# Number Theory

Martina Tscheckl

January 4, 2016

Please send feedback to martina@tscheckl.eu.

## Contents

# Organizatorial stuff

Dates (in TUGrazOnline):

| | | | |
|---|---|---|---|
| Mon | 14:15–15:45 | C208 | Exercises (starting 19.10. first exercise class) |
| Tue | 14:15–15:45 | C307 | Lecture (starting 20.10. first (real) lecture) |
| Wed | 08:15–09:45 | C208 | Lecture |

From now until 15.12. lectures by Martin Widmer. Then C. Frei.
End: oral exams
Exercises: Find details on website of the instructor Dijana Kreso. math.tugraz.at/~kreso

# 0 Basics

$$\mathbb{N} = \{1, 2, \dots\} \tag{1}$$
$$\mathbb{N}_0 = \mathbb{N} \cup \{0\} \tag{2}$$
$$\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\} \tag{3}$$

## 0.1 Divisibility

**Definition 0.1.1.** *Let $a, b \in \mathbb{Z}$. $a$ divides $b$ (written $a \mid b$) if $\exists q \in \mathbb{Z} : b = qa$. Some properties: Let $a, b, c \in \mathbb{Z}$. Then the following statements hold:*

$$a \mid b \Rightarrow ac \mid bc \tag{4}$$
$$a \mid b \wedge b \mid c \Rightarrow a \mid c \tag{5}$$
$$a \mid b \wedge b \mid a \Leftrightarrow a = b \tag{6}$$
$$a \mid b \wedge a \mid c \Rightarrow a \mid (b + c) \tag{7}$$

**Definition 0.1.2** (Remainder)**.** *Let $a \in \mathbb{Z}$, $b \in \mathbb{N}$. Then there are unique $q, r \in \mathbb{Z}$ such that*

$$a = qb + r \text{ and } 0 \le r < b.$$

**Remark.** 1. $b \mid a \Leftrightarrow r = 0$

2. $q = \lfloor \frac{a}{b} \rfloor$ *(largest integer $\le \frac{q}{b}$)*

3. *we will somtimes write: $a \mod b := c$*

**Definition 0.1.3.** *Let $a_1, a_2, \dots, a_n, d \in \mathbb{Z}$. $d$ is a greatest common divisor (gcd) of $a_1, \dots, a_n$ if $d \mid a_i\ \forall 1 \le i \le n$, and for every $e \in \mathbb{Z}$ with $e \mid a_i\ \forall 1 \le i \le n$, $e \mid d$.*

**Remark.** 1. *a* gcd *of $a_1, \dots, a_n$ is unique up to sign*

2. *we write $d = \gcd(a_1, \dots, a_n)$ if $d$ is a* gcd *of $a_1, \dots, a_n$*

3. *for $a_1, \dots, a_n \in \mathbb{Z}$, a* gcd *exists and can be written as a linear combination of $a_1, \dots, a_n$, i.e., $\exists x_1, \dots, x_n \in \mathbb{Z}$ such that*

$$\gcd(a_1, \dots, a_n) = x_1 a_1 + \dots + x_n a_n$$

4. $\gcd(a_1, \dots, a_n) = \gcd(\gcd(a_1, \dots, a_{n-1}), a_n)$

5. *if* $a \mid bc$ *and* $\gcd(a, b) = 1$ *then* $a \mid c$.

6. *let* $a' := \frac{a}{\gcd(a,b)}$, $b' = \frac{b}{\gcd(a,b)}$. *Then* $\gcd(a', b') = 1$

Hier verwendest du :=, sonst aber nur =, evtl. einheitlich machen für alle Definitionen?

**Algorithm 1** Compute the gcd of two integers: Euclidean algorithm

**Given:** $a, b \in \mathbb{Z}$. $|a| \geq |b|$
**Find:** $a := \gcd(a, b)$
  replace $a$ by $|a|$, $b$ by $|b|$
  **while** $b \neq 0$ **do**
    write $a = qb + r$, $0 \leq r < b$
    $a := b$
    $b := r$
  **end while**
  **return** $a$

The algorithm is correct, since $\gcd(a, b) = \gcd(b, a \mod b)$.
The algorithm terminates because $b$ decreases in each step.
The algorithm is fast: $(\mathcal{O}(\log b))$

sollte ausgebessert werden, 1. $O(logn)$ steps, 2. stimmt nur wenn $|r| \leq b/2$

The Euclidean algorithm also allows us to find $x, y$ such that $\gcd(a, b) = ax + by$ by doing all computations backwards.

**Example.** $\gcd(56, 22) = ?$

$$\begin{aligned}
a &= 56, b = 22 \\
56 &= 2 \cdot 22 + 12 \\
a &= 22, b = 12 \neq 0 \\
22 &= 1 \cdot 12 + 10 \\
a &= 12, b = 10 \neq 0 \\
12 &= 1 \cdot 10 + 2 \\
a &= 10, b = 2 \neq 0 \\
10 &= 5 \cdot 2 + 0 \\
a &= 2, b = 0 && \Rightarrow \gcd(56, 22) = 2
\end{aligned}$$

*Doing the computations backwards:*

$$\begin{aligned}
&2 = 12 - 10 = 12 - (22 - 12) = -22 + 2 \cdot 12 = -22 + 2(56 - 2 \cdot 22) = 2 \cdot 56 - 5 \cdot 22 \\
&x = 2, y = -5
\end{aligned}$$

**Application** (linear diophantine equations)**.** *Let* $a, b, c \in \mathbb{Z}$, $a, b, c \neq 0$. *Find all* $(x, y) \in \mathbb{Z}^2$ *which satisfy*

$$ax + by = c. \tag{8}$$

**Existence of solution** let $d = \gcd(a, b)$.

$$(d \mid a \Rightarrow d \mid xa) \wedge (d \mid b \Rightarrow d \mid yb)$$

$$\Rightarrow d \mid xa + yb = c$$
$$\Rightarrow eq.\ (8)$$

can have solutions only if $d \mid c$.

**Solution in case** $d = 1$ Let $x_0, y_0 \in \mathbb{Z}$ such that $ax_0 + by_0 = 1$ using the Euclidean algorithm. Then from $acx_0 + bcy_0 = c$ the solution $(cx_0, cy_0)$ of (eq. (8)) follows: for all $n \in \mathbb{Z} : (x, y) := (cx_0 + nb, cy_0 + na)$ is a solution.

Indeed,

$$ax + by = acx_0 + anb + bcy_0 - bna = c \qquad \checkmark$$

These $(x, y)$ are all solutions: let $(x, y)$ be a solution. Then

$$\begin{aligned} ax + by &= c \\ acx_0 + bcy_0 &= c \\ \Rightarrow a(x - cx_0) &= b(cy_0 - y) \end{aligned}$$

$$\gcd(a, b) = 1 \ \Rightarrow\ b \mid x - cx_0 \ \Rightarrow\ x = cx_0 + nb, n \in \mathbb{Z}$$
$$\Rightarrow\ a \mid cy_0 - y \ \Rightarrow\ y = cy_0 + ma, m \in \mathbb{Z}$$
$$c = ax + by = acx_0 + anb + bcy_0 + bma$$
$$= c + (n + m)ab \Rightarrow (n + m)ab = 0 \Rightarrow m = -n$$

**Solutions in the general case** Assume $d = \gcd(a, b)$ and $d \mid c$, let

$$a' = \frac{a}{d} \qquad b' = \frac{b}{d} \qquad c' := \frac{c}{d}$$

Then $\gcd(a', b') = 1$ and the solution to (eq. (8)) is exactly the solution of $a'x + b'y = c'$.

## 0.2 Primes

1. Beistriche für bessere Lesbarkeit
2. faustregel, vor und nach "i.e." gehört eigentlich beistrich

**Definition 0.2.1.** $p \in \mathbb{N}$, $p > 1$ *is a* prime number *if the only positive divisors of* $p$ *are* 1 *and* $p$*, i.e.,* $a \in \mathbb{N}$, $a \mid p \Rightarrow a \in \{1, p\}$. $\mathbb{P} := \{primes\} \subset \mathbb{N}, \mathbb{P} = \{2, 3, 5, 7, 11, 13, \ldots\}$. *$p$ prime and $p \mid ab \Rightarrow p \mid a$ or $p \mid b$*

**Theorem 0.2.2** (Fundamental theorem of arithmetic)**.** *Every $n \in \mathbb{N}$ can be written uniquely (up to reordering) as a product of primes. i.e. there are distinct primes $p_1, \ldots, p_l$, and $\alpha_1, \ldots, \alpha_l \in \mathbb{N}$ such that $n = p_1^{\alpha_1} \ldots p_l^{\alpha_l}$*

*Sketch.*

**Existence** let $p_0 > 1$ be the smallest divisor $> 1$ of $n$. Then $p_0$ is prime. $n = p_0 n_0$, induction $\checkmark$

**Uniqueness** let $p_1 \ldots p_m = q_1 \ldots q_l = n$, $p_i, q_j$ primes. $p_1 \mid q_1 \ldots q_l \Rightarrow \exists i : p_1 \mid q_i$, both prime $\Rightarrow p_1 = q_i$, wlog: $i = 1$. $p_1 \ldots p_m = q_1 \ldots q_l$, induction $\checkmark$

□

**Theorem 0.2.3** (Euclid). *There are $\infty$-many primes.*

*Proof.* Given primes $p_1, \dots, p_n \in \mathbb{P}$. We construct one more prime

$$N := p_1 \cdot \dots \cdot p_n + 1.$$

Assume $P$ is a prime factor of $N$. If $P \in \{p_1, \dots, p_n\}$ then $P \mid N$ and $P \mid p_1 \dots p_n \Rightarrow P \mid 1$ ↯ □

**Remark** (prime factors and gcds). *Let $a_1, \dots, a_n \in \mathbb{Z}$, write*

$$a_i = \prod_{p \in \mathbb{P}} p^{\alpha_{p,i}},\ \alpha_{p,i} \in \mathbb{N}_0,$$

*almost all $a_i = 0$, then*

$$\gcd(a_1, \dots, a_n) = \prod_{p \in \mathbb{P}} p^{\min_{1 \le i \le n}\{\alpha_{p,i}\}}$$

## 0.3 Congruences

All rings are commutative with 1.

**Definition 0.3.1.** *Let $a, b \in \mathbb{Z}$, $n \in \mathbb{N}$. Then $a$* is congruent to *$b$ $\pmod n$, $a \equiv b \pmod n$, if $n \mid a - b$. We write $\bar{a} = [a]_n := \{b \in \mathbb{Z} : b \equiv a(\mod n)\}$*

**Remark.** 1. *Congruence* mod$n$ *is an* equivalence relation

2. *$\bar{0}, \bar{1}, \dots, \overline{n-1}$ is a partition of $\mathbb{Z}$.*

3. *if $a \equiv b \pmod n$, $c \equiv d \pmod n$, then $-a \equiv -b \pmod n$, $a \mathbin{\substack{+\\ \cdot \\ -}} d \pmod n$.*

**Definition 0.3.2.** $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n := \{[a]_n : a \in \mathbb{Z}\} = \{\bar{0}, \bar{1}, \dots, \overline{n-1}\}$ *residue class ring modulo $n$*

**Remark.** *$\mathbb{Z}_n$ is a ring with operation $\bar{a} \mathbin{\substack{+\\ \cdot \\ -}} \bar{b} := \overline{a \mathbin{\substack{+\\ \cdot \\ -}} b}$ (well defined due to item 3 of section 0.3) $\mathbb{Z}_n^\times = \{\bar{a} \in \mathbb{Z}_n : \exists \bar{b} \in \mathbb{Z}_n : \bar{a}\bar{b} = \bar{1}\}$ ... group of units* $\mod n$

**Lemma 0.3.3.** *Let $a \in \mathbb{Z}$. Then $\bar{a} \in \mathbb{Z}_n^\times \Leftrightarrow \gcd(a, n) = 1$.*

*Proof.*

"$\Rightarrow$" $\bar{a}\bar{b} = \bar{1} \Leftrightarrow ab \equiv 1 \pmod n \Leftrightarrow n \mid ab - 1$
$\Rightarrow$ no prime factor of $n$ divides $a$
$\Rightarrow \gcd(a, n) = 1$.

"$\Leftarrow$" $1 = \gcd(a, n) = ax + ny \Rightarrow \bar{1} = \bar{a}\bar{x}$

□

**Remark.** *The inverse of $\bar{a}$ can be computed by the Euclidean algorithm.*

**Example** (Simultaneous congruences). *Find $x \in \mathbb{Z}$ such that*

$$x \equiv 2 \pmod 3 \tag{9}$$
$$x \equiv 1 \pmod 5 \tag{10}$$
$$x \equiv 0 \pmod 7 \tag{11}$$

**Theorem 0.3.4** (Chinese remainder theorem (CRT)). *Let*

$$n_1, \dots, n_l \in \mathbb{N} \textit{ subject to } \gcd(n_i, n_j) = 1 \ \forall i \neq j$$

$$x_1, \dots, x_l \in \mathbb{Z}.$$

*Then*

$$\exists x \in \mathbb{Z} \textit{ such that } x \equiv x_i \pmod{n_i} \ \forall 1 \leq i \leq l$$

*where $x$ is unique modulo $n_1 \cdot \dots \cdot n_l$.*

*Proof.* How to compute $x$? For $i \in \{1, \dots, l\}$, let

$$N_i := \prod_{j \neq i} n_j = n_1 \dots n_{i-1} n_{n+1} \dots n_l$$

and let

$$N := \prod_i n_i = n_1 N_1 = n_2 N_2 = \dots = n_l N_l$$

because $\gcd(n_i, N_i) = 1 \Rightarrow N_i$ in invertible mod$n_i$. Let

$$m_i N_i \equiv 1 \pmod{n_i}$$

and let

$$x := N_1 m_1 x_1 + \dots + N_l m_l x_l.$$

We have $N_i m_i x_i \equiv 0 \pmod{n_j, j \neq i}$ □

**Example.**

$$n_1 = 3, \qquad n_2 = 5, \qquad n_3 = 7$$

$$x_1 = 2, \qquad x_2 = 1, \qquad x_3 = 0$$

$$N_1 = 35, \qquad N_2 = 21, \qquad N_3 = ?$$

$$\overline{m}_1 = \overline{35}^{-1} \pmod 3 = \overline{2}^{-1} \pmod 3 = \overline{2} \pmod 3 \Rightarrow m_1 = 2$$

$$\overline{m}_2 = \overline{21}^{-1} \pmod 5 = \overline{1}^{-1} \pmod 5 = \overline{1} \pmod 5 \Rightarrow m_2 = 1$$

$$\begin{aligned} x &= 35 \cdot 2 \cdot 2 + 21 \cdot 1 \cdot 1 + 0 \\ &= 140 + 21 \\ &= 161 \\ &\equiv 56 \pmod{105} \end{aligned}$$

**Example** (more abstract CRT). *Let $n_1, \dots, n_l \in \mathbb{N}$, with $\gcd(n_i, n_j) = 1$ $\forall i \neq j$. There is a ring isomorphism $f : \mathbb{Z}_{n_1 \dots n_l} \overset{\cong}{\mapsto} \mathbb{Z}_{n_1} \times \dots \times \mathbb{Z}_{n_l}$ that satisfies $f([a]_{n_1 \dots n_l}) = ([a]_{n_1}, \dots, [a]_{n_l})$ $\forall a \in \mathbb{Z}$. In particular: $\mathbb{Z}^\times_{n_1 \dots n_l} \cong \mathbb{Z}^\times_{n_1} \times \dots \times \mathbb{Z}^\times_{n_l}$ (restrict $f$ to $\mathbb{Z}^\times_{n_1 \dots n_l}$)*

