$L_i(\underline{\underline{x}}) = x_i \quad \text{for } i = 1, \dots, n \quad \text{and} \quad L_{n+1}(\underline{\underline{x}}) = \alpha_1 x_1 + \cdots \alpha_n x_n - x_{n+1}.$$

Then $L_1, \dots, L_{n+1}$ are $n+1$ linearly independent linear forms with algebraic coefficients. Then with $\underline{\underline{x}} = (q_1, \dots, q_n, p)$,

$$|L_1(\underline{\underline{x}})\cdots L_n(\underline{\underline{x}})| = |q_1\cdots q_n| \cdot |\alpha_1 q_1 + \cdots + \alpha_n q_n - p|$$

hence, since $q_1, \dots, q_n$ are all non-zero,

$$|L_1(\underline{\underline{x}}) \cdots L_n(\underline{\underline{x}})| < \frac{1}{|q_1 \cdots q_n|^{\delta}} < \frac{1}{|\underline{\underline{x}}|^{\delta/2}}$$

for q sufficiently large. By the Subspace Theorem $\underline{\underline{x}} = (q_1, \dots, q_n, p)$ lies in one of finitely many proper linear subspaces $T_1, \dots, T_w$ of $\mathbb{R}^{n+1}$, say T_1. Further there exist rationals $c_1, \dots, c_{n+1}$ such that T_1 is defined by

$$c_1 x_1 + \cdots + c_{n+1} x_{n+1} = 0.$$

Since T_1 is a proper subspace not all of the c_i's are zero. Suppose first that $c_j \neq 0$ for some j with $1 \leq j \leq n$. Then, without loss of generality, we may suppose that $c_n \neq 0$. We have

$$c_n q_n = -c_1 q_1 - \cdots - c_{n-1} q_{n-1} - c_{n+1} p$$

hence

$$c_n \alpha_n q_n = -c_1 \alpha_n q_1 - \cdots - c_{n-1} \alpha_n q_{n-1} - c_{n+1} \alpha_n p$$

Thus

$$\begin{aligned} |c_n|\,|\alpha_1 q_1 + \cdots + \alpha_n q_n - p| &= |c_n \alpha_1 q_1 + \cdots + c_n \alpha_n q_n - c_n p| \\ &= |(c_n \alpha_1 q_1 - c_1 \alpha_n q_1) + \cdots + (c_n \alpha_{n-1} q_{n-1} - c_{n-1} \alpha_n q_{n-1}) - (c_n p + c_{n+1} \alpha_n p)| \\ &= |(c_n \alpha_1 - c_1 \alpha_n) q_1 + \cdots + (c_n \alpha_{n-1} - c_{n-1} \alpha_n) q_{n-1} - (c_n + c_{n+1} \alpha_n) p| \\ &= |c_n + c_{n+1} \alpha_n| \left| \left(\frac{c_n \alpha_1 - c_1 \alpha_n}{c_n + c_{n+1} \alpha_n} \right) q_1 + \cdots + \left(\frac{c_n \alpha_{n-1} - c_{n-1} \alpha_n}{c_n + c_{n+1} \alpha_n} \right) q_{n-1} - p \right| \end{aligned}$$

Put $\alpha_i' = \frac{c_n \alpha_i - c_i \alpha_n}{c_n + c_{n+1} \alpha_n}$ for $i = 1, \dots, n-1$, then

$$|c_n||\alpha_1 q_1 + \cdots + \alpha_n q_n - p| = |c_n + c_{n+1} \alpha_n||\alpha_1' q_1 + \cdots + \alpha_{n-1}' q_{n-1} - p|$$

Thus

$$\|\alpha_1' q_1 + \cdots + \alpha_{n-1}' q_{n-1}\| < \frac{C}{|q_1 \cdots q_n|^{1+\delta}} < \frac{1}{|q_1 \cdots q_{n-1}|^{1+\frac{\delta}{2}}}$$

provided that $q = \max\{q_1, \dots, q_n\}$ is sufficiently large. Note that $1, \alpha_1', \dots, \alpha_{n-1}'$ are $\mathbb{Q}$-linearly independent since if $\lambda_1 \alpha_1' + \cdots + \lambda_{n-1} \alpha_{n-1}' + \lambda_n = 0$ with $\lambda_i \in \mathbb{Q}$ for $i = 1, \dots, n$ then

$$\lambda_1(c_n \alpha_1 - c_1 \alpha_n) + \cdots + \lambda_{n-1}(c_n \alpha_{n-1} - c_{n-1} \alpha_n) + \lambda_n(c_n + c_{n+1} \alpha_n) = 0$$

$$c_n(\lambda_1 \alpha_1 + \cdots + \lambda_{n-1} \alpha_{n-1}) - (c_1 \lambda_1 + \cdots + c_{n-1} \lambda_{n-1} - c_{n+1} \lambda_n) \alpha_n + \lambda_n c_n = 0$$