## 0.4 Arithmetic functions

**Definition 0.4.1.** *$f : \mathbb{N} \to \mathbb{C}$ is an* arithmetic function*. $f$ is* multiplicative *if $\forall m, n$ it holds that* $\gcd(m,n) = 1$. *We have $f(mn) = f(m)f(n)$. $f$ is* completely multiplicative *if $\forall m, n : f(mn) = f(m)f(n)$. Let $f : \mathbb{N} \to \mathbb{C}$. Its* summatory function *is $S_f(n) := \sum_{d|n} f(d)$.*

*Proof.* If $\gcd(m,n) = 1$ and $d \mid mn$, then $\exists$ unique $d_1, d_2$ such that $d = d_1 \cdot d_2$ with $d_1 \mid m$, $d_2 \mid n$.

$$S_f(mn) = \sum_{d|mn} f(d) = \sum_{d_1|m} \sum_{d_2|n} f(d_1 d_2) = \sum_{d_1|m} f(d_1) \sum_{d_2|n} f(d_2) = S_f(m) S_f(n)$$

□

**Example.**

$$\tau(n) := S_1(n) = \sum_{d|n} 1 \qquad \ldots \textit{number of divisors of } n$$
$$\sigma(n) := S_{id}(n) = \sum_{d|n} d \qquad \ldots \textit{divisor sum of } n$$

**Definition 0.4.2.** *The function $\phi(n) := |\mathbb{Z}_n^\times|$ is called* Euler's $\phi$-function*.*

**Remark.** *1. $\phi(n) = |\{0 \le a < n : \gcd(a,n) = 1\}|$*

*2. $\phi$ is multiplicative (CRT: $\gcd(m,n) = 1$. $\mathbb{Z}_{nm}^\times \cong \mathbb{Z}_n^\times \times \mathbb{Z}_m^\times$)*

*3. $\phi(p) = p - 1$ ($\mathbb{Z}_p$ is a field)*

**Lemma 0.4.3.** $\phi(p^n) = p^n - p^{n-1}$

*Proof.*

$$\begin{aligned}\phi(p^n) &= |\{0 \le a < p^n\}| - |\{0 \le a < p^n : \gcd(a, p^n) \ne 1\}| \\ &= p^n - |\{0 \le a < p^n : p|a\}| \\ &= p^n - p^{n-1}\end{aligned}$$

□

**Proposition 0.4.4.** *If $n = p_1^{\alpha_1} \ldots p_l^{\alpha_l}$ with $p_i \ne p_j$ primes, $\alpha_i \in \mathbb{N}$. Then*

$$\phi(n) = \prod_{i=1}^{l} p_i^{\alpha_i} (1 - \frac{1}{p_i}) = n \prod_{p|n} (1 - \frac{1}{p})$$

**Theorem 0.4.5** (Euler-Fermat)**.** *Then $a^{\phi(n)} \equiv 1 \mod n$. In particular: $a^{p-1} \equiv 1 \mod p \ \forall p \nmid a$ (little Fermat).*

*Proof 1.* Lagrange's Theorem, $G = \mathbb{Z}_n^\times, \bar{a} \in G \Rightarrow \bar{a}^{|G|} = \bar{1}, |G| = \phi(n)$. □

*Proof 2.* $\prod_{x \in \mathbb{Z}_n^\times} x = \prod_{x \in \mathbb{Z}_n^\times} (\bar{a}x) = \bar{a}^{\phi(n)} \prod_{x \in \mathbb{Z}_n^\times} x \Rightarrow a^{\phi(n)} \equiv 1 \mod n$ □

**Definition 0.4.6.** *The Möbius function $\mu : \mathbb{N} \to \{-1, 0, +1\}$ is defined as*

$$\mu(n) = \begin{cases} (-1)^l & n = p_1 \ldots p_l, p_i \ne p_j, i \ne j, p_i \textit{ primes} \\ 0 & \textit{otherwise i.e. if } \exists p : p^2 \mid n \end{cases}$$

**Remark.**

1. $\mu(1) = 1,\ \mu(2) = -1,\ \mu(3) = -1,\ \mu(4) = 0,\ \mu(5) = -1,\ \mu(6) = 1,\ \dots$
2. $\mu$ *is* multiplicative

**Lemma 0.4.7.**

$$S_\mu(n) = \begin{cases} 1 & \textit{if } n = 1 \\ 0 & \textit{if } n > 0 \end{cases}$$

*Proof.*

$$S_\mu(1) = \sum_{d \mid 1} \mu(d) = \mu(1) = 1$$

By multiplicativity, it suffices to prove $S_\mu(p^n) = 0\ \forall p, n$.

$$\begin{aligned} S_\mu(p^n) &= \sum_{d \mid p^n} \mu(d) \\ &= \sum_{i=0}^{n} \mu(p^i) \\ &= \mu(1) + \mu(p) + 0 + \dots + 0 \\ &= 0 \end{aligned}$$

□

**Theorem 0.4.8** (Möbius inversion formula)**.** *Let* $f : \mathbb{N} \to \mathbb{C}$*. Then*

$$f(n) = \sum_{d \mid n} \mu(d) S_f(\frac{n}{d}).$$

*Proof.*

$$\begin{aligned} \sum_{d \mid n} \mu(d) S_f\left(\frac{n}{d}\right) &= \sum_{d \mid n} \mu(d) \sum_{e \mid \frac{n}{d}} f(e) \\ &= \sum_{e \mid n} f(e) \sum_{\substack{d \mid n \\ s.t. e \mid \frac{n}{d}}} \mu(d) \end{aligned}$$

For the next step we use $d \mid n \wedge e \mid \dfrac{n}{d} \Leftrightarrow ed \mid n \Leftrightarrow e \mid n \wedge d \mid \dfrac{n}{e}$

$$\begin{aligned} &= \sum_{e \mid n} f(e) \sum_{d \mid \frac{n}{e}} \mu(d) \\ &= f(n) \end{aligned}$$

since $\displaystyle\sum_{d \mid \frac{n}{e}} \mu(d) = \begin{cases} 1 & \frac{n}{e} = 1 \\ 0 & \text{otherwise} \end{cases}$

□

## 0.5 Structure of $\mathbb{Z}_n^\times$

$$n = p_1^{\alpha_1} \dots p_l^{\alpha_l} \text{ with } p_i \neq p_j, i \neq j, \alpha_i \in \mathbb{N} \text{ where } p_i \text{ are primes}$$

From the CRT it follows that $\mathbb{Z}_n^\times \cong \mathbb{Z}_{p_1^{\alpha_1}}^\times \times \dots \times \mathbb{Z}_{p_l^{\alpha_l}}^\times$. So we only consider prime powers $p^\alpha$, $p \in \mathbb{P}$, $\alpha \in \mathbb{N}$

### 0.5.1 Case 1: $\alpha = 1$

**Theorem 0.5.1.** *$\mathbb{Z}_p^\times$ is cyclic, i.e. $\mathbb{Z}_p^\times \cong \mathbb{Z}_{(p-1)}$*

*Proof.* Use structure theorem for finite abelian groups. If $G$ is a finite abelian group then $\exists d_1, \dots d_l \in \mathbb{N}$ such that $1 < d_1 \mid d_2 \mid d_3 \mid \cdots \mid d_l$, and $G \cong \mathbb{Z}_{d_1}^\times \times \cdots \times \mathbb{Z}_{d_l}^\times$ thus, $\mathbb{Z}_p^\times \cong \mathbb{Z}_{d_1}^\times \times \cdots \times \mathbb{Z}_{d_l}^\times$ (every element $x \in \mathbb{Z}_{d_1}^\times \times \cdots \times \mathbb{Z}_{d_l}^\times$ satisfies $d_l x = 0 \Rightarrow$ every $x \in \mathbb{Z}_p^\times$ satisfies $x^{d_l} = 1$). $x^{d_l} - 1$ is a polynomial of degree $d_l$ over the field $\mathbb{Z}_p \Rightarrow x^{d_l} - 1$ has $\le d_l$ roots $\Rightarrow p - 1 \le d_l$, but $p - 1 = d_1 \dots d_l \Rightarrow l = 1, p - 1 = d_l$ □

**Remark.** *The same proof shows: Let $F$ be a field, $G \le F^\times$, $|G| < \infty$. Then $G$ is cyclic.*

### 0.5.2 Case 2: $\alpha \ge 2$; $p \ge 3$

Denote $|x|$ as the order of $x$ in $\mathbb{Z}_{p^\alpha}^\times$; i.e. $|x| = \min\left\{l \in \mathbb{N} : x^l \equiv 1 \mod p^\alpha\right\}$

$\left|\mathbb{Z}_{p^\alpha}^\times\right| = \phi(p^\alpha) = p^{\alpha-1}(p-1)$, find $x, y \in \mathbb{Z}_{p^\alpha}^\times$ such that $|x| = p^{\alpha-1}$, $|y| = p - 1$ then $|xy| = |x||y| = p^{\alpha-1}(p-1)$, since $\gcd(|x|, |y|) = 1$

**Lemma 0.5.2.**

$$(1+p)^{p^{n-1}} \begin{cases} \equiv 1 & \mod p^n \\ \not\equiv 1 & \mod p^{n+1} \end{cases}$$

*Proof.* Proof by induction

$n = 1$ ✓

$n \to n + 1$

$$\begin{aligned} (1+p)^{p^{n-1}} &= 1 + ap^n, p \nmid a \\ (1+p)^{p^n} &= (1 + ap^n)^p \\ &= 1 + pap^n + \sum_{i=2}^{p-1} \binom{p}{i} (ap^n)^i + (ap^n)^p \end{aligned}$$

$$p^{np} \mid \bullet, \quad np \ge n + 2, \quad (\text{or } p \ge 3), \quad p^{2n+1} \mid \bullet, \quad 2n + 1 \ge n + 2$$

$$p \mid \binom{p}{i} = \frac{p!}{i!(p-i)!}, 1 \le i < p \Rightarrow (1+p)^{p^n} \equiv 1 + ap^{n+1} \pmod{p^{n+2}}, p \nmid a$$

□

$2\times$ Lemma: $x = 1 + p$ satisfies $|x| = p^{\alpha-1}$, now find $y$.

1. $\exists z \in \mathbb{Z} : |\bar{z}| = p - 1$ is $\mathbb{Z}_p^\times$
2. let $l := |E|$ is $\mathbb{Z}_{p^\alpha}^\times$
3. Then $p^\alpha \mid z^l - 1 \Rightarrow z^l \equiv 1 \mod p$
4. $\Rightarrow p - 1 \mid l$.
5. Let $y := z^{\frac{l}{p-1}}$, then $|\bar{y}| = p - 1$.

We have proven: Theorem: $\mathbb{Z}_{p^\alpha}^\times$ is cyclic, i.e. $\mathbb{Z}_{p^\alpha}^\times \cong \mathbb{Z}_{p^{\alpha-1}(p-1)}$, if $p \geq 3, \alpha \geq 1$.
$p = 2$: $\mathbb{Z}_{2^\alpha}^\times \cong \{0, \alpha = 1 \quad \mathbb{Z}_2, \alpha = 2 \quad \mathbb{Z}_2 \times \mathbb{Z}_{p^{\alpha-2}}, \alpha \geq 3\}$

**Corollary 0.5.3.** *Let $m \in \mathbb{N}$. Then $\mathbb{Z}_m^\times$ is cyclic iff $m$ has one of the following forms:*

- $m = 2$
- $m = 4$
- $m = p^\alpha, p \geq 3, \alpha \in \mathbb{N}$
- $m = 2p^\alpha, p \geq 3, \alpha \in \mathbb{N}$

In these cases a generator of $\mathbb{Z}_m^\times$ is called *a primitive root modulo m.*

# New Lecturer

Chapter 1:

1. Approximation to algebraic numbers; Wolfgang M. Schmidt, 1972 L'Ehseignement Mathématique
2. Lectures Notes in Mathematics 785; W.M.Schmidt, Springer
3. LNM 1467, W.M.S., Springer
4. For section 2 (continued fractions) he will strictly follow the lecture notes of MT421 of Professor James McKee

# 1 Diophantine Approximation

## 1.1 Dirichlet's Theorem

Let $\alpha \in \mathbb{R}$. As $\mathbb{Q}$ is dense in $\mathbb{R}$ any $\alpha \in \mathbb{R}$ can be approximated arbitrarily well, by rational numbers $p/q$ ($p \in \mathbb{Z}, q \in \mathbb{N} = \{1, 2, 3, \dots\}$).
The question is how well can we approximate $\alpha$ in terms of the denominator $q$, e.g., is it true that for every $\alpha \in \mathbb{R}$ there exists infinitely many $p/q \in \mathbb{Q}$ $q \in \mathbb{N}$) such that $|\alpha - \frac{p}{q}| < \frac{1}{q^2}$?

The answer is no!
Take $\alpha = r/s (s \in \mathbb{N})$ a rational number. Then

$$|\alpha - \frac{p}{q}| = |\frac{r}{s} - \frac{p}{q}| = |\frac{qr - ps}{sq}| \overset{\geq}{\text{provided}} \ \alpha \neq \frac{p}{q} \frac{1}{sq} > \frac{1}{q^2} \text{ provided } q > s.$$

This shows that we have only finitely many solutions $p/q \in \mathbb{Q}$ for $|\alpha - \frac{p}{q}| < \frac{1}{q^2}$.

**Theorem 1.1.1** (Dirichlet's Theorem)**.** *Suppose $\alpha, Q \in \mathbb{R}$ and $Q > 1$. Then $\exists p, p \in \mathbb{Z} s.t. 0 < q < Q$ and $|q\alpha - p| \leq \frac{1}{Q}$.*

*Proof.* for $\xi \in \mathbb{R}$ put $\{\xi\} = \xi - \lfloor \xi \rfloor$. so $0 \le \{\xi\} \le 1$. First suppose $Q \in \mathbb{Z}$. Consider the $Q+1$ numbers $0, 1, \{\alpha\}, \{2\alpha\}, \dots, \{(Q-1)\alpha\}$.
They all lie in $[0,1]$. We split it up in $Q$ subintervals:

$$[0,1] = [0, \frac{1}{Q}] \cup \left[\frac{1}{Q}, \frac{2}{Q}\right] \cup \cdots \cup \left[\frac{Q-1}{Q}, 1\right]$$

By the pigeon hole principle two of the previous numbers lie in the same subinterval. Thus $\exists r_1, r_2, s_1, s_2 \in \mathbb{Z}$ with $0 \le r_1 < r_2 \le Q-1$ such that $|(r_1\alpha - s_1) - (r_2\alpha - s_2)| \le \frac{1}{Q}$. Then with $q = r_2 - r_1$ and $p = s_2 - s_1$ we get $|q\alpha - p| \le \frac{1}{Q}$ and $0 < q < Q$. This proves the Theorem when $Q \in \mathbb{Z}$. Now suppose $Q \notin \mathbb{Z}$. We apply the previous with $Q' = \lfloor Q \rfloor + 1 > 1$. Hence, $\exists p, q \in \mathbb{Z}$ with $|q\alpha - p| \le \frac{1}{Q'}$ and $0 < q < Q'$, and so $|q\alpha - p| \le \frac{1}{Q}$ and $0 < q < Q$. □

**Corollary 1.1.2.** *Suppose $\alpha \in \mathbb{R}/\mathbb{Q}$. Then there exist infinitely many solutions $p/q \in \mathbb{Q}$ ($q \in \mathbb{N}$) of $|\alpha - \frac{p}{q}| < \frac{1}{q^2}$.*

*Proof.* Take $Q_1 > 1$. By Theorem 1.1.1 we get $(p_1, q_1) \in \mathbb{Z}^2$ with $0 < q_1 < Q$, and $|q_1\alpha - p_1| \le \frac{1}{Q_1}$. Thus $|\alpha - \frac{p_1}{q_1}| \le \frac{1}{q_1 Q_1} < \frac{1}{q_1^2}$
Next take $Q_2 = |\alpha - \frac{p_1}{q_1}|^{-1} + 1$. Then Theorem 1.1.1 again yields $\frac{p_2}{q_2} \in \mathbb{Q}$ with $|\alpha - \frac{p_2}{q_2}| < \frac{1}{q^2}$ and $|\alpha - \frac{p_2}{q_2}| \le \frac{1}{q_r Q_2} \le \frac{1}{Q_2} < |\alpha - \frac{p_1}{q_1}|$. So $\frac{p_2}{q_2}$ is a better approx then $\frac{p_1}{q_1}$. Repeating this process indefinitely proves the claim. □

**Theorem 1.1.3** (Pell-equation)**.** *Suppose $m \in \mathbb{N}$ is not a square (i.e., $m \ne n^2 \forall n \in \mathbb{Z}$).*
*Then*

$$x^2 - my^2 = 1$$

*has infinitely many solutions $(x, y) \in \mathbb{Z}^2$.*

*Proof.* Apply Corollary 1.1.2 with $\alpha = \sqrt{m}$. So $\alpha \in \mathbb{R}/\mathbb{Q}$. We get $|\alpha - \frac{p}{q}| < \frac{1}{q^2}$ and $|\alpha + \frac{p}{q}| \overset{\le}{triangleinequality} 1 + 2\alpha$. Thus

$$|p^2 - mq^2| = q^2 |\alpha - \frac{p}{q}| \cdot |\alpha + \frac{p}{q}| < 1 + 2\sqrt{m}.$$

Hence, there exists $k \in \mathbb{Z}$ with $|k| < 1 + 2\sqrt{m}$. such that $p^2 - mq^2 = k$ for infinitely many $(p, q) \in \mathbb{Z}^2$ and $p/q$ all distinct.
As $m$ is not a square we have $k \ne 0$.

Let $S$ be the set of solutions $(p, q) \in \mathbb{Z}^2$ of $p^2 - mq^2 = k$. The map $S \to (\mathbb{Z}/k\mathbb{Z}) \times (\mathbb{Z}/k\mathbb{Z})$. This map is not injective ($S = \infty$) hence, $\exists (p_1, q_1) \ne (p_2, q_2)$ both in $S$ such that $p_1 \cong p_2, q_1 \cong q_2 \pmod{k}$. (MOD)
Now we compute

$$\begin{aligned} k^2 &= (p_1^2 - mq_1^2)(p_2^2 - mq_2^2) && (12)\\ &= (p_1 + \sqrt{m}q_1)(p_2 - \sqrt{m}q_2) && (13)\\ &= (r - \sqrt{m}s)(r + \sqrt{m}s) = r^2 - ms^2 && (14)\\ \text{where } r &= p_1p_2 - mq_1q_2 && (15)\\ s &= p_1q_2 - q_1p_2 = \frac{1}{q_1q_2}(\frac{p_1}{q_1} - \frac{p_2}{q_2}) \ne 0. && (16) \end{aligned}$$

because of (MOD) $k \mid s$. Hence, $k^2 \mid s^2$. Thus $k^2 \mid r^2$. Hence $k \mid r$. Then $x = \frac{r}{k}$ and $y = \frac{s}{k}$ are both integers and

$$x^2 - my^2 = 1.$$

We have one solution but we need infinitely many! To this end we replace $m$ by $md^2$ ($d \in \mathbb{N}$). The above argument yields a solution $(x', y') \in \mathbb{Z}^2$ of $x'^2 - md^2y'^2 = 1$. Thus, $(x, y) = (x', dy')$ is a new solution of $x^2 - my^2 = 1$.
(Critical: $s \neq 0$) □

## 1.2 Continued fractions

Let $\theta \in \mathbb{R}$. Put $a_0 = \lfloor\theta\rfloor$. If $a_0 \neq \theta$ then we find $\theta_1 > 1$ such that

$$\theta = a_0 + \frac{1}{\theta_1}$$

and we put $a_1 = \lfloor\theta_1\rfloor$. If $a_1 \neq \theta_1$ then we can find $\theta_2 > 1$ such that

$$\theta_1 = a_1 + \frac{1}{\theta_2}$$

and we put $a_=\lfloor\theta_2\rfloor$. This process can be continued indefinitely, unless $a_n = \theta_n$ for some $n$. Note that $a_0$ can be zero or negative but $a_1, a_2, a_3, \ldots$ are all positive integers.
We call this process the *continued fraction process*. The $a_i$ are called *partial quotients* of $\theta$.

**Example.**

$$\theta = \frac{19}{11}$$

*Then* $a_0 = \lfloor\theta\rfloor = 1$
*Now* $\theta = \frac{19}{11} = a_0 + \frac{1}{\theta_1} = 1 + \frac{8}{11} = 1 + \frac{1}{\frac{11}{8}}$
*So* $\theta_1 = \frac{11}{8}$.
*Thus* $a_1 = \lfloor\theta_1\rfloor = 1$.
*Now*

$$\theta_1 = \frac{11}{8} = a_1 + \frac{1}{\theta_2} = 1 + \frac{3}{8} = 1 + \frac{1}{\frac{8}{3}}$$

*Thus* $\theta_2 = \frac{2}{3}$ *and* $a_2 = \lfloor\theta_2\rfloor = 2$
*and so on...*

*If the continued fraction process terminates then we have*

$$\theta = a_0 + \frac{1}{\theta_1} \tag{17}$$

$$= a_0 + \frac{1}{a_2 + \frac{1}{\theta_2}} \tag{18}$$

$$= a_0 + \frac{1}{a_2 + \frac{1}{a_3 + \frac{1}{\theta_3}}} \qquad \ldots = a_0 + \frac{1}{a_1 + \frac{1}{\ddots}} \tag{19}$$

*In this case we write* $\theta = [a_0, \dots, a_n]$.
*We use the same notation when the* $a_i$ *are any real numbers, not necessarily integers.*
*In particular*

$$\theta = [a_0, \dots, a_i, \theta_{i+1}]$$

*where* $a \le i < n$.

If the continued fraction process does not terminate then we write $\theta = [a_0, a_1, a_2, \dots]$.
Note that in this case, for every $n \ge 0$, we have

$$\theta = [a_0, \dots, a_n, \theta_{n+1}]$$

where $a_0, \dots, a_n$ are integers but $\theta_{n+1}$ is not! For $n \ge 0$ we set

$$\frac{p_n}{q_n} = [a_0, \dots, a_n]$$

where $\gcd(p_n, q_n) = 1$. We shall say that $\frac{p_n}{q_n}$ is the $n$-th convergent of $\theta$. We will prove that $\frac{p_n}{q_n} \to \theta$ as $n \to \infty$. Next we shall see that $p_n, q_n > 0$ both satisfy the same simple recurrence relation $x_n = a_n x_{n-1} + x_{n-2}$ with different starting values.

**Lemma 1.2.1.** *Let* $a_0, a_1, a_2, \dots$ *be a sequence of integers with* $a_i > 0$ $(i > 0)$. *Define* $p_n, q_n$*:*

$$p_0 = a_0 \tag{20}$$
$$q_0 = 1 \tag{21}$$
$$p_1 = a_0 a_1 + 1 \tag{22}$$
$$q_1 = a_1 \tag{23}$$
$$p_n = a_n p_{n-1} + p_{n-2} \text{ for } n \ge 2 \tag{24}$$
$$q_n = a_n q_{n-1} + q_{n-2} \text{ for } n \ge 2. \tag{25}$$

*Then:*

1. $p_n q_{n+1} - p_{n+1} q_n = (-1)^{n+1}$
2. $\gcd(p_n, q_n) = 1$
3. $p_n / q_n = [a_0, \dots, a_n]$
4. *If the* $a_i$ *are produced by the continued fraction process for* $\theta$*, then, for every* $n \ge 1$, $\frac{p_n}{q_n}$ *is the n-th convergent of* $\theta$ *and*

$$\theta = \frac{p_n \theta_{n+1} + p_{n-1}}{q_n \theta_{n+1} + q_{n-1}}$$

*Proof.* 1. We use induction on $n$. For $n = 0$ we note that

$$p_0 q_1 - p_1 q_0 = a_0 a_1 - a_0 a_1 - 1 = -1.$$

So the result holds for $n = 0$.
Now suppose result holds for $n = m - 1$.
consider case $n = m$. Using the recurrence relation, we set

$$p_m q_{m+1} - p_{m+1} q_m = p_m(a_m q_m + q_{m-1}) - q_m(a_m p_m + p_{m-1}) \tag{26}$$
$$= p_m q_{m-1} - p_{m-1} q_m = -(-1)^m = (-1)^{m+1}. \tag{27}$$

This proves claim for $n = m$.

2. Immediate from (a)

3. (c) + (d):
Remark about $\frac{p_n}{q_n}$ in (d) follows directly from (c). We prove the rest of (d), along with (c), using induction on $n$. Remember that (c) a priori does not require that the $a_i$ are produced by the continued fraction process. Consider base case $n = 1$. For (c) note that $\frac{p_1}{q_1} = a_0 + \frac{1}{a_1} = [a_0, a_1]$. For (d) we note that

$$\frac{p_1\theta_2 + p_0}{q_1\theta_2 + q_0} = \frac{(a_0a_1 + 1)\theta_2 + a_0}{a_1\theta_2 + 1} = a_0 + \frac{\theta_2}{a_1\theta_2 + 1} = a_0 + \frac{1}{a_1 + \frac{1}{\theta_2}} = \theta$$

Next suppose (c) and (d) both hold for $n = m - 1$, and consider $n = m$. Using (d) with $n = m - 1$ we get

$$[a_0, \ldots, a_m] = \frac{p_{m-1}a_m + p_{m-2}}{q_{m-1}a_m + q_{m-2}} = \frac{p_m}{q_m} \text{ by recurrence ralation.}$$

This proves (c) for $n = m$.
To prove (d) with $n = m$ we observe that

$$\theta = [a_0, \ldots, a_m, \theta m + 1] \tag{28}$$
$$= [a_0, \ldots, a_m + \frac{1}{\theta_{m+1}}] \tag{29}$$
$$\overset{=}{(d) for n} = m - 1 \frac{p_{m-1}(a_m + \frac{1}{\theta_{m+1}}) + p_{m-2}}{q_{m-1}(a_m \frac{1}{\theta_{m-1}}) + q_{m-2}} \tag{30}$$
$$\overset{=}{rec.rel} \frac{p_m + p_{m-1}(\frac{1}{\theta_{m+1}})}{q_m + q_{m-1}(\frac{1}{\theta_{m+1}})} \tag{31}$$
$$= \frac{p_m\theta_{m+1} + p_{m-1}}{q_m\theta_{m+1} + q_{m-1}} \tag{32}$$

which is (d) for $n = m$.

□

Next we deduce some properties of continued fraction convergents.

**Theorem 1.2.2.** *Let $\theta = [a_0, a_1, a_2, \ldots]$ with convergents $\frac{p_n}{q_n}$. For (a) - (d) we assume that the continued fraction proves does not terminate*

1. *For all $n \in \mathbb{N}_0$, $\theta$ lies between $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$.*

2. *For all $n \in \mathbb{N}_0 : |\theta - \frac{p_n}{q_n}| \leq \frac{1}{q_n q_{n+1}}$*

3. *For $n \geq 1$ we have $q_{n+2} \geq 2 \cdot q_n$*
4. $\frac{p_n}{q_n} \to \theta$ *as $n \to \infty$*
5. *The continued fraction process terminates if and only if $\theta$ is rational.*

*Proof.* 1. Note $\theta = [a_0, \ldots, a_n, \theta_{n+1}] = [a_0, \ldots, a_n + \frac{1}{\theta_{n+1}}$ where $0 < \frac{1}{\theta_{n+1}} < \frac{1}{a_{n+1}}$. So that $\theta$ lies between $[a_0, \ldots, a_n]$ and $[a_0, \ldots, a_n + \frac{1}{a_{n+1}}]$. But $[a_0, \ldots, a_n + \frac{1}{a_{n+1}} = [a_0, \ldots, a_{n+1}]$. This shows (a).

2. By (a) we have $|\theta - \frac{p_n}{q_n}| \leq |\frac{p_n}{q_n} - \frac{p_{n+1}}{q_{n+1}}| = |\frac{p_n q_{n+1} - p_{n+1} q_n}{q_n q_{n+1}}| \underset{=}{Lemma 1.2.1(a)} \frac{1}{q_n q_{n+1}}$

3. Follows from the fact that $a_i > 0 (i > 0)$ using Lemma 1.2.1.

4. Follows from (b) and (c)

5. Only if part is obvious.
Conversely suppose $\theta = \frac{a}{b} \in \mathbb{Q}$ but the process does *not* terminate. Taking $n$ such that $q_n > b$ yields

$$|\theta - \frac{p_n}{q_n}| \frac{a}{b} \neq \frac{p_n}{q_n} as q_n > b and \gcd(p_n, q_n) = 1 \underset{\geq}{} \frac{1}{bq_n} > \frac{1}{q_n q_{n+1}}$$

contradicting (b).

□

**Example.** *Take $\theta = \frac{16}{9}$. We have $a_0 = 1$. Then $\theta = 1 + \frac{7}{9}$ so $\theta_1 = \frac{9}{7}$ and $a_1 = 1$. From $\theta_1 = \frac{9}{7} = 1 + \frac{2}{7}$ we get $\theta_2 = \frac{7}{2}$ and $a_2 = 3$. Form $\theta_2 = \frac{7}{2} = 3 + \frac{1}{2}$ we get $\theta_3 = 2$ and $a_3 = 2$. Thus $\theta = \frac{16}{9} = [1, 1, 3, 2]$ and the convergents are $\frac{p_0}{q_0} = \frac{1}{1}$, $\frac{p_1}{q_1} = 1 + \frac{1}{1} = \frac{2}{1}$, $\frac{p_2}{q_2} = 1 + \frac{1}{1+\frac{1}{3}} = 1 + \frac{1}{\frac{4}{3}} = \frac{7}{4}$ and $\frac{p_3}{q_3} = \frac{16}{9}$.*
*Let's check some of the properties claimed.*
$p_1 q_2 + p_2 q_1 = 2 \cdot 4 - 7 \cdot 1 = 1 \checkmark, p_2 q_3 - p_3 q_2 = 7 \cdot 9 - 16 \cdot 4 = -1 \checkmark, \frac{p_2 \theta_3 + p_1}{q_2 \theta_3 + q_1} = \frac{7 \cdot 2 + 2}{4 \cdot 2 + 1} = \frac{16}{9} = \theta \checkmark$

We now show that convergents give best-possible rational approximations.