Since $c_n \neq 0$ and $1, \alpha_1, \dots, \alpha_n$ are linearly independent over $\mathbb{Q}$ we see that $\lambda_1 = \cdots = \lambda_n = 0$. Thus by induction, $|q_1|, \dots, |q_{n-1}|$ are bounded.

It remains to consider the possibility that $c_1 = \cdots = c_n = 0$ and $c_{n+1} \neq 0$. Then $c_{n+1} p = 0$ hence $p = 0$. In this case

$$|q_1 \cdots q_n|^{1+\delta} |\alpha_1 q_1 + \cdots + \alpha_n q_n| < 1.$$

Thus

$$|q_1 \cdots q_n|^{1+\delta} |\alpha_n| \left| \frac{\alpha_1}{\alpha_n} q_1 + \cdots + \frac{\alpha_{n-1}}{\alpha_n} q_{n-1} + q_n \right| < 1$$

Put $\alpha_i' = \frac{\alpha_i}{\alpha_n}$ for $i = 1, \dots, n$. Our result now follows by induction since then

$$|q_1 \cdots q_{n-1}|^{1+\delta/2} \|\alpha_1' q_1 + \cdots + \alpha_{n-1} q_{n-1}\| < 1$$

for $q = \max_i |q_i|$ sufficiently large. $\square$

Theorem 19. *Let α_{ij} be real algebraic numbers for $i = 1, \dots, n$ and $j = 1, \dots, m$ and suppose that $1, \alpha_{i1}, \dots, \alpha_{im}$ are linearly independent over $\mathbb{Q}$ for $i = 1, \dots, n$. Let $\delta > 0$. Then there are only finitely many m-tuples $(q_1, \dots, q_m)$ of non-zero integers for which*

$$|q_1 \cdots q_m|^{1+\delta} \prod_{i=1}^{n} \|\alpha_{i1}q_1 + \cdots + \alpha_{im}q_m\| < 1.$$

Note: We have been looking at the "height" of a rational number to be q but this does not make sense since we are not taking into account the numerator. So a better height would be $H(p/q) = \max(|p|, |q|)$, $(p, q) = 1$.

Definition:

For any algebraic number α we define the height of α, denoted $H(\alpha)$, to be the maximum of the absolute values of the coefficients of the minimal polynomial for α over $\mathbb{Q}$. Here we are taking the minimal polynomial in $\mathbb{Z}[x]$ and of content 1.

Note: This is the naive height.

Instead of approximating an algebraic number by rationals we can approximate it by other algebraic numbers.

Theorem 20. *Let n be a positive integer and $\epsilon > 0$. If α is an algebraic number of degree greater than n then there are only finitely many algebraic numbers β of degree at most n for which*

$$|\alpha - \beta| < H(\beta)^{-n-1-\epsilon}.$$

Proof:

Let m be the degree of β over $\mathbb{Q}$. Put $\alpha_j = \alpha^j$ for $j = 1, \dots, m$. Certainly $1, \alpha_1, \dots, \alpha_m$ are linearly independent over $\mathbb{Q}$ since α is of degree $n > m$. Let $P(x)$ be the minimal polynomial for β. Then we claim that

$$|P(\alpha)| \leq H(\beta)C|\alpha - \beta|,$$

where C is a positive number which depends on α and m only. To see this let $P(x) = a_m x^m + \cdots + a_1 x + a_0 = a_m(x - \beta_1) \cdots (x - \beta_m)$ and we may suppose that $\beta_1 = \beta$. Then

$$\begin{aligned}
|P(\alpha)| &= |a_m||\alpha - \beta_1| \cdots |\alpha - \beta_m| \\
&\leq |\alpha - \beta| \cdot |a_m| \prod_{i=2}^{m} \max\{2|\alpha|, 2|\beta_i|\} \\
&\leq |\alpha - \beta| 2^{m-1} (\max\{1, |\alpha|\})^{m-1} |a_m| \prod_{i=2}^{m} \max\{1, |\beta_i|\} \\
&\leq |\alpha - \beta| \cdot C_1 \cdot C_2 H(\beta)
\end{aligned}$$

where C_1 depends only on m and α and C_2 depends on m only, and C_1 and C_2 are both positive. By Corollary 16

$$|P(\alpha)| > \frac{C_3(m, \delta)}{H(\beta)^{m+\delta}}$$

The result now follows on noting $m \leq n$. $\square$.

There is another extension of Roth's Theorem due to Leveque where we approximate by algebraic numbers from a fixed field. Let $[K:\mathbb{Q}]=n$. Let α be an algebraic number of degree $d\geq 2$ over K. Let $\epsilon>0$, then there are only finitely many algebraic numbers β for which $|\alpha-\beta<H(\beta)^{-2-\epsilon}$.

Applications to Diophantine Equations

Let $F(x,y)\in\mathbb{Z}[x,y]$ be a binary form of degree n, so

$$F(x,y)=a_nx^n+a_{n-1}x^{n-1}y+\cdots+a_1xy^{n-1}+a_0y^n.$$

Put $f(x)=F(x,1)$ and suppose

$$f(x)=a_n(x-\alpha_1)\cdots(x-\alpha_n)$$

with $\alpha_1,\ldots,\alpha_n$ distinct. For example, let us suppose that $f(x)$ is irreducible and $n\geq 3$. Let m be a non-zero integer. The equation

$$F(x,y)=m$$

in integers x and y is known as a Thue equation.

Example:

$x^3-2y^3=6,\quad (x,y)=(2,1)$ is a solution, in fact, the only solution.

The fact that there are only finitely many solutions is a consequence of Roth's Theorem.

Plainly there are only finitely many solutions with $y=0$, so suppose (x,y) is a solution with $y\neq 0$. Then

$$|m|=|F(x,y)|=|a_n||x-\alpha_1y|\cdots|x-\alpha_ny|$$

Thus

$$\frac{|m|}{|y|^n}=|a_n|\left|\alpha_1-\frac{x}{y}\right|\cdots\left|\alpha_n-\frac{x}{y}\right|.$$

Suppose, without loss of generality, that $\frac{x}{y}$ is closest to α_1 among the roots of $f(x)$. Then

$$\left|\alpha_1-\frac{x}{y}\right|=\frac{|m|}{|y|^n}\cdot\frac{1}{|a_n|}\cdot\frac{1}{\left|\alpha_2-\frac{x}{y}\right|\cdots\left|\alpha_n-\frac{x}{y}\right|}$$

Note that

$$\left|\alpha_2-\frac{x}{y}\right|\geq|\alpha_2-\alpha_1|-\left|\alpha_1-\frac{x}{y}\right|\geq\frac{|\alpha_2-\alpha_1|}{2}$$

hence that

$$\left|\alpha_2-\frac{x}{y}\right|\cdots\left|\alpha_n-\frac{x}{y}\right|\geq\prod_{i=2}^{n}\frac{|\alpha_i-\alpha_1|}{2}$$

Observe that the roots $\alpha 1,\ldots,\alpha_n$ are distinct since f is irreducible over $\mathbb{Q}$. Therefore

$$\left|\alpha_1-\frac{x}{y}\right|\leq\frac{|m|}{|y|^n}\cdot\frac{1}{|a_n|}\prod_{i=2}^{n}\frac{2}{|\alpha_i-\alpha_1|}<\frac{C(m,f)}{|y|^n}$$

where C is a positive number which depends on m and f. by Roth's Theorem with $\epsilon=\frac{1}{2}$ say there is a positive number $C_1(\alpha_1)$, which depends on α_1, such that

$$\left|\alpha_1-\frac{x}{y}\right|>\frac{C_1(\alpha_1)}{|y|^{2+\frac{1}{2}}}$$

therefore

$$\frac{C(m,f)}{C_1(\alpha_1)} > |y|^{n-(2+\frac{1}{2})}.$$

But $n \geq 3$ and so $|y|$ is bounded thus $|x|$ is also bounded.