**Theorem 1.2.3.** *Let $\theta$ be an irrational real number, and let $\frac{p_n}{q_n}$ be the convergents $(n \geq 0)$ with partial quotients $a_n (n \geq 0)$.*
*Then*

1. $|\theta - \frac{p_n}{q_n}|$ *strictly decreases as $n$ increases.*
2. *the convergents give successively closer approximations to $\theta$.*
3. $\frac{1}{(a_{n+1}+2) q_n^2} < |\theta - \frac{p_n}{q_n}| < \frac{1}{a_{n+1} q_n^2} \leq \frac{1}{q_n^2}$
4. *If $p, q \in \mathbb{Z}$ with $0 < q < q_{n+1}$ then*

$$|q\theta - p| \geq |q_n \theta - p_n|$$

*Moreover, "=" only if $(p, q) = (p_n, q_n)$.*
*(In this sense convergents are best-possible approximations.)*

5. *If $(p, q) \in \mathbb{Z} \times \mathbb{N}$ and $|\theta - \frac{p}{q}| < \frac{1}{2 \cdot q^2}$ then $\frac{p}{q}$ is a convergent to $\theta$.*

*Proof.* 1. From Lemma 1.2.1(d) we have $\theta = \frac{p_n\theta_{n+1}+p_{n-1}}{q_n\theta_{n+1}+q_{n-1}}$. Using Lemma 1.2.1(a) we get

$$|q_n\theta - p_n| = |\frac{q_n p_n \theta_{n+1} + q_n p_{n-1} - p_n q_n \theta_{n+1} - p_n q_{n-1}}{q_n\theta_{n+1} + q_{n-1}}| \tag{33}$$

$$= \frac{1}{q_n\theta_{n+1} + q_{n-1}} \tag{34}$$

$$< \frac{1}{q_n + q_{n-1}} \tag{35}$$

$$= \frac{1}{(a_n + 1)q_{n-1} + q_{n-2}} \tag{36}$$

$$< \frac{1}{\theta_n q_{n-1} + q_{n-2}} \tag{37}$$

$$= |q_{n-1}\theta - p_{n-1}| \tag{38}$$

This shows (a) and (b) because the $q_n$ are increasing.

c We use $a_{n+1}q_n^2 < \theta_{n+1}q_n^2 + q_n q_{n-1} < (a_{n+1} + 2)q_n^2$ and combine it with the equation (proof part (a)),

$$|\theta - \frac{p}{q}| = \frac{1}{q_n^2\theta_{n+1} + q_n q_{n-1}}$$

d) By Lemma 1.2.1(a) we can find $\begin{pmatrix} u \\ v \end{pmatrix} in\mathbb{Z}^2$ such that

$$\begin{pmatrix} p_n & p_{n+1} \\ q_n & q_{n+1} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix}.$$

As $0 < q < q_{n+1}$ we have $u \neq 0$. If $v = 0$ then $(p, q) = u \cdot (p_n, q_n)$ and the claim is trivial. ($u = 1 \Rightarrow$ equality, $u > 1 \Rightarrow$ strictly ¿)
So let's assume $v \neq 0$. Then $u$ and $v$ cannot both be negative (as $q > 0$) nor both be positive (as $q < q_{n+1}$). So they have opposite signs.
By Theorem 1.2.2(a) also $q_n\theta - p_n$ and $q_{n+1}\theta p_{n+1}$ have opposite signs. Hence, $|q\theta - p| = |u(q_n\theta p_n) + v(q_{n+1}\theta - p_{n+1}| > |q_n\theta - p_n|$.

e) Take $n$ with $q_n \leq q < q_{n+1}$. Then

$$\begin{aligned} |\frac{p}{q} - \frac{p_n}{q_n}| &\leq |\theta - \frac{p}{q}| + |\theta - \frac{p_n}{q_n}| \\ &= \frac{|q\theta - p|}{q} + \frac{|q_n\theta - p_n|}{q_n} \\ &\overset{\leq}{(d)} (\frac{1}{q} + \frac{1}{q_n})|q\theta - p| \\ &\leq \frac{2}{q_n}\frac{1}{2q} \\ &= \frac{1}{qq_n} \end{aligned}$$

Hence, $\frac{p}{q} = \frac{p_n}{q_n}$.

$\square$

**Remark.**

- *(d) implies that if $p,q \in \mathbb{Z}$, $0 < q \le p_n$ then*

$$|\theta - \frac{p}{q}| \ge |\theta - \frac{p_n}{q_n}| \cdot \frac{p_n}{q} \\ \ge |\theta - \frac{p_n}{q_n}|$$

*with "=" only if $\frac{p}{q} = \frac{p_n}{q_n}$.*

- *We say $\alpha \in \mathbb{R} \smallsetminus \mathbb{Q}$ is* badly approximable *if*

$$\exists c > 0 \text{ such that } |\alpha - \frac{p}{q}| > \frac{c}{q^2} \forall (p,q) \in \mathbb{Z} \times \mathbb{N}$$

- *By (c) and (d) we see that $\theta = [a_0, a_1, a_2, \dots]$ is badly approximable if and only if the partial quotients $a_i$ are uniformly bounded, i.e., $\exists M > 0$ such that $a_i < M \forall i$.*

- *(c) suggests that the "worst-approximable" number is $\theta = [1,1,1,\dots]$. That's indeed the case c.f Exercise sheet 2 # 5,6 (using that $\theta = 1 + \frac{1}{1+\frac{1}{1+\dots}} = 1 + \frac{1}{\theta}$. So $\theta^2 - \theta - 1 = 0$.*
*So $\theta = \frac{1 \pm \sqrt{5}}{2}$ but $a_0 = 1$ so $\theta = \frac{1+\sqrt{5}}{2}$.*

*Counting Diophantine approximations 1:*

Let $\alpha \in \mathbb{R} \smallsetminus \mathbb{Q}$ and let $\phi : [1,\infty) \to (0,\infty)$ be decreasing. Consider the number of "$\phi$-good" approximations:

$$N_\alpha(\phi, Q) = \#\{\frac{p}{q} \in \mathbb{Q}; |\alpha - \frac{p}{q}| < \phi(q), 1 \le q \le Q\}$$

We put $S_\alpha(\phi, Q) = \{(x,y) \in \mathbb{R}^2 : |\alpha - \frac{x}{y}| < \phi(y), 1 \le y \le Q\}$. Then

$$N_\alpha(\phi, Q) = \#\{(p,q) \in \mathbb{Z} \times \mathbb{N} : \gcd(p,q)\} \cap S_\alpha(\phi, Q)$$

Note that by Corollary 1.1.2 we have $N_\alpha(\phi, Q) \to \infty$ as $Q \to \infty$ provided $\phi(y) \ge \frac{1}{y^2}$, and by Exercise sheet 2, even when $\phi(y) \ge \frac{1}{\sqrt{5}y^2}$. If $\phi$ decays slowly enough then one can easily show that

$$N_\alpha(\phi, Q) = 2 \cdot \underbrace{\int_1^Q y\phi(y)dy}\, VolS_\alpha(\phi, Q) (\text{It } \underbrace{o(1)} \text{ tends to 0 as } Q \to \infty) \text{ as Q } \to \infty$$

More specifically, using tools we develop in Chapter 3, one can easily show that

$$\#\mathbb{Z}^2 \cap S_\alpha(\phi, Q) = 2 \cdot \int_1^Q y\phi(y)dy + \mathcal{O}(Q),$$

using Möbius-inversion, one can show that

$$N_\alpha(\phi, Q) = \frac{2}{S(2)} \cdot \int_1^Q y\phi(y)dy + \mathcal{O}(Q \log Q).$$

So we get an asymptotic formula

$$N_\alpha(\phi, Q) \sim \frac{2}{S(2)} VolS_\alpha(\phi, Q)$$

provided

$$\frac{Q\log Q}{\int_1^Q y\phi(y)dy} \to 0 \text{ as } Q \to \infty.$$

So, e.g., if $\phi(y) \geq \frac{(\log y)^2}{y}$.

However, the case when $\phi(y)$ decays much quicker is more interesting. Serge Lang in 1967 proved that if $\alpha$ is real quadratic then

$$N_\alpha(\frac{1}{x^2}, Q) = c_\alpha \cdot \log(Q) + \mathcal{O}(1).\ (c_\alpha > 0).$$

He mentioned that it would seem quite difficult to prove an asymptotic result for algebraic $\alpha$, let alone transcended.
Adams showed

$$N_e(\frac{1}{x^2}, Q) = c_e \cdot \frac{\log Q}{\log\log Q} + \mathcal{O}(1)(c_e > 0)$$

where $e = 2.7122\ldots$
Lang and Adams both used continuous fractions expansion. How can one prove asymptotics for $N_\alpha(\phi, Q)$? Here is an example.

**Example.** *Suppose $\phi(x) = \frac{1}{2x^2}$. Consider the continuous fraction expansion $\alpha = [a_0, a_1, a_2, \ldots]$. By Theorem 1.2.3 we know $|\alpha - \frac{p}{q}| < \phi(q) \Rightarrow \frac{p}{q}$ is a convergent. Moreover, if $\frac{p}{q} = \frac{p_n}{q_n}$ is the $n$-th convergent then $|\alpha - \frac{p}{q}| < \frac{1}{a_{n+1}q^2}$. So if all $a_i > 1$ then $|\alpha - \frac{p}{q}| < \phi(q) \forall$ convergent $\frac{p}{q}$.*
*Hence, $N_\alpha(\phi, Q) = \#\{n : q_n \leq Q\}$.*
*So, we need to compute the number of convergents $\frac{p_n}{q_n}$ with $q_n \leq Q$. We shall soon see that this is rather simple if $\alpha = [b, a, b, a, b, a, \ldots]$ with $a \mid b$.*
*We will get back to this after Theorem 1.2.5.*

A continued fraction $[a_0, a_1, a_2, \ldots]$ is called *periodic* if

$$\exists k \in \mathbb{N} \text{ and } L \in \mathbb{N}_0 \text{ such that } a_{k+l} a_l \forall l \geq L.$$

In this case we write $[a_0, a_1, a_2, \ldots] = [a_0, \ldots, a_L, \overline{a_{L+1}, \ldots, a_{L+k-1}}]$.

**Theorem 1.2.4.** *$\theta = [a_0, a_1, a_2, \ldots]$ is periodic $\iff$ $\theta$ is real quadratic ($\theta$ is real quadratic means $\exists D \in \mathbb{Z}[x] \setminus 0$ with $D(\theta) = 0$, but $\theta \notin \mathbb{Q}$ and $\theta \in \mathbb{R}$)*

See Ex Sheet 2 #3 for a special instance.
A proof can be found, e.g., in Hardy & Wright "The Theory of numbers", Oxford University press

Let's go back to the problem of computing $p_n, q_n$ of the $n$-th convergent. The general recursion formula is unhandy. But in certain cases there is a simple explicit formula. Consider $\theta = [b, a, b, a, \ldots] = [\overline{b, a}]$ and suppose $b = a \cdot c$ for some $c \in \mathbb{N}$. Now $\theta = b + \frac{1}{a + \frac{1}{b + \frac{1}{\cdots}}} = b + \frac{1}{a + \frac{1}{\theta}}$. Thus $\underbrace{a\theta^2 - ab\theta - b}\theta^2 - b\theta - c = 0$, so $\theta = \frac{b + \sqrt{b^2 + 4c}}{2}$ and we put $\bar{\theta} = \frac{b - \sqrt{b^2 + 4c}}{2}$.

**Theorem 1.2.5.** *The $p_n$ and $q_n$ of the $n$-th convergent $\frac{p_n}{q_n}$ of $\theta = [\overline{b, a}](b = ac)$ are give by*

$$p_n = c^{-\lfloor \frac{n+1}{2} \rfloor} \cdot U_{n+2}, q_n = c^{-\lfloor \frac{n+q}{2} \rfloor} \cdot u_{n+1}$$

*where*

$$u_n = \frac{\theta^n - \bar{\theta}^n}{\theta - \bar{\theta}}.$$

*(Recall:* $\theta = \frac{b+\sqrt{b^2+4c}}{2}, \bar{\theta} = \frac{b-\sqrt{b^2+4c}}{2}$*, so* $\theta - b\theta - c = 0, \bar{\theta}^2 - b\bar{\theta} - c = 0$*)*

*Proof.* For $n = 0, 1$ we note that

$$q_0 = q = u_1 \tag{39}$$

$$q_1 = a = \frac{b}{c} = \frac{u_2}{c} \tag{40}$$

$$p_0 = b = \theta + \bar{\theta} = u_2 \tag{41}$$

$$p_1 = ab + 1 = \frac{b^2 + c}{c} = \frac{(\theta + \bar{\theta})^2 - \theta\bar{\theta}}{c} = \frac{u_3}{c} \tag{42}$$

Put $\omega_{n+2} = c^{-\lfloor \frac{n+1}{2} \rfloor} u_{n+2}$.

So we need to show that $p_n = \omega_{n+2}$.

Using that $\theta^{n+2} = b\theta^{n+1} + c\theta^n$ and $\bar{\theta}^{n+2} = b\bar{\theta}^{n+1} + c\bar{\theta}^n$ and hence $u_{n+2} = \frac{\theta^{n+2} - \bar{\theta}^{n+2}}{\theta - \bar{\theta}} = bu_{n+1} + cu_n$.

Moreover, $u_{2m+2} = c^m \omega 2m + 2$, $u_{2m+1} = c^m \omega_{2m+1}$. Inserting this into the above, distinguishing $n$ even or odd yields:

$$\omega_{2m+2} = b\omega_{2m+1} + \omega_{2m} \tag{43}$$

$$\omega_{2m+1} = a\omega_{2m} + \omega_{2m-1} \tag{44}$$

Hence, $p_n$ and $\omega_{n+2}$ satisfy the same recurrence relation. and here the same two starting values, so $p_n = \omega_{n+2}$.

Similar for $q_n$. $\square$

*Counting Diophantine Approximation 2:*

We can use Theorem 1.2.5 to show that if $\theta = [\overline{b, a}]$ with $b = ac, a > 1$ then

$$N_\theta(\frac{1}{2x^2}, Q) = \frac{\log Q}{\log(\frac{Q}{\sqrt{c}})} + \mathcal{O}(1)$$

Indeed, we have already seen, that

$$N_\theta(\frac{1}{2x^2}, Q) = \#\{n : q_n \leq Q\}$$

By Theorem 1.2.5 we know

$$q_n \leq Q \iff c^{-\lfloor \frac{n+1}{2} \rfloor} \frac{\theta^n - \bar{\theta}^n}{\theta - \bar{\theta}} = \left(\frac{\theta}{\sqrt{c}}\right)^n \left(1 - \left(\frac{\bar{\theta}}{\theta}\right)^n\right) \epsilon \leq Q$$

where $\epsilon = \begin{cases} \frac{1}{\theta - \bar{\theta}} & 2 \mid n \\ \frac{1}{\sqrt{c}(\theta - \bar{\theta}} & 2 \nmid n \end{cases}$

$$\iff n \log\left(\frac{\theta}{\sqrt{c}}\right) + \log\left(1 - \left(\frac{\bar{\theta}}{\theta}\right)^n\right) + \log \epsilon \leq \log Q$$

Using Taylor series expansion we see that

$$|\log\left(1 - \left(\frac{\bar{\theta}}{\theta}\right)^n\right)| \leq |\frac{\bar{\theta}}{\theta - \bar{\theta}}|$$

This proves the claim.

## 1.3 Liouville's Theorem

Let $\alpha \in \mathbb{C}$. If $\exists D(x) \in \mathbb{Z}[x]$, $D \neq 0$ and $D(\alpha) = 0$ then we say $\alpha$ is *algebraic*. In this case $\exists D(x) = a_0 x^d + \cdots + a_d \in \mathbb{Z}[x]$ with

- $D(\alpha) = 0$
- $a_0 > 0$
- $\gcd(a_0, \ldots, a_d) = 1$
- $\deg D(x)$ minimal

Imposing all these condition renders $D$ unique; We write $D_\alpha(x)$ and call this the *minimal polynomial* of $\alpha$. If $\alpha$ is algebraic then we say $\deg D_\alpha$ is the *degree of* $\alpha$.