Note: Since Roth's theorem is ineffective we can't determine $C_1(\alpha_1)$ from the proof and so we can't use the theorem to find all solutions of a Thue equation.

In 1909, Thue proved that "Thue equations" have only finitely many solutions and he deduced his result from the following:

Let α be an algebraic number of degree n with $n \geq 3$. Let $\epsilon > 0$. There exist only finitely many rationals $\frac{p}{q}$ with $q > 0$ for which

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^{\frac{n}{2}+1+\epsilon}}.$$

In 1921, Siegel replaced $\frac{n}{2}+1$ by $\min_{s\in\mathbb{Z}^+} s + \frac{n}{s+1}$ hence we can take $2\sqrt{n}$.

In 1947, Dyson and Gelfond independently showed that one can replace $2\sqrt{n}$ by $\sqrt{2n}$.

In 1955, Roth proved that we can take $2+\epsilon$.

Question: How can one overcome the "ineffectiveness" in Roth's theorem? This is still open.

There are three approaches to the problem that have been fruitful. The first is due to Thue and it depends on examining Pade' approximates to hypergeometric functions and it works for some n-th roots of rationals.

In 1964, Baker proved that

$$\left|\alpha - \frac{p}{q}\right| > \frac{c}{q^k}, \tag{14}$$

where $\alpha = \sqrt[3]{2}$, $c = 10^{-6}$ and $k = 2.955$. Baker used the fact that $128 = 2^7$ is close to $125 = 5^3$. He also proved (14) with $\alpha = \sqrt[3]{19}$, $k = 10^{-9}$ and $k = 2.56$.

Chudnovsky 1983 refined Baker's work:

$$\left|\sqrt[3]{2} - \frac{p}{q}\right| > \frac{1}{q^{2.43}}$$

for $q > C$ with C effectively computable.

Easton 1986 proved that

$$\left|\sqrt[3]{2} - \frac{p}{q}\right| > \frac{10^{-6}}{q^{2.8}}$$

Bennett showed for $q > 3$

$$\left|\sqrt[3]{2} - \frac{p}{q}\right| > \frac{1}{4q^{2.5}}$$

This approach works for some n-th roots!

Bombieri 1982 and Bombieri + Mueller 1983 showed that in some cases one can make Thue's original argument effective.

The only general effective improvement on the Liouville estimates is due to Baker 1968 and it follows from his work on estimates for linear forms in the logarithms of algebraic numbers. In 1986, Baker and Stewart proved the following by this method:

Let a be a positive integer which is not a perfect cube. Let ϵ be the fundamental unit in the ring of algebraic integers of $\mathbb{Q}(\sqrt[3]{a})$. Here the fundamental unit is the smallest unit larger than 1 in the ring. Then for all rationals $\frac{p}{q}$ with $q > 0$,

$$\left|\sqrt[3]{a} - \frac{p}{q}\right| > \frac{C}{q^k},$$

where $C = \frac{1}{3ac_1}$ and $k = 3 - \frac{1}{c_2}$ where $c_1 = e^{(50 \log\log\epsilon)^2}$ and $c_2 = 10^{12} \log\epsilon$. For example, if $a = 14$ then $C = 10^{-11000}$ and $k = 2.9999999999998$.

A Thue equation is a special case of a norm form equation. Let K be an algebraic number field of degree d over $\mathbb{Q}$. Let $\phi_q, \ldots, \phi_d$ be the isomorphic embeddings of K into $\mathbb{C}$. For any element α in K we denote $\phi_i(\alpha)$ by $\alpha^{(i)}$. The norm of α, Norm(α) is $\alpha^{(1)} \cdots \alpha^{(d)}$.

Let $\alpha_1, \ldots, \alpha_n \in K$. Consider the linear form

$$M(\underline{\underline{x}}) = \alpha_1^{(1)} X_1 + \cdots + \alpha_n^{(1)} X_n.$$

Then

$$N(M(\underline{\underline{x}})) = \prod_{i=1}^{d} (\alpha_1^{(i)} X_1 + \cdots + \alpha_n^{(i)} X_n)$$

is the norm form associatd to M and K.