**Example.**
- $\alpha = 0, D_\alpha(x) = x$
- $\alpha = \sqrt{2} + 1, D_\alpha(x) = (x-1)^2 - 2 = x^2 - 2x - 1$
- $\alpha = \frac{1}{\sqrt{2}}, D_\alpha(x) = 2x^2 - 1$

**Theorem 1.3.1** (1.3.1 Liouville's Theorem)**.** *Suppose $\alpha$ is a real, algebraic number of degree $d$. Then $\exists c(\alpha) > 0$ such that*

$$|\alpha - \frac{p}{q}| > \frac{c(\alpha)}{q^d}$$

*for every $(p, q) \in \mathbb{Z} \times \mathbb{N}$ with $\alpha \neq \frac{p}{q}$.*

*Proof.* Suppose $|\alpha - \frac{p}{q}| > 1$ then the claim holds for every $c(\alpha) > 1$. Now suppose $|\alpha - \frac{p}{q}| \leq 1$. Taylor series expansion at $D_\alpha$ about $\alpha$ gives:

$$D_\alpha(x) = \sum_{i=1}^{d} (x - \alpha)^i \frac{1}{i!} D_\alpha^{(i)}(\alpha)$$

Hence,

$$|D_\alpha\left(\frac{p}{q}\right)| = |\sum_{i=1}^{d} \left(\frac{p}{q} - \alpha\right)^i \frac{1}{i!} D_\alpha^{(i)}(\alpha)| \overset{\leq}{(D)label} |\frac{p}{q} - \alpha| \frac{1}{c(\alpha)}$$

where

$$c(\alpha) = \left(1 + \sum_{i=1}^{d} \frac{1}{i!} |D_\alpha^{(i)}(\alpha)|\right)^{-1}$$

Now if $D_\alpha$ has a rational root then it must have degree one, so have only *one* root. Thus $D_\alpha\left(\frac{p}{q}\right) \neq 0$ unless $\alpha = \frac{p}{q}$. Hence, if $\alpha \neq \frac{p}{q}$ we get

$$|D_\alpha\left(\frac{p}{q}\right)| = |\frac{\text{non-zero integer}}{q^d} \geq \frac{1}{q^d}.$$

Combing this with (D)label yields

$$|\alpha - \frac{p}{q}| > \frac{c(\alpha)}{q^d}.$$

□

We say a real number $\alpha$ is a *Liouville number* if for every $n \in \mathbb{N}$

$$0 < |\alpha - \frac{p}{q}| < \frac{1}{q^n}$$

has a solution. $p, q \in \mathbb{Z}$ with $q > 1$.

**Example.** *$\alpha = \sum_{k=1}^{\infty} 10^{-k^k}$ is a Liouville number. Let $n \in \mathbb{N}$ and put $p = \sum_{k=1}^{n} 10^{n^n - k^k}$ and $q = 10^{n^n}$. Then $0 < |\alpha - \frac{p}{q}| = \sum_{k>n} 10^{-k} \leq 2 \cdot 10^{-(n+1)^{(n+1)}} < 10^{-n^{(n+1)}} = q^{-n}$*

**Corollary 1.3.2** (1.3.2)**.** *Every Liouville number is transcendental (i.e., not algebraic).*

*Proof.* Immediate from Theorem 1.3.1 (Liouville's Theorem). □

Algebraic numbers are enumerable and thus have Lebesgue measure zero. It's not difficult to show that the set of Liouville numbers, while *not* enumerable, also has measure zero. In fact "most" real numbers are "not very far" from badly approximable as the following theorem shows.

**Theorem 1.3.3** (Khintchine)**.** *Suppose $\psi : \mathbb{N} \to (0, \infty)$ is monotone decreasing (not necessarily strictly). The set*

$$A_\psi = \{\alpha \in \mathbb{R} : |\alpha - \frac{p}{q}| < \frac{\psi(q)}{q} \text{ has } \infty\text{-many solutions } (p,q) \in \mathbb{Z} \times \mathbb{N}\}$$

*has a Lebesgue measure zero of $\sum_{q=1}^{\infty} \psi(q)$ converges and has full Lebesgue measure (i.e. the complement has measure zero) if $\sum_{q=1}^{\infty} \psi(q)$ diverges.*

We will not prove this Theorem. (For a proof see e.g. Glyn Harman "Metric number theory".)

**Example.**

- *Take $\psi(q) = \frac{1}{q}$. We already know that $A_\psi = \mathbb{R} \setminus \mathbb{Q}$. And indeed $\sum \psi(q)$ diverges...*
- *$\psi(q) = \frac{1}{q \log(q-1)}$. Then $\sum \psi(q)$ diverges and thus $A_\psi$ has full measure.*
- *$\psi(q) = \frac{1}{q(\log(q+1))^{1+\epsilon}} (\epsilon > 0)$ then $\sum \psi(q)$ converges, so $A_\psi$ has measure zero.*

## 1.4 4 Thue-Siegel-Roth theorem

In Section 1 we have seen that $\infty$-many solutions $\frac{p}{q}$ to $\left|\sqrt{2} - \frac{p}{q}\right| < \frac{1}{q^2}$ leads to $\infty$-many solutions $(x,y) \in \mathbb{Z}^2$ of $x^2 - 2y^2 = 1$. What about $x^3 - 2y^3 = 1$? Starting as for $x^2 - 2y^2$ we get

$$y^3 \left|\frac{x}{y} - 2^{1/3}\right| \underbrace{\left|\frac{x}{y} - 2^{1/3}\omega\right|}_{\geq \operatorname{Im}\omega} \underbrace{\left|\frac{x}{y} - 2^{1/3}\omega^2\right|}_{\geq (\operatorname{Im}\omega)^2}$$

where $\omega = e^{\frac{2\pi i}{3}}$.

So to get boundedness of $x^3 - 2y^3$ for $\infty$-many $(x,y)$ we need $\exists c > 0$ such that

$$\left|\frac{x}{y} - 2^{1/3}\right| < \frac{c}{y^3}$$

has $\infty$-many solutions $(x,y) \in \mathbb{Z} \times \mathbb{N}$.

Theorem 1.3.3 tells us that we would be extremely lucky if that were the case. And erven if so, we still would lack the group structure for $\mathbb{Z}+\sqrt{2}\mathbb{Z}$ (closed under multiplication but $\mathbb{Z}+2^{1/3}\mathbb{Z}$ is not). On the other hand, suppose we could show that

$$\left|\frac{x}{y} - 2^{1/3}\right| < 1/y^{\lambda}$$

has only finitely many solutions $(x,y) \in \mathbb{Z}\times\mathbb{N}$ for some fixed $\lambda < 3$. As $x^3 - 2y^3 = 1$, and $y \neq 0$ yields:

$$\left|\frac{x}{y} - 2^{1/3}\right| < \frac{1}{2^{1/3}(\operatorname{Im}\omega)^2 y^3}$$

We would conclude that $x^3 - 2y^3 = 1$ has only finitely many solutions $(x,y) \in \mathbb{Z}^2$. Note that "*deg*" $2^{1/3} = 3(D(x) = x^3 - 2)$ and so Liouville's Theorem yields only $\lambda = 3$ not $\lambda < 3$. So the big challenge is to improve Liouville's Theorem. After Liouville it has taken 65 years until the first breakthrough was obtained by Axel Thue in 1909.

**Theorem 1.4.1** (Thue)**.** *Let $\alpha$ be a real algebraic number of degree $d \geq 2$, and let $\lambda > \frac{d}{2} + 1$. Then $\exists c = c(\alpha, \lambda) > 0$ such that*

$$\left|\alpha - \frac{p}{q}\right| > \frac{c}{q^{\lambda}}, \quad \forall (p,q) \in \mathbb{Z} \times \mathbb{N}.$$

- Note that for $d = 2$ Liouville is stronger.
- Given $\alpha$ and $\lambda$ there is no method to determine a feasible value for $c$. This is in strong contrast to Liouville's Theorem.

Just as for $x^3 - 2y^2 = 1$ one can now very easily show that if $f(X,Y) = a_0(X - \alpha_1 Y)\cdots(X - \alpha_d Y) \in \mathbb{Q}[X,Y]$ with $a_0 \neq 0, d \geq 3$, and $\alpha_1, \ldots, \alpha_d$ pairwise distinct, and $b \in \mathbb{Q} \smallsetminus \{0\}$, then

$$f(x,y) = b$$

has only finitely many solutions $(x,y) \in \mathbb{Z}^2$.
Wrong if $d = 2$:

$$X^2 - 2Y^2 = 1$$

or $b = 0$:

$$X^3 - Y^3 = 0$$

or $\alpha_1, \ldots, \alpha_d$ not pairwise distinct:

$$(X - Y)^5 = 1$$

We will show that Theorem 1.4.1 implies even the following stronger result.

**Theorem 1.4.2** (1.4.2 Generalized Thue equations)**.** *Let $f(X,Y) = a_0(X-\alpha_1 Y)\cdots(X-\alpha_d Y) \in \mathbb{Q}[X,Y]$ with $a_0 \neq 0, d \geq 3$ and $\alpha_1,\cdots,\alpha_d$ pairwise distinct. Let $g(X,Y) \in \mathbb{Q}[X,Y]$ of total degree $< \frac{d}{2} - 1$.*
*Then there are only finitely many $(X,Y) \in \mathbb{Z}^2$ with*

$$f(x,y) = g(x,y)$$

*and $g(x,y) \neq 0$.*

**Example.**

$$x^5 - 2y^5 = x - y$$

*has only finitely many solutions $(x,y) \in \mathbb{Z}^2$. Indeed if $x - y = 0$ then $x^5 - 2y^5 = 0$ thus $x = y = 0$. Note Theorem can go wrong if $\alpha_1 = \alpha_2$:*

$$(X^2 - 2Y^2)^2 = 1.$$

*(assuming Theorem 1.4.1).* If $y = 0$ then we have at most d possibilities for x. So we can assume $y \neq 0$. We claim that

$$|x| \leq c_1 |y|$$

for some $c_1 = c_1(f,g)$. Clearly true when $|x| \leq |y|$, so let's assume $|x| > |y|$. Then we write

$$f(x,y) = \sum_{i=0}^{d} a_i x^{d-i} y^i = \sum_{j+k \leq d-1} b_{jk} x^j y^k = g(x,y)$$

Dividing by $x^{d-i}$ yields

$$a_0 x = -\sum_{i=0}^{d} a_i \frac{y^i}{x^{i-1}} + \sum_{j+k \leq d-1} b_{jk} x^{j-d+1} y^k$$

We have

$$\left| \frac{y^i}{x^{i-1}} \right| \leq |y|$$

and

$$\left| \frac{y^k}{x^{d-1-j}} \right| \leq |y|^{j+k-(d-1)} \leq 1$$

Therefore $|x| \leq c_1|y|$, e.g. with $c_1 = \frac{1}{|a_0|}\left(\sum |a_i| + \sum |b_{jk}|\right) + 1$. From

$$f(x,y) = g(x,y), (\star)$$

we get

$$|\alpha_0| \prod_{j=1}^{d} \left| \frac{x}{y} - \alpha_i \right| \leq c_2 |y|^{e-d}$$

where $c_2 = c_2(c_1, g)$ and $e < \frac{d}{2} - 1$. So assume $(\star)$ has $\infty$-many solutions $(x,y) \in \mathbb{Z}^2$. Then $\exists i$, say $i = 1$, such that $\left|\frac{x}{y} - \alpha_1\right| \leq \mu := \frac{1}{2}\min_{j \neq i}\{|\alpha_j - \alpha_1|\} > 0$ for $\infty$-many $(x,y)$ of these solutions of $(\star)$. Now

$$\left| \frac{x}{y} - \alpha_j \right| \geq \left| |\alpha_j - \alpha_i| - \left| \frac{x}{y} - \alpha_1 \right| \right| \geq 2\mu - \mu = \mu > 0$$

Hence, we conclude

$$\left|\frac{x}{y} - \alpha_1\right| \le \frac{c_2}{|a_0|}\mu^{1-d}|y|^{e-d}, (\star\star)$$

for these solutions $(x, y)$. Here we can assume $y > 0$ (just replace $x$ by $-x$). Now let $d_1$ be the degree of $\alpha_1$. As $f(x,1) \in \mathbb{Q}[x]$, $f(x,1) \neq 0$ and $f(\alpha_1, 1) = 0$. Thus $d_1 \le d$. Moreover, $d - e > \frac{d}{2} + 1$ and this $\exists\lambda$ such that

$$d - e > \lambda > \frac{d_1}{2} + 1.$$

If $d_1 \ge 2$ then Theorem 1.4.1 implies that $(\star)$ has only finitely many solutions $(x, y) \in \mathbb{Z}^2$. Finally suppose $d_1 = 1$. Then $\alpha_1 = \frac{p}{q}$, and $(\star\star)$ yields:

$$\left|x - \frac{p}{q}y\right| \le c_3 y^{e-d+1} \le c_3 y^{-\frac{d}{2}}.$$

Thus $x = \frac{p}{q}y = \alpha_1 y$ for $y$ large enough. But then $0 = f(x,y) = g(x,y)$ a contradiction. $\square$

After Thue came Siegel (1921) who improved the exponent $\frac{d}{2} + 1$ to $2\sqrt{d}$. This was slightly improved by Dyson and Gelfand (1947) to $\sqrt{2d}$. Finally in 1955 came Roth:

**Theorem 1.4.3** (1.4.3 (Roth))**.** *Let $\alpha$ be a real, algebraic irrational number, and $\lambda > 2$. Then $\exists c = c(\alpha, \lambda) > 0$ such that*

$$\left|\alpha - \frac{p}{q}\right| \ge \frac{c}{q^\lambda}, \quad \forall (p,q) \in \mathbb{Z} \times \mathbb{N}.$$

By Corollary 1.1.2 $\lambda > 2$ is best-possible. But if we allow more general functions $\phi(q)$, not only powers of $q$, then an improvement might be possible. However, since 1955 nobody was able to replace $q^{-\lambda}$ by a function $\phi(q)$ that decays more slowly, e.g. $\phi(q) = q^{-2}(\log q)^{-1}$.
However, back to the case where $\phi(q)$ is a power of $q$.
From Theorem 1.3.3. we know that for a generic real $\alpha$

$$\left|\frac{p}{q} - \alpha\right| < q^{-\lambda}$$

has only finitely many solutions $p, q \in \mathbb{Z} \times \mathbb{N}$ provided $\lambda > 2$. Any by Corollary 1.1.2 every irrational real number has $\infty$-many solutions when $\lambda = 2$. And so from Roth's Theorem we see an algebraic irrational behaves "essentially" like a generic number.

Roth's Theorem has various new applications to, e.g., Diophantine equations and transcendence. Let's consider just one now transcendence result:
Take $\alpha = \sum_{k=1}^\infty 2^{-3^k}$; put $q_n = 2^{3^n}$ and $p_n = q_n \sum_{k=1}^n 2^{-3^k}$. Then $0 < \left|\alpha - \frac{p_n}{q_n}\right| = \sum_{k=n+1}^\infty 2^{-3^k} < 2 \cdot 2^{-3^{n+1}} = 2 \cdot 2 \cdot q_n^{-1}$ so by Roth's Theorem $\alpha$ is transcendental.

How does one prove results like Roth's Theorem of the kind

$$\left|\alpha - \frac{p}{q}\right| \ge \phi(q)?$$

The idea is to find good rational approximations.

$$\left|\alpha - \frac{p_n}{q_n}\right| \leq \delta_n$$

with $\delta_n$ "pretty small". Then

$$\left|\alpha - \frac{p}{q}\right| \geq \left|\frac{p_n}{q_n} - \frac{p}{q}\right| - \left|\alpha - \frac{p_n}{q_n}\right|$$

If

$$\frac{p_n}{q_n} \neq \frac{p}{q} \tag{45}$$

then

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{1}{qq_n} - \delta_n.$$

If we are lucky then $\delta_n < \frac{1}{qq_n}$ and we get a positive lower bound. How do we find these $\frac{p_n}{q_n}$?
Usually this is a difficult task, but sometimes one can easily see these approximations $\frac{p_n}{q_n}$. Here is an example.
Take again $\alpha = \sum_{k=1}^{\infty} 2^{-3^k}$. Then we can take again $q_n = 2^{3^n}$, $p_n = q_n \sum_{k=1}^{n} 2^{-3^k}$; so $\left|\alpha - \frac{p_n}{q_n}\right| < 2 \cdot q_n^{-3}$. Hence, if

$$\frac{p_n}{q_n} \neq \frac{p}{q}$$

then

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{1}{qq_n} - \frac{2}{q_n^3}$$

If $q_n^2 > 4 \cdot q$ then

$$\frac{1}{qq_n} - \frac{2}{q_n^3} \geq \frac{q}{2 \cdot qq_n}$$

As $\frac{p_n}{q_n}$ tends strictly monotonously to $\alpha$, we have $\frac{p_n}{q_n} \neq \frac{p}{q}$ or $\frac{p_{n+1}}{q_{n+1}} \neq \frac{p}{q}$ Let $m$ be minimal with $q_m > 4 \cdot q$. Hence

$$q_m^{\frac{2}{3}} = q_{m-1}^2 \leq 4 \cdot q < q_m^2$$

If $\frac{p_m}{q_m} \neq \frac{p}{q}$ we take $n = m$ and $n = m + 1$ else. We conclude

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{1}{2qq_n} \geq \frac{1}{2qq_{m+1}} \geq \frac{1}{2q}\frac{1}{q_m^3} \geq \frac{1}{2q}\frac{1}{(4q)^{\frac{9}{2}}} = 2^{-10}q^{-\frac{11}{2}}$$

In this example everything works out nicely, e.g., (ref*) could easily be guaranteed by using $\frac{p_n}{q_n}$ tending strictly monotonously to $\alpha$. However, in Roth's Theorem (ref*) becomes the major-problem.