For example, if $K = \mathbb{Q}(\sqrt[4]{2})$ and

$$M(x_1, x_2) = x_1 - \sqrt[4]{2} x_2$$

then

$$N(M(x_1, x_2)) = x_1^4 - 2x_2^4.$$

If

$$M(x_1, x_2, x_3) = x_1 + \sqrt[4]{2} x_2 + \sqrt[4]{4} x_3$$

then

$$N(M(x_1, x_2, x_3)) = x_1^4 - 2x_2^4 + 4x_3^4 - 4x_1^2 x_3^2 + 8x_1 x_2^2 x_3.$$

Let $m \in \mathbb{Z}\backslash\{0\}$. $N(M(\underline{\underline{x}})) = m$ is said to be a norm form equation in integers $x_1, \ldots, x_n$. Observe that if $\alpha_1, \ldots, \alpha_n$ are algebraic integers then $N(M(\underline{\underline{x}}))$ is a homogeneous polynomial of degree d in the variables $x_1, \ldots, x_n$ with integer coefficients.

We wish to study solutions in the integers of the equation $N(M(\underline{\underline{x}})) = m$. Put $\mathcal{M} = \{M(x_1, \ldots, x_n) \,|\, (x_1, \ldots, x_n) \in \mathbb{Z}^n\}$. Note that $\mathcal{M}$ is a $\mathbb{Z}$-module since it is as additive abelian group under $+$ and for all $r, s \in \mathbb{Z}$ and $m, n \in \mathcal{M}$ we have $rm \in \mathcal{M}$ and

i) $r(m + n) = rm + rn$
ii) $(r + s)m = rm + sm$
iii) $r \cdot (sm) = (r \cdot s) \cdot m$
iv) $1 \cdot m = m$

Therefore $N(M(\underline{\underline{x}})) = m$ can be rewritten as $N(\mu) = m$ for $\mu \in \mathcal{M}$

We'll now discuss finitely generated $\mathbb{Z}$-module in K. Our first step will be to show that these objects have a basis. That is a set of generators $\{\alpha_1, \ldots, \alpha_t\}$ which is $\mathbb{Z}$-linearly indpendent. In particular if $c_1\alpha_1 + \cdots + c_n\alpha_n = 0$ with $c_1, \ldots, c_n \in \mathbb{Z}$ then $c_1 = \cdots = c_n = 0$.

Theorem 21. *If an abelian group has no non-zero torsion element and it is finitely generated then it has a basis.*

Proof:

Let $\alpha_1, \dots, \alpha_s$ be a system of generators for the group. Denote the abelian group by M. Then $M = \{\alpha_1, \dots, \alpha_s\}$. Observe that if $k \in Z$ then $M = \{\alpha_1 + k\alpha_2, \alpha_2, \dots, \alpha_s\}$. To see this put $\alpha_1' = \alpha_1 + k\alpha_2$. Suppose that $\alpha \in M$ and

$$\alpha = c_1\alpha_1 + \cdots + c_s\alpha_s \quad \text{with } c_i \in \mathbb{Z}$$

then

$$\alpha = c_1\alpha_1' + (c_2 - kc_1)\alpha_2 + c_3\alpha_3 + \cdots + c_s\alpha_s.$$

Thus

$$M = \{\alpha_1, \dots, \alpha_s\} = \{\alpha_1', \alpha_2, \dots, \alpha_s\}.$$

If $\alpha_1, \dots, \alpha_s$ are $\mathbb{Z}$-linearly independent then $\{\alpha_1, \dots, \alpha_s\}$ form a basis as required. If not then there exist integers $c_1, \dots, c_s$ not all zero for which

$$c_1\alpha_1 + \cdots + c_s\alpha_s = 0.$$

We may suppose that $\{\alpha_1, \dots, \alpha_s\}$ are chosen such that the smallest non-zero of the c_i's is minimal over generators $\{\alpha_1, \dots, \alpha_s\}$. Suppose, without loss of generality, that c_1 has the smallest non-zero absolute value among the c_i's. Then $c_1 \mid c_i$ for $i = 1, \dots, s$ for if not we may suppose, without loss of generality, that $c_1 \nmid c_2$. Then $c_2 = qc_1 + r$ with $0 < r < |c_1|$. We now replace α_1 by $\alpha_1' = \alpha_1 + q\alpha_2$ and then by our earlier remarks

$$M = \{\alpha_1, \dots, \alpha_s\} = \{\alpha_1', \alpha_2, \dots, \alpha_s\}$$

and since $c_1\alpha_1 + \cdots + c_s\alpha_s = 0$ we have

$$c_1\alpha_1' + r\alpha_2 + c_3\alpha_3 + \cdots + c_s\alpha_s = 0.$$

This contradicts the minimal choice of $\alpha_1, \dots, \alpha_s$. Thus $c_1 \mid c_i$ for $i = 1, \dots, s$. In particular,

$$\alpha_1 + \frac{c_2}{c_1}\alpha_2 + \cdots + \frac{c_s}{c_1}\alpha_s = 0$$

hence $\alpha_1 = b_2\alpha_2 + \cdots + b_s\alpha_s$ with $b_i \in \mathbb{Z}$ for $i = 2, \dots, s$. In particular, $M = \{\alpha_2, \dots, \alpha_s\}$. We repeat the argument with $\alpha_2, \dots, \alpha_s$. We continue until we get a $\mathbb{Z}$-linearly independent set. Since M is finitely generated the process terminates after finitely many steps. □

Return to the case when $[K : \mathbb{Q}] < \infty$. If $\mathcal{M}$ is a finitely generated $\mathbb{Z}$-module in K then $\mathcal{M}$ has a basis. Since the characteristic of K is zero there are no non-trivial torsion elements. Since $[K : \mathbb{Q}] < \infty$, $\mathcal{M}$ will be finitely generated.

The number of generators in a basis for such a $\mathbb{Z}$-module is said to be the rank. The rank is well defined since two bases for such a $\mathbb{Z}$-module have the same number of elements. In fact, if the rank is m then there is an invertible $m \times m$ matrix with integer entries which transforms one base to the other.

We say that $\mathcal{M}$ is a full module or a module of full rank if the rank of $\mathcal{M}$ is equal to $[K : \mathbb{Q}]$.

Theorem 22. *The norm form $N(\alpha_1 x_1 + \cdots + \alpha_n x_n)$ is irreducible over $\mathbb{Q}$ if and only if $K = \mathbb{Q}(\alpha_2/\alpha_1, \ldots, \alpha_n/\alpha_1)$.*

Proof:

Since $N(\alpha_1 x_1 + \cdots + \alpha_n x_n) = N(\alpha_1) N\left(x_1 + \frac{\alpha_2}{\alpha_1} x_2 + \cdots + \frac{\alpha_n}{\alpha_1} x_n\right)$ we may suppose without loss of generality that $\alpha_1 = 1$.

Put $L = \mathbb{Q}(\alpha_1, \ldots, \alpha_n)$. Then

$$\begin{aligned} N(x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) &= N_K((x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) \\ &= N_L(x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n)^{[K:L]} \end{aligned}$$

Thus if $N(x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n)$ is irreducible then $[K : L] = 1$ so $K = L$.

On the other hand, if $K = L$ then $K = \mathbb{Q}(\beta)$ for some $\beta \in K$ so $\beta \in L$. But then there exist rationals $c_2, \ldots, c_n$ for which $\beta = c_2\alpha_2 + \cdots + c_n\alpha_n$. Let $[K : \mathbb{Q}] = d$. Since β has degree d over $\mathbb{Q}$ then $N(x + \beta y)$ is irreducible over $\mathbb{Q}$. Therefore

$$N(x + \beta y) = N(x + c_2\alpha_2 y + \cdots + c_n\alpha_n y)$$

is irreducible hence $N(x + \alpha_2 x_2 + \cdots + \alpha_n x_n)$ is also irreducible over $\mathbb{Q}$. □

An irreducible binary form with integer coefficients $F(x, y)$ over $\mathbb{Q}$ can be written as a norm form $N(\alpha_1 x + \alpha_2 y)$. Further the associated Thue equation has only finitely many solutions if the degree of the form is ≥ 3.

On the other hand, for forms in more than two variables the two notions are different. In particular, there are irreducible forms which are not norm forms. In fact, this is the generic situation.

Definition:

A full module in K which contains 1 and is a ring is said to be an order of K.

For example the algebraic integers of K form an order of K.

Notice that if Θ is an order of K and $\mu \in \Theta$ then $\mu^h \in \Theta$ for $h = 1, 2, \ldots$. For each $\mathbb{Z}$-module $\mathcal{M}$ in K there is a non-zero integer c such that cm is an algebraic integer for all $m \in \mathcal{M}$. Thus $c\mu^h$ is an algebraic integer for $h = 1, 2, \ldots$. Therefore μ is an algebraic integer. Hence every order of K is a subset of the order of algebraic integers of K. As a consequence we call the ring of algebraic integers of K the maximal order of K.