## 1.5 5 Simultaneous Diophantine approximation and the Subset Theorem

Suppose $\alpha_1, \dots, \alpha_n$ are real numbers. Theorem 1.1.1 can be generated to yield a solution $(x_1, \dots, x_n, y) \in \mathbb{Z}^n \times \mathbb{N}$ at the system

$$\left|\frac{x_i}{y} - \alpha\right| \leq \frac{1}{y \cdot Q}(1 \leq i \leq n), 0 < y < Q.$$

(c.f. Exercise sheet 4). This in turn yields $\infty$-many solutions $(x_1, \dots, x_n, y) \in \mathbb{Z}^n \times \mathbb{N}$ of the system

$$\left|\frac{x_i}{y} - \alpha_i\right| < \frac{1}{y^{1+\frac{1}{n}}}(1 \leq i \leq n).$$

provided at least one of the $\alpha_i$'s is irrational. So Corollary 1.1.2 extends to simultaneous approximation. A much deeper fact is that Roth' Theorem also extends to simultaneous approximation.

For $\underline{x} \in \mathbb{R}^n$ we write $\|\underline{x}\| = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$ for the Euclidean length.

**Theorem 1.5.1** (Subspace Theorem, Schmidt)**.** *Suppose $L_i(\underline{x}) = \sum_{j=1}^n a_{ij} x_j (1 \leq i \leq n)$ are linearly independent linear forms with algebraic coefficients $a_{ij}$. Let $\delta > 0$. Then the solutions $\underline{x} \in \mathbb{Z}^n \smallsetminus \underline{0}$ of*

$$|L_1(\underline{x}) \dots L_n(\underline{x})| < \|\underline{x}\|^{-\delta}$$

*lie in finitely many proper subspaces of $\mathbb{Q}^n$.*

**Remark.** *linearly independent linear forms means the coefficient vectors $(a_{i1}, \dots, a_{in})$ are linearly independent over $\mathbb{C}$.*

**Corollary 1.5.2** (1.5.2)**.** *Let $\delta > 0$, suppose $\alpha_1, \dots, \alpha_n$ are algebraic and $1, \alpha, \dots, \alpha_n$ are linearly independent over $\mathbb{Q}$. Then there are only finitely many $(x_1, \dots, x_n, y) \in \mathbb{Z}^n \times \mathbb{N}$ with*

$$(5.1)\left|\frac{x_i}{y} - \alpha_i\right| < \frac{1}{y^{1+\frac{1}{n}+\delta}}(1 \leq i \leq n) \tag{46}$$

*Proof.* (assuming Theorem 1.5.1) Put $\underline{X} = (X_1, \dots, X_n, Y)$, $L_i(\underline{X}) = \alpha_i Y - X_i (1 \leq i \leq n)$, $L_n(\underline{X}) = Y$. These $n+1$ linear forms in $n+1$ unknowns are linearly independent. With $\underline{x} = (x_1, \dots, x_n, y)$ the solutions of (5.1) yield

$$|L_1(\underline{x}) \dots L_{n+1}(\underline{x})| < \frac{1}{y^\delta} < \frac{1}{\|\underline{x}\|^{\frac{\delta}{2}}}$$

if $y$ is large enough. so by Theorem 1.5.1 (in $n+1$ dimensions), we set that the solutions lie in finitely many prober sub spaces at $\mathbb{Q}^{n+1}$. Pick one of these (of co-dimension I say). It is given by an equation $c_1 x_1 + \dots + c_n x_n + c_{n+1} y = 0$ where $c_i \in \mathbb{Q}$ not all zero. On this subspace we have

$$(c_1\alpha_1 + \dots + c_n\alpha_n + c_{n+1})y = c_1(\alpha_1 y - x_1) + \dots + c_n(\alpha_n y - x_n).$$

Put $\gamma = c_1\alpha_1 + \dots + c_n\alpha_n + c_{n+1}$. By $\mathbb{Q}$-linearly independence of $1, \alpha_1, \dots, \alpha_n$ we have $\gamma \neq 0$. Hence,

$$|\gamma||y| \leq |c_1||\alpha_1 y - x_1| + \dots + |c_n||\alpha_n y - x_n| \leq (|c_1| + \dots + |c_n|)\frac{1}{y^{1+\frac{1}{n}+\delta}} \leq |c_1| + \dots + |c_n|$$

So $|y|$ is bounded and we are done. $\square$

In applications one sometimes needs a "$p$-adic" version of the subspace Theorem in which one approximates with respect to also the so called $p$-adic absolute values.

**Definition** (Absolute values)**.** *An absolute value on a field $K$ is a map $|\bullet| : K \to [0,\infty)]$ such that*

- $|x| = 0 \iff x = 0$
- $|x \cdot y| = |x| \cdot |y|$
- $|x + y| \le |x| + |y|$

**Example.**

- *$K$ arbitrary.* $|x| = \begin{cases} 0 & x = 0 \\ 1 & x \neq 0 \end{cases}$ *the* trivial absolute value*.*
- *$K = \mathbb{Q}$, $|\bullet|$ =* standard absolute value *on $\mathbb{Q}$. To distinguish it from other absolute values let's write it as $|\bullet| = |\bullet|_\infty$.*
- *$K = \mathbb{Q}$ and let $p \in \mathbb{N}$ be a prime number. If $x \in \mathbb{Q}, x \neq 0, \pm 1$, then $\exists$ a unique prime factorisation $x = \pm p_1^{a_1} \dots p_s^{a_s}$ where $p_1, \dots, p_s$ primes and $a_i \in \mathbb{Z} \setminus 0$. For any prime $p \in \mathbb{N}$ write $ord_p(x)$ for the exponent of $p$ in the prim-factorisation of $x$ (e.g. $ord_{p_i} x = a_i$). For $x = \pm 1$ we put $ord_p x = 0 \forall p_i$. The $p$-*adic absolute value $1 \cdot 1_p$ *on $\mathbb{Q}$ is defined by*

$$|x|_p = \begin{cases} 0 & : x = 0 \\ p^{-ord_p(x)} & : x \neq 0 \end{cases}$$

*The multiplicativity is clear. Note that $ord_p(x_1+x_2) \ge \min\{ord_p(x_1), ord_p(x_2)\}$. Hence, $|x_1 + x_2|_p = p^{-ord_p(x_1+x_2)} \le p^{-\min\{ord_p(x_1), ord_p(x_2)\}} \underbrace{=}$ strong triangle inequality* $\max |x_1|_p, |x_2|_p\}$

$|x_1|_p + |x_2|_p$ *An absolute value that satisfies the strong triangle inequality is called non-Archimedean.*

**Definition 1.5.3.** *We set $M_\mathbb{Q} = \{primes\ in\ \mathbb{N}\} \cup \{\infty\}$. Then for each $v \in M_\mathbb{Q}$ we get an absolute value $|\cdot|_v$. Note that if $v \in M_\mathbb{Q}$ and $p$ a prime, $a \in \mathbb{Z}$, then*

$$|\pm p^a|_v = \begin{cases} p & : v = p \\ p^a & : v = \infty \\ 1 & : v \neq p, v \neq \infty \end{cases}$$

*Hence*

$$\prod_{v \in M_\mathbb{Q}} |1 \pm p^a|_v = 1$$

*and so by multiplicativity we conclude*

$$\prod_{v \in M_\mathbb{Q}} |x|_v = 1$$

*for all $x \in \mathbb{Q}$, $x \neq 0$. (PF) This is the so-called product formula (PF) on $\mathbb{Q}$.*

Next, we want to introduce a notion of "arithmetic complexity" on elements in $\mathbb{Q}^{n+1}$, the so-called projective height:

$$H_{\mathbb{P}^n} : \mathbb{Q}^{n+1} \to [1, \infty)$$

defined by

$$H_{\mathbb{P}^n}(\underline{x}) = \prod_{v \in M_{\mathbb{Q}}} |\underline{x}|_v$$

where $|\underline{x}|_v = \max\{|x_0|_v, \dots, |x_n|_v\}$.

**Example.** *If $\underline{x} = (x_0, \dots, x_n)$ where $x_0, \dots, x_n \in \mathbb{Z}$ and $\gcd(x_0, \dots, x_n) = 1$. Then $|x|_p = 1$ for all primes p. Hence,*

$$H_{\mathbb{P}^n}(\underline{x}) = \max\{|x_0|_\infty, \dots, |x_n|_\infty\}.$$

*Note that $H_{\mathbb{P}^n}(\lambda \cdot \underline{x}) = H_{\mathbb{P}^n}(\underline{x}) \forall \lambda \in \mathbb{Q} \setminus 0$.*

**Theorem 1.5.4** (1.5.3 p-adic Subspace Theorem, Schlickewei and Schnidt)**.** *Let $\delta > 0$ and let $S \subset M_{\mathbb{Q}}$ be finite and with $\infty \in S$. For $v \in S$ let $L_{v_1}, \dots, L_{v_n}$ be n linearly independent linear forms in n variables with coefficients in $\mathbb{Q}$. Then the set of solutions $\underline{x} \in \mathbb{Q}^{n+1} \setminus 0$ of*

$$\prod_{v \in S} \prod_{i=1}^{n} \frac{|L_{v_i}(\underline{x})_v|}{|\underline{x}|_v} < H_{\mathbb{P}^{n-1}}(\underline{x})^{-n-\delta}$$

*lie in finitely many proper subspace of $\mathbb{Q}^n$.*

An interesting consequence is a finiteness result for $S$-unit equations.
$S$-integers and $S$-units:
Let $v$ be a non-Archimedean absolute value or a field $K$. Then

$$O_v = \{x \in K : |x|_v \leq 1\}$$

is called the valuation ring of $v$. It is indeed a ring, e.g., $|x|_v, |y|_v \leq 1$ then

$$|x + y|_v \leq \max\{|x|_v, |y|_v\} \leq 1.$$

In particular, if $K = \mathbb{Q}$ and $v = p$ then $O_v$ is a sub-ring of $\mathbb{Q}$.

Now let $S \subset M_{\mathbb{Q}}$ be finite and $\infty \in S$. We define the set of *$S$-integers $O_S$* to be

$$O_S = \cap_{v \notin S} O_v.$$

As $\infty \in S$, this is an intersection of rings, hence a ring.

**Example.** *If $S = \{\infty\}$, then $O_S = \mathbb{Z}$. If $S = \{\infty, p_1, \dots, p_s\}$ then*

$$O_S = \{\frac{m}{p_1^{a_1} \dots p_s^{a_s}} : m \in \mathbb{Z}, a_1, \dots, a_s \in \mathbb{N}_0\}$$

We say $x \in \mathbb{Q}$ is an $S$-unit if $x \neq 0$ and $x, x^{-1}$ are both in $O_S$. So if $S = \{\infty\}$, then $\pm 1$ are the only $S$-units. If $S = \{\infty, p_1, \dots, p_s\}$ then $x$ is an $S$-unit $\iff$ $x = \pm \prod_{p \in S \setminus \infty} p_p^a$ (and $ap \in \mathbb{Z}$).

**Theorem 1.5.5** (1.5.4 $S$-unit equation)**.** *Let $S \subset M_{\mathbb{Q}}$ be finite, and $\infty \in S$. Let $\alpha_0, \dots, \alpha_n$ be non-zero and in $\mathbb{Q}$. Then*

$$\alpha_0 x_0 + \cdots + \alpha_n x_n = 0$$

*has only finitely many solutions $\underline{x} = (x_0, \dots, x_n)$ if:*

- *$x_0, \dots x_n$ are $S$-units*
- *we identify proportional solutions (i.e., $\underline{x} = \lambda \underline{x}$ for $\lambda \in \mathbb{Q} \smallsetminus 0$).*
- *no proper sub-sum vanishes, i.e., $\sum_I \alpha_i x_i \neq 0$ for all $\emptyset \subsetneq I \subsetneq \{0, 1, \dots, n\}$.*

**Remark.** *$S = \{\infty, p\}$ $x_0 + x_1 + x_2 + x_3 = 0$ then $x_0 = -x_1 = 1$, $x_2 = -x_3 = p^a$ $(a \in \mathbb{Z})$ are solutions in $S$-units. So non-vanishing condition is needed!*

**Example.** *The exponential Diophantine equation*

$$3^x + 5^y - 7^z = 1$$

*has solutions, e.g., $(x, y, z) = (0, 0, 0)$ or $(x, y, z) = (1, 1, 1)$. However, with $S = \{\infty, 3, 5, 7\}$ each solution $(x, y, z)$ yields a solution $u_0 = 3x$, $u_1 = 5y$, $u_2 = -7z$, $u_3 = -1$ of the $S$-unit equation $u_0 + u_1 + u_2 + u_3 = 0$. These solutions are all non-proportional. Moreover, no sub-sum vanishes unless $xyz = 0$ but then we easily see that $x = y = z = 0$. Hence, Theorem 1.5.4 yields finiteness.*

*assuming Theorem 1.5.3.* Induction on $n$. If $n = 1$ then $\alpha_0 x_0 + \alpha_1 x_1 = 0$, so all solutions are proportional to $(1, -\frac{\alpha_0}{\alpha_1}$. Now suppose the claim holds for all $S$-unit equations in $\leq n$ variables. As $x_i$ are $S$-units we have

$$|x_i|_v = 1 \forall v \notin S.$$

By the product formula (PF)

$$1 = \prod_{v \in M_{\mathbb{Q}}} |x_i|_v = \prod_{v \in S} |x_i|_v,$$

and thus

$$\prod_{v \in S} \prod_{i=0}^{n} |x_i|_v = 1.$$

Let $\underline{\tilde{x}} = (x_0, \dots, x_{n-1})$. For each $v \in S$ pick $i(v)$ with $0 \leq i(v) \leq n-1$. So we get $n^{\#S}$ such tuples $(i(v))_{v \in S}$. Choose one of those tuples and consider all solutions of $\alpha_0 x_0 + \cdots + \alpha_n x_n = 0$ with

$$|\underline{\tilde{x}}|_v = |x_{i(v)}|_v$$

Choose the set of linear forms

$$\{L_{v_j} : 1 \leq j \leq n\} = \{X_0, X_1, \dots, X_{n-1}, \frac{\alpha_0}{\alpha_n} X_0 + \cdots + \frac{\alpha_{n-1}}{\alpha_n} X_{n+1}\} \smallsetminus \{X_{i(v)}\}$$

Then

$$\prod_{v \in S} \prod_{j=1}^{n} \frac{|L_{v_j}(\underline{\tilde{x}})|_v}{|\underline{\tilde{x}}|_v} = \frac{1}{H_{\mathbb{P}_{n+1}}(\underline{\tilde{x}})^{n+1}}$$