The units ϵ in an order $\mathcal{O}$ in a field K are the elements for which there exist ε_1 in $\mathcal{O}$ with

$$\epsilon\epsilon_1 = 1.$$

Note that $1 = N(1) = N(\epsilon\epsilon_1) = N(\epsilon)N(\epsilon_1)$.

Since $\mathcal{O}$ is an order ϵ and ϵ_1 are algebraic integers and thus $N(\epsilon)$, $N(\epsilon_1)$ are rational integers. Thus $N(\epsilon) = \pm 1$.

Suppose ϵ is in $\mathcal{O}$ and $N(\epsilon) = \pm 1$. Then ϵ is an algebraic integer and so is the root of a polynomial with integer coefficients of the form

$$x^d + \cdots + a_1 x + N(\epsilon) = 0.$$

Therefore

$$\epsilon^{d-1} + \cdots + a_1 = \frac{-N(\epsilon)}{\epsilon} = \frac{\pm 1}{\epsilon}.$$

Thus $\frac{1}{\epsilon}$ is in $\mathcal{O}$ and thus ϵ is a unit.

Proposition 6. *Let $\mathcal{O}$ be an order of a number field K. Then the group of units is infinite except when $K = \mathbb{Q}$ or when K is an imaginary quadratic extension of $\mathbb{Q}$.*

Proof: This is a consequence of Dirichlet's unit theorem extended to orders. (Borevich + Shafarevich).

Proposition 7. *Let $\mathcal{M}$ be a finitely generated abelian group with no non-zero element of finite order. All subgroups N of $\mathcal{M}$ have a finite number of generators and so possess a basis. Further if $\omega_1, \ldots, \omega_m$ is a basis for $\mathcal{M}$ then there is a basis $\eta_1, \ldots, \eta_k$ of N such that*

$$\begin{aligned}
\eta_1 &= c_{11}\omega_1 + c_{12}\omega_2 + \cdots + c_{1m}\omega_m \\
\eta_2 &= \qquad\quad c_{22}\omega_2 + \cdots + c_{2m}\omega_m \\
&\vdots \\
\eta_k &= \qquad\quad c_{kk}\omega_k + \cdots + c_{km}\omega_m
\end{aligned}$$

where c_{ij}'s are in $\mathbb{Z}$, $c_{ij} > 0$ and $k \leq m$.

Proof: Similar to the proof of Theorem 12.

Thus a submodule of a module of K is a finitely generated $\mathbb{Z}$-module.

Let $\mathcal{M}$ be a finitely generated $\mathbb{Z}$-module in a number field K. Suppose that $\mathcal{M}$ is a full module. Define $\mathcal{O}_\mathcal{M}$, the stabilizer of $\mathcal{M}$ to be the set of λ in K for which $\lambda\mathcal{M} \subseteq \mathcal{M}$. In particular $\lambda\mu \in \mathcal{M}$ for each $\mu \in \mathcal{M}$. ($\mathcal{O}_\mathcal{M}$ is also called the coefficient ring of $\mathcal{M}$)

Proposition 8. *If $\mathcal{M}$ is a full module in K then $\mathcal{O}_\mathcal{M}$ is an order of K.*

Proof:

First note that $\mathcal{O}_\mathcal{M}$ is a ring since it is a subring of K. Note that if $a, b \in \mathcal{O}_\mathcal{M}$ then $a-b$ and ab are in $\mathcal{O}_\mathcal{M}$. Plainly $\mathcal{O}_\mathcal{M}$ is non-empty since $1 \in \mathcal{O}_\mathcal{M}$. Next we observe that $\mathcal{O}_\mathcal{M}$ is a $\mathbb{Z}$-module since it is an additive abelian group under addition and properties i) - iv) hold. To prove that $\mathcal{O}_\mathcal{M}$ is an order, it remains to show that $\mathcal{O}_\mathcal{M}$ is a full $\mathbb{Z}$-module in K.