By Theorem 1.5.3 the solutions $\underline{\tilde{x}}$ lie in finitely many proper subspaces. Take one of these then all elements in this subspace satisfy an equation

$$c_0x_0 + \dots c_{n-1}x_{n-1} = 0 (c_i \in \mathbb{Q}, \text{ not all } = 0!)$$

Let $J_0$ be the set of $i$ with $c_i \neq 0$. Then

$$\sum_{i \in J_0} c_i x_i = 0 \text{ marker(S)} \tag{47}$$

is an $S$-unit equation in $\leq n$ unknowns. For every solution of "marker(S)" there is a set $J \subset J_0$, $J \neq \varnothing$, such that

$$\sum_{i \in J} c_i x_i = 0$$

and <u>no</u> sub-sum vanishes. By the induction hypotheses, up to proportionality, we get only finitely many solutions. Moreover, the number of possible choices $J$ is finite. Therefore it suffices to consider solutions $\{x_i\}_{i\in J}$ that are proportional to a fixed $\{u_i\}_{i\in J}$, i.e., $x_i = \xi u_i (i \in J)$. Returning to our initial equation $\sum_{i=0}^n \alpha_i x_i = 0$ we get

$$\xi(\sum_{i\in J} \alpha_i u_i) + \sum_{i \notin J} \alpha_i x_i) = 0$$

If $\sum_{i\in J} \alpha_i x_i \neq 0$ then the above is an $S$-unit equation in $1 + (n+1) - \#J \leq n$ unknowns, namely $\xi, x_i\ (i \notin J)$. By the induction hypothesis we get only finitely many non-proportional solutions $\{x_i\}_{i=0}^n$ for which no sub-sum vanishes. Finally, if $\sum_{i\in J} \alpha_i x_i = 0$ then $\sum_{i \notin J} \alpha_i x_i = 0$ and we ignore these solutions by assumption of the Theorem. □

## 1.6 6. Further generalizations and open problems

Let $\alpha, \beta \in \mathbb{R} \setminus \mathbb{Q}$ and consider the linearly independent linear forms $L_1 \underline{\mathrm{x}} = x_0\alpha - x_1$, $L_2(\underline{\mathrm{x}}) = x_0\beta - x_2$, $L_3(\underline{\mathrm{x}}) = x_0$ If $\alpha, \beta$ are algebraic then Theorem 1.5.1 implies that the solutions $\underline{\mathrm{x}} \in \mathbb{Z}^3 \setminus \underline{0}$ of

$$|L_1(\underline{\mathrm{x}}) \cdot L_2(\underline{\mathrm{x}}) \cdot L_3(\underline{\mathrm{x}})| < \|\underline{\mathrm{x}}\|^{-\delta}\ (\delta > 0)$$

lie in finitely many proper subspaces of $\mathbb{Q}^3$.

However, in the following is a long-standing conjecture.

**Conjecture 1.6.1** (Littlewood-conjecture, around 1920). *Let $\alpha, \beta \in \mathbb{R} \setminus \mathbb{Q}$ and $\varepsilon > 0$. Then $\exists \underline{x} \in \mathbb{Z}^3$ such that*

$$0 < |L_1(\underline{x}) L_2(\underline{x}) L_3(\underline{x})| < \varepsilon$$

**Remark.** *Note that the conjecture is obviously true if $\alpha = \beta$ (by Corollary 1.1.2) or if $\alpha$ or $\beta$ are <u>not</u> badly approximable.*

Let's consider again approximations to <u>one</u> real $\alpha$. So far our approximations were $\frac{p}{q} \in \mathbb{Q}$. If we replace $\mathbb{Q}$ by a smaller or larger set then we get now interesting problems.

**Open Problem 1.6.2** (1.6.2). *Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and $\lambda < 2$. Does $\left|\alpha - \frac{p}{q}\right| < q^{-\lambda}$ have $\infty$-many solutions $(p, q) \in \mathbb{Z} \times \mathbb{N}$ with:*

- *$p$ and $q$ are both square-free.*
  *Best-result (Hoath-Brown 1984): Yes, if $\lambda < \frac{5}{3}$.*

- *$q$ is prime?*
  *Best-result (Matomaki, 2009): Yes, if $\lambda < \frac{4}{3}$.*

Let's now consider problems in which $\mathbb{Q}$ is replaced by a certain subset $A$ of

$$\overline{\mathbb{Q}} = \{\alpha \in \mathbb{C} : \alpha \text{ algebraic}\}.$$

If we assume that $A \subset \mathbb{R}$ then we still can use the (usual) absolute value on $\mathbb{R}$ to measure

$$|\alpha - x| \ (x \in A).$$

But usually we have no "natural denominators" for $x \in A$. But there is a natural way to interpret the original setting that easily generalizes from $A = \mathbb{Q}$ to $A = \overline{\mathbb{Q}} \cap \mathbb{R}$. To this end we introduce the so-called multiplicative absolute Weil height:

$$H : \mathbb{Q} \to [1, \infty)$$

defined by

$$H(\alpha) = M(D_\alpha(x))$$

where $D_\alpha(x) = a_0(x - \alpha_1)\dots(x - \alpha_d) \in \mathbb{Z}[x]$ is the minimal polynomial of $\alpha$ and

$$M(D_\alpha(x)) = |a_0| \cdots \prod_{i=1}^{n} \max\{1, |\alpha_i|\}$$

$M$ is called the Mahler-measure.

**Example.** 
- $\alpha = \frac{p}{q} \in \mathbb{Q}$ *($q > 0$, $\gcd(p,q) = 1$).*
  $\deg(\alpha) = 1$*, $D_\alpha = qx - p$. So* $H(\alpha) = M(D_\alpha) = q \max\{1, \left|\frac{p}{q}\right|\} = \max\{q, |p|\} = H_{\mathbb{P}^1}((1, \alpha))$.

- $\alpha = 2^{\frac{1}{d}}$*, $D_\alpha = x^d - 2$ (2-Eisenstein),* $\deg \alpha = d$*,* $H(\alpha) = M(D_\alpha)^{\frac{1}{d}} = \prod_{i=1}^{d} \max\{1, \left|\xi_d^{i-1} 2^{\frac{1}{d}}\right|\} = 2^{\frac{1}{d}}$

One can easily show (c.f. sheet 4) that

$$\#\{\alpha \in \overline{\mathbb{Q}} : \deg d \le d, H(\alpha) \le X\} < \infty \ \forall d \in \mathbb{N} \ X \ge 1. \tag{48}$$

Back to Diophantine approximation with $A = \mathbb{Q}$. As

$$a + m - \frac{p}{q} = a - \left(\frac{p - mq}{q}\right)$$

we can assume $\alpha \in (0,1)$ So all good enough approximations $\frac{p}{q}$ lie also in $(0,1)$. Now if $\frac{p}{q} \in (0,1)$ then

$$H\left(\frac{p}{q}\right) = q.$$

So

$$\left|\alpha - \frac{p}{q}\right| < \phi(q) \iff \left|\alpha - \frac{p}{q}\right| < \phi\left(H\left(\frac{p}{q}\right)\right).$$

So now the denominator plays no role any more and we can write more easily:

$$|\alpha - x| < \phi(H(x)) \tag{49}$$

(49) makes sense as long as $x - \alpha \in \mathbb{R}$, so $x \in \mathbb{R}$, and $x \in \overline{\mathbb{Q}}$. So let's assume $A \subset \overline{\mathbb{Q}} \cap \mathbb{R}$. However, if $x_1, x_2, x_3, \ldots$ is a sequence of pairwise distinct solutions of (49) then we want to conclude that $x_i \to \alpha$ (with respect to $|\bullet|$). Now as $\phi(t) \to 0$ as $t \to \infty$ but we don't know a priori that $H(x_i) \to \infty$. So cannot conclude from (49) that $x_i \to \alpha$. But if $A \subset \mathbb{Q}_{(d)} = \{\alpha \in \overline{\mathbb{Q}} : \deg \alpha \le d\}$ then (48) tells us that $H(x_i) \to \infty$ and so $x_i \to \alpha$.

More generally this is true if

$$\#\{\alpha \in A : H(\alpha) \le X\} < \infty \ \forall X \ge 1.$$

In this case we say $A$ has property $N$.

**Theorem 1.6.3** (1.6.3 Wirsing 1961 )**.** *Let $d \in \mathbb{N}$, $d > 1$ and $\alpha \in \mathbb{R} \setminus \mathbb{Q}_{(d)}$. Then $\exists \infty$-many $x \in \mathbb{Q}_{(d)}$ with*

$$|\alpha - x| < H(x)^{-(\frac{d+3}{2})}$$

**Conjecture 1.6.4** (1.6.4 Wirsing's Conjecture, around 1961 )**.** *Suppose $\alpha \in \mathbb{R} \setminus \mathbb{Q}_{(d)}$, $(d \in \mathbb{N})$ and $\lambda < d+1$ then $\exists \infty$-many $x \in \mathbb{Q}_{(d)}$ with*

$$|\alpha - x| < H(x)^{-\lambda}.$$

**Theorem 1.6.5** (1.6.5 Davenport and Schmidt )**.** *Wirsing's conjecture (1.6.4) holds for $d \le 2$.*

Instead of taking $A = \mathbb{Q}_{(d)} \cap \mathbb{R}$ let's replace $\mathbb{Q}_{(d)}$ with the smallest field that contains $\mathbb{Q}_{(d)}$; let's call this field $\mathbb{Q}^{(d)}$. Unfortunately, nobody knows whether $\mathbb{Q}^{(d)}$ has property $(N)$, except when $d \le 2$.

**Theorem 1.6.6** (1.6.6 Bombien-Zannier, 2001 )**.** *$\mathbb{Q}^{(2)}$ has Property $(N)$.*

**Open Problem 1.6.7** (1.6.7)**.** *Find an analogue of Corollary 1.1.2 for $A = \mathbb{Q}(2) \cap \mathbb{R}$. How quickly can $\phi : [1, \infty) \to (0, \infty)$ decay if for every $\alpha \in \mathbb{R} \setminus A$*

$$|\alpha - x| < \phi(H(x))$$

*has $\infty$-many solutions $x \in A$? It is not difficult to show an inequality in the other direction provided $\alpha$ is algebraic, e.g., if $\alpha \in \overline{\mathbb{Q}} \setminus \mathbb{Q}^{(2)}$ then*

$$|\alpha - x| > (2 \cdot H(\alpha) H(x))^{-\deg \alpha \cdot 2^{(2H(x))}}$$

*How much can this be improved?*

# 2 Geometry of Numbers

References:

- J.W.S. Cassels "An Introduction to the Geometry of Numbers"
- W.M. Schmidt Lecture Notes M. 785 and 1467

## 2.1 Basic notions

Let $R$ be an integral domain (with 1), and $n \in \mathbb{N}$. We write $\mathrm{Mat}_n(R) = \{n \times n$ matrices with entries in $R\}$ and $\mathrm{GL}_n(R) = \{A \in \mathrm{Mat}_n(R) : A$ is invertibel and $A^{-1} \in \mathrm{Mat}_n(R)\}$. Then $(\mathrm{GL}_n(R), \cdot)$ is a group. We say $u \in R$ is a unit (in $R$) if $\exists u' \in R$ such that $u' \cdot u = 1$.
If $A \in \mathrm{GL}_n(R)$ then $(\det A)^{-1} = \det A^{-1} \in R$. So $\det A$ is a unit in $R$. On the other hand if $A \in \mathrm{Mat}_n(R)$ and $\det A$ is a unit in $R$ then $A^{-1} = (\det a)^{-1} \operatorname{adj}(A) \in \mathrm{Mat}_n(R)$ as the adjungate matrix $\operatorname{adj}(A)$ of $A$ clearly is in $\mathrm{Mat}_n(R)$. Thus we have

$$\mathrm{GL}_n(R) = \{A \in \mathrm{Mat}_n(R) : \det A \text{ is a unit in } R\}.$$

In particular, $\mathrm{GL}_n(\mathbb{Z}) = \{A \in \mathrm{Mat}_n(\mathbb{Z}) : \det A = \pm 1\}$.

Let $n \in \mathbb{N}$. A <u>lattice</u> $\Lambda$ in $\mathbb{R}^n$ is a set of the form

$$\Lambda = A\mathbb{Z}^n = \{Ax : x \in \mathbb{Z}^n\}$$

where $A \in \mathrm{GL}_n(\mathbb{R})$. The column vectors of $A$ are called a <u>basis</u> of $\Lambda$.

**Lemma 2.1.1** (2.1.1 )**.** *Let $A, B \in \mathrm{GL}_n(\mathbb{R})$. Then*

$$A\mathbb{Z}^n = B\mathbb{Z}^n \iff \exists T \in \mathrm{GL}_n(\mathbb{Z}) \text{ such that } B = AT$$

*Proof.* "$\Leftarrow$" If $B = AT$ with $T \in \mathrm{GL}_n(\mathbb{Z})$ then $T\mathbb{Z}^n = \mathbb{Z}^n$. Hence, $B\mathbb{Z}^n = A\mathbb{Z}^n$.
"$\Rightarrow$" If $A\mathbb{Z}^n = B\mathbb{Z}^n$ then each column vector of $B$ lies in $A\mathbb{Z}^n$, thus $\exists T \in \mathrm{Mat}_n(\mathbb{Z})$ such that $B = AT$. Similarly $\exists T' \in \mathrm{Mat}_n(\mathbb{Z})$ such that $A = BT'$. Hence, $A = ATT'$. Thus $T' = T^{-1}$. So $T \in \mathrm{GL}_n(\mathbb{Z})$. ☐

By Lemma 2.1.1 we see that if $\Lambda = A\mathbb{Z}^n$ then $|\det A|$ is uniquely determined by $\Lambda$. We call it the <u>determinant</u> of $\Lambda$

$$\det \Lambda = |\det A|$$

Let $b_1, \dots, b_n$ be a basis of $\Lambda$ and $v \in \Lambda$. We set

$$F_v = [0,1) \cdot b_1 + \cdots + [0,1) \cdot b_n + v$$

and call it a fundamental cell of $\Lambda$.
Note that

- $\det \Lambda = \mathrm{Vol}\, F$
- $\mathbb{R}^n = \cap_{v \in \Lambda}^{disjointwith \bullet}$ is a partition of $\mathbb{R}^n$ (cf sheet 5).

Recall that $C \subset \mathbb{R}^n$ is called <u>convex</u> if:

$$x, y \in C \implies tx + (1-t)y \in C \forall t \in [0,1]$$

We say $C$ is symmetric if:

$$x \in C \implies -x \in C$$

Recall that every convex set is measurable.

Let $C$ be a convex, compact and symmetric set in $\mathbb{R}^n$ with the origin in the interior of $C$. Let $\Lambda$ be a lattice in $\mathbb{R}^n$. Then we define the successive minima $\lambda_1, \dots, \lambda_n$ of $\Lambda$ with respect to $C$ by

$$\lambda_i = \inf\{\lambda : \lambda C \cap \Lambda \text{ contains } i \text{ linearly independent vectors}\}.$$

Note that $0 < \lambda_1 \le \lambda_2 \le \dots \le \lambda_n < \infty$.