Let $\gamma \in \mathcal{M}$ with $\gamma \neq 0$. Then $\alpha\gamma \in \mathcal{M}$ for all $\alpha \in \mathcal{O}_\mathcal{M}$. Thus $\gamma\mathcal{O}_\mathcal{M} \subseteq \mathcal{M}$. Hence $\gamma\mathcal{O}_\mathcal{M}$ is a subgroup of $\mathcal{M}$ hence a submodule of $\mathcal{M}$. Therefore it has a basis of the form given by Proposition 7. Note that $\mathcal{O}_\mathcal{M} = \gamma^{-1}(\gamma\mathcal{O}_\mathcal{M})$. It remains to show that $\gamma\mathcal{O}_\mathcal{M}$ is a full module.

Let $\{\alpha_1, \ldots, \alpha_d\}$ be a basis for K over $\mathbb{Q}$. Recall that $\mathcal{M}$ is full and so $\mathcal{M} = \{\mu_1, \ldots, \mu_d\}$. Let $\alpha \in K$. We can write

$$\alpha\mu_i = \sum_{j=1}^{d} a_{ij}\mu_i \quad \text{with } a_{ij} \in \mathbb{Q}$$

for $i = 1, \ldots, d$ and $j = 1, \ldots, d$. By clearing denominators we see that there is an integer c_i such that

$$c_i\alpha\mu_i = \sum_{j=1}^{d} (c_i a_{ij})\mu_i \quad \text{with } c_i a_{ij} \in \mathbb{Z}.$$

Then take $c = c_1 \cdots c_d$ and we see that $c\alpha \in \mathcal{O}_\mathcal{M}$. In particular there exist non-zero integers $c^{(1)}, \ldots, c^{(d)}$ such that $c^{(i)}\alpha_i \in \mathcal{O}_\mathcal{M}$ for $i = 1, \ldots, d$. Thus $\mathcal{O}_\mathcal{M}$ is full and hence is an order. $\square$

Let $\mathcal{M}$ be a full module in K. Let $\mathcal{U}_{\mathcal{M}}$ be the group of units ϵ in $\mathcal{O}_{\mathcal{M}}$ for which $N(\epsilon) = 1$. $\mathcal{U}_{\mathcal{M}}$ is a subgroup of the groups of units in $\mathcal{O}_{\mathcal{M}}$ of index 1 or 2. Thus $\mathcal{U}_{\mathcal{M}}$ is infinite except if $K = \mathbb{Q}$ or K is an imaginary quadratic extension by Proposition 6 and 8. Now notice that if $\mu \in \mathcal{M}$ is a solution of

$$N(\mu) = a \tag{15}$$

for some integer a and ϵ is in $\mathcal{U}_{\mathcal{M}}$ then

$$N(\epsilon\mu) = N(\epsilon)N(\mu) = a.$$

Thus if $\mathcal{M}$ is a full module in K and K is not exceptional then whenever (15) has one solution $\mu \in \mathcal{M}$ it has infinitely many solutions in $\mathcal{M}$. But this is not the only situation where we can have infinitely many solutions to (15).

Suppose that L is a subfield of K and that L is not exceptional and that $\mathcal{M}_0$ is a module of K which is proportional to a full module of L. Say $\mathcal{M}_0 = \gamma\mathcal{L}$ where $\mathcal{L}$ is a full module of L and γ is a non-zero element of K. Since L is not exceptional there exists an integer b for which there are infinitely many $\lambda \in \mathcal{L}$ such that

$$N_L(\lambda) = b.$$

But then notice

$$N(\gamma\lambda) = N(\gamma)N(\lambda) = N(\gamma) \cdot N_L(\lambda)^{[K:L]} = N(\gamma) \cdot b^{[K:L]}.$$

So, take $a = N(\gamma)b^{[K:L]}$.

Definition:

A module $\mathcal{M}$ in K is said to be degenerate if it contains a submodule proportional to a full module in a subfield of K which is neither $\mathbb{Q}$ nor an imaginary quadratic field.

Note that if $\mathcal{M}$ is degenerate then there are integers a for which the equation $N(\mu) = a$ has infinitely many solutions $\mu \in \mathcal{M}$.

Theorem 23. *(Schmidt's Norm Form Theorem, 1972)*
Let $\mathcal{M}$ be a module of K. There exists an $a \in \mathbb{Q}$ for which $N(\mu) = a$ has infinitely many solutions $\mu \in \mathcal{M}$ if and only if $\mathcal{M}$ is degenerate.

Proof:

We already proved $\Leftarrow$. To prove $\Rightarrow$ apply Schmidt's Subspace Theorem.