**Example.**

- *$\Lambda = \mathbb{Z}^n$, $C = [-1,1]^n$. Then $\lambda_1 = \dots = \lambda_n = 1$.*
- *$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \mathbb{Z}^2$, $C = [-1,1]^2$. Then $\lambda_1 = 1$, $\lambda_2 = 2$.*
- *$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \mathbb{Z}^2$, $C = [-1,1] \times [-2,2]$. Then $\lambda_1 = \lambda_2 = 1$.*

**Theorem 2.1.2.** *Let $\Lambda$ be in $\mathbb{R}^n$. Then $\Lambda$ is a lattice in $\mathbb{R}^n$ if and only if:*

*i) $(\Lambda, +)$ is a group*

*ii) $\Lambda$ contains $n$ linearly independent vectors*

*iii) $\Lambda$ is discrete, $\#S \cap \Lambda < \infty \forall$ compact $S \subset \mathbb{R}^n$.*

*Proof.* First suppose $\Lambda$ is a lattice. Then i) and ii) are clear. And iii) is clear, at least if $\Lambda = \mathbb{Z}^n$. But if $\Lambda = A\mathbb{Z}^n$ then $\#\Lambda \cap S = \#\mathbb{Z}^n \cap A^{-1}S$ as $x \mapsto A^{-1}x$ is continuous we have $S$ compact $\implies A^{1-}S$ compact. So this proves the first direction.
Now let's suppose $\Lambda \subset \mathbb{R}^n$ such that i), ii), iii) hold. We use induction on $n$.
Let $n = 1$. Then $\Lambda$ contains a non-zero vector $b$ that is closest to the origin (using ii) and iii)). By i) we easily see that $\Lambda = b\mathbb{Z}$. So $\Lambda$ is a lattice in $\mathbb{R}^1$.
Now suppose the claim holds in $\mathbb{R}^m$ if $m < n$. Let $u_1, \dots, u_n$ be $n$ linearly independent vectors in $\Lambda$. Consider the subspace $U = \langle u_1, \dots, u_{n-1} \rangle_{\mathbb{R}}$; thus $\dim U = n-1$. Let $\tilde{e_1}, \dots, \tilde{e_{n-1}}$ be an orthonormal basis of $U$. Let

$$O \in \mathrm{O}_n(\mathbb{R}) = \{A \in \mathrm{GL}_n(\mathbb{R}) : A^T A = I_n\} \text{ (the orthogonal group)}$$

with

$$O(\tilde{e_i}) = e_i \; (1 \le i \le n-1)$$

where $e_1, \dots, e_n$ is the canonical basis of $\mathbb{R}^n$. Hence, $O(U) = \mathbb{R}^{n-1} \times \{0\}$, and

$$O(\Lambda \cap U) \subset \mathbb{R}^{n-1} \times \{0\}$$

is a discrete additive group that contains the $n-1$ linearly independent vectors $O(u_1), \dots, O(u_{n-1})$. Let $\Pi : \mathbb{R}^n \to \mathbb{R}^{n-1}$, $\Pi(x) = (x_1, \dots, x_{n-1})$. Then $\Pi \circ O(\Lambda \cap U)$ is also a discrete additive group that contains $n-1$ linearly independent vectors and it is also in $\mathbb{R}^{n-1}$. Hence, by induction hypothesis $\Pi \circ O(\Lambda \cap U)$ is a lattice in $\mathbb{R}^{n-1}$. So $\exists A_{n-1} \subset \mathrm{GL}_{n-1}(\mathbb{R})$ such that $\Pi \circ O(\Lambda \cap U) = A_{n-1}\mathbb{Z}^{n-1}$. So

$$O(\Lambda \cap U) = \left( \underbrace{\begin{matrix} A_{n-1} & 0 \\ 0 & 0 \end{matrix}}_{} \tilde{A} \right) \mathbb{Z}^n$$

Now let $\mu = \inf\{|w_n| : w = (w_1,\dots,w_n) \in O(\Lambda \setminus U)\}$. Suppose $v_1, v_2, v_3, \dots \in O(\Lambda \setminus U)$ with

$$|v_{in}| \to \mu \text{ (as } i \to \infty$$

where $v_{in}$ is the last coordinate.

Adding elements from $O(U)$ does not change the last coordinate. Hence, we can assume that the vectors

$$(v_{i1},\dots,v_{in-1}) \in [0,1)a_1 + \dots + [0,1)a_{n-1}$$

where $a_i$ = column vector of $A_{n-1}$. In particular, the first $n-1$ coordinates of $v_i$ are bounded in absolute value. But the absolute value of the last coordinate also tends to $\mu$, so is also bounded. As $\Lambda$ is discrete by iii) also $O(\Lambda)$ is discrete. Thus the sequence $v_i$ contains only finitely many vectors, in particular

$$\exists v \in O(\Lambda \setminus U) \text{ such that } v_n \overset{after v \to -v}{=} \mu \text{ and } \mu > 0.$$

Let $u \in O(\Lambda)$. Then also $u' = u - \left[\frac{u_n}{\mu}\right] \cdot v$ is in $O(\Lambda)$ ($O(\Lambda)$ is a group as $\Lambda$ is). So $0 \le u'_n < \mu$, so by minimality of $\mu$, $u'_n = 0$. Hence, $u' \in O(\Lambda) \cap \mathbb{R}^{n-1} \times \{0\} = O(\Lambda) \cap O(U) = O(\Lambda \cap U)$. Now

$$u = u' + \left[\frac{u_n}{\mu}\right] \cdot v \in O(\Lambda \cap U) + \mathbb{Z} \cdot v$$

and thus

$$u \in \tilde{A}\mathbb{Z}^n + v \cdot \mathbb{Z} = \underbrace{[(\tilde{a_1}) \dots (\tilde{a_{n-1}})(v)]}\mathbb{Z}^n = A\mathbb{Z}^n$$

where $\tilde{a_i}$ = $i$-th column vector of $\tilde{A}$. Thus $O(\Lambda) \subset A\mathbb{Z}^n$. Clearly (as $O(\Lambda)$ is a group) also $A\mathbb{Z}^n \subset O(\Lambda)$. Now the rows of $A$ are linearly independent. Thus $A \in \mathrm{GL}_n(\mathbb{R})$. So $O(\Lambda)$ is a lattice and thus $\Lambda$ is a lattice. □

**Corollary 2.1.3** (2.1.3 )**.** *Let $n \in \mathbb{N}$, $m_1,\dots,m_n \in \mathbb{N}$ and $a_{ij} \in \mathbb{Z}$ $(1 \le i,j \le n)$. Then*

$$\Lambda = \{x \in \mathbb{Z}^n : \sum_{j=1}^{n} a_{ij}x_j \equiv 0 \bmod m_i (1 \le i \le n)\}$$

*is a lattice in $\mathbb{R}^n$.*

*Proof.* As $\Lambda \subset \mathbb{Z}^n$, it is discrete. Clearly $0 \in \Lambda$, and if $x_1, x_2 \in \Lambda$ then $x_1 + x_2 \in \Lambda$. So $\Lambda$ is a discrete additive subgroup of $\mathbb{Z}^n$. Moreover, the $n$ linearly independent vectors $(m,0,\dots,0),\dots,(0,\dots,0,m)$ where $m = m_1 \cdot \dots \cdot m_n$ are all in $\Lambda$. Hence, by Theorem 2.1.2 we conclude that $\Lambda$ is a lattice in $\mathbb{R}^n$. □

## 2.2 The Theorems of Blichfeldt and Minkowski

Minkowski's First and Second Theorem are possibly some of the most useful theorems in number theory. We will deduce Minkowski's First Theorem via Blichfeldt's Theorem which is of interest for its own sake.

**Theorem 2.2.1** (Blichfeldt )**.** *Let $\Lambda$ be a lattice in $\mathbb{R}^n$, and let $S \subset \mathbb{R}^n$ be measurable such that $\operatorname{Vol} S > \det \Lambda$ ($\operatorname{Vol} S = \infty$ is allowed). Then*

$$\exists x_1, x_2 \in S, x_1 \ne x_2 \text{ and } x_1 - x_2 \in \Lambda.$$

*Proof.* Let $b_1, \dots, b_n$ be a basis of $\Lambda$ and let $F = [0,1)b_1 + \dots + [0,1)b_n$ be a fundamental cell. Thus $\operatorname{Vol} F = \det \Lambda$, and if $x \in \mathbb{R}^n$ then $\exists$ unique $v \in \Lambda$ and $\theta \in F$ such that

$$x = V + \theta.$$

Now for each $v \in \Lambda$ consider

$$\mathcal{R}(v) = \{\theta \in F : v + \theta \in S\}$$

Hence,

$$\sum_{v \in \Lambda} \operatorname{Vol}(\mathcal{R}(v)) = \operatorname{Vol} S$$

Now if $\operatorname{Vol} S > \det \Lambda$ then (2.1) implies $\sum_{v \in \Lambda} \operatorname{Vol}(\mathcal{R}(v)) > \det \Lambda = \operatorname{Vol} F$. But $\cup_{v \in \Lambda} \mathcal{R}(v) \subset F$; so the union cannot be disjoint. Thus $\exists v_1, v_2 \in \Lambda, v_1 \neq v_2$ such that $\theta_0 \in \mathcal{R}(v_1) \cap \mathcal{R}(v_2)$. Hence, the points $x_1 = v_1 + \theta_0, x_2 = v_2 + \theta_0$ are both in $S$ and $x_1 - x_2 = v_1 - v_2 \in \Lambda \setminus 0$. $\square$

**Theorem 2.2.2** (Mnkowski's First Theorem )**.** *Let $\Lambda$ be a lattice in $\mathbb{R}^n$, and let $S \subset \mathbb{R}^n$ be convex and symmetric. Suppose that either*

- $\operatorname{Vol} S > 2^n \det \Lambda$ *($\operatorname{Vol} S = \infty$ allowed)*
  *or*
- $\operatorname{Vol} S \geq 2^n \det \Lambda$ *and $S$ is compact*

*Then $S$ contains a non-zero lattice point.*

**Remark.** *$2^n$ is sharp, take $\Lambda = \mathbb{Z}^n$ and $S = (-1,1)^n$, then $\operatorname{Vol} S = 2^n$, $\det \Lambda = 1$, $S$ is symmetric and convex, but $S \cap \Lambda = \{0\}$.*

*Proof.* First suppose $\operatorname{Vol} S > 2^n \cdot \det \Lambda$. Now $\operatorname{Vol}\left(\frac{1}{2}S\right) = 2^{-n} \operatorname{Vol} S > \det \Lambda$. By Theorem 2.2.1 (with the set $\frac{1}{2}S$) we see that $\exists x_1, x_2 \in S$, $x_1 \neq x_2$ such that $\frac{1}{2}x_1 - \frac{1}{2}x_2 \in \Lambda \setminus 0$. But $S$ is symmetric, thus $-x_2 \in S$. As $S$ is convex we conclude that $\frac{1}{2}x_1 + \frac{1}{2}(-x_2) \in S$. This proves the first part.
Now suppose $S$ is compact and $\operatorname{Vol} S = 2^n \det \Lambda$. If $v \in \Lambda \setminus S$ then $\exists \varepsilon_v > 0$ such that $B_{\varepsilon_v}(v) \cap S = \varnothing$ ($S^c$ is open!) where $B_r(y) = \{x \in \mathbb{R}^n : |x - y| < r\}$. As $S$ is compact $\exists R > 0$ such that $\lambda S \subset B_R(0)$ for all $\lambda$ with $0 < \lambda < 2$. So $(\Lambda \setminus S) \cap B_R(0)$ is finite by Theorem 2.1.2 and hence $\exists \varepsilon > 0$ such that

$$B_\varepsilon(v) \cap S = \varnothing \ \forall v \in (\Lambda \setminus S) \cap B_R(0).$$

Hence, $\exists \lambda > 1$ such that

$$\lambda S \cap \Lambda = S \cap \Lambda.$$

By the first part we know that $\lambda S$ contains a non-zero lattice point, and this completes the proof. $\square$

**Corollary 2.2.3.** *Let $\Lambda$ be a lattice in $\mathbb{R}^n$ and let $a_{ij} \in \mathbb{R}$ $(1 \leq i, j \leq n)$. Suppose $c_1, \dots, c_n > 0$ and $c_1 \dots c_n \geq |\det A| \det \Lambda$. Then $\exists u \in \Lambda \setminus 0$ such that (2.2)*
$$\begin{cases} \left|\sum_{j=1}^n a_{ij} u_j\right| \leq c_1 \\ \left|\sum_{j=1}^n a_{ij} u_j\right| < c_i \ (2 \leq i \leq n). \end{cases}$$

*Proof.* First suppose $\det A \neq 0$. Then $\mathcal{L} = A - \Lambda$ is a lattice in $\mathbb{R}^n$ with $\det \mathcal{L} = |\det A| \det \Lambda$. Then (2.2) means we are looking for a non-zero lattice point $x \in \mathcal{L}$ such that

$$|x_1| \leq c_1$$

$$|x_i| < c_i \ (2 \leq i \leq n)$$

These inequalities define a symmetric, convex set of points $x \in \mathbb{R}^n$ with volume $2^n c_1 \cdots c_n$. So if $c_1 \cdots c_n > |\det(A)| \cdot \det(\Lambda)$ then we can apply Theorem 2.2.2 and the claim follows at once.
Next let $0 < \varepsilon < 1$. Then the set

$$S_\varepsilon : \begin{cases} |x_1| \leq c_1 + \varepsilon < c_1 + 1 & \\ |x_i| < c_i & \text{for } (2 \leq i \leq n) \end{cases}$$

still has a non-zero lattice point in $\mathcal{L}$. But these sets $S_\varepsilon$ all lie in $S_1$ which lies in a compact set and hence has only finitely many lattice points. Hence, there must be a non-zero lattice point of $\mathcal{L}$ in $S_0$. This proves the Corollary if $\det(A) \neq 0$.
Now if $\det(A) = 0$ then (2.2) defines a set of points $u \in \mathbb{R}^n$ of infinite volume and so Theorem 2.2.2 applies again and yields the claim. □

**Corollary 2.2.4** (Lagrange's four-square Theorem)**.** *Every positive integer is the sum of four squares.*

*Proof.* First we observe that

$$\begin{aligned}
(x_1^2 + x_2^2 + x_3^2 + x_4^2) \cdot (y_1^2 + y_2^2 + y_3^2 + y_4^2) = \\
&= (x_1 y_1 + x_2 y_2 x + x_3 y_3 + x_4 y_4)^2 + (-x_1 y_2 + x_2 y_1 - x_3 y_4 + x_4 y_3)^2 \\
&+ (-x_1 y_3 + x_2 y_4 + x_3 y_1 - x_{-} 2 y_2)^2 + (-x_1 y_4 - x_2 y_3 + x_3 y_2 + x_4 y_1)^2.
\end{aligned}$$

Now $1 = 1^2 + 0^2 + 0^2 + 0^2$. So it suffices to prove the claim for primes $p$. And we can assume $p \neq 2$ since $2 = 1^2 + 1^2 + 0^2 + 0^2$.
Now $a^2$ and $-(b^2 + 1)$ run through exactly $\frac{p+1}{2}$ distinct residue classes modulo $p$ as $a \bmod b$ run through an entire system of residue classes. $(0^2, 1^2, \ldots, \left(\frac{p-1}{2}\right)^2$ are all distinct in $\mathcal{F}_p$.)
Hence, they have a common residue class; thus

$$\exists a, b \in \mathbb{Z} \text{ such that } a^2 + b^2 + 1 \equiv 0 \bmod p.$$

With this choice of $a$ and $b$ consider

$$\Lambda = A \cdot \mathbb{Z}^4 \text{ where } A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a & b & p & 0 \\ b & -a & 0 & p \end{pmatrix}.$$

So $\Lambda$ is a lattice in $\mathbb{R}^4$ with $\det(\Lambda) = p^2$.
Next consider the convex, symmetric set

$$S = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 : x_1^2 + x_2^2 + x_3^2 + x_4^2 < 2p\}$$

Then $\mathrm{Vol}(S) = \frac{\pi^2}{2}(2p)^4 = (2\pi)^2 p^4 > 16 \cdot p^2 = 2^4 \det(\Lambda)$. By Theorem 2.2.2 there exists an $x \in \Lambda \cap S$ with $x \neq 0$.
Now

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = A \cdot z = \begin{pmatrix} z_1 \\ z_2 \\ az_1 + bz_2 + pz_3 \\ bz_1 - az_2 + pz_4 \end{pmatrix} \text{ for some } z \in \mathbb{Z}^4 \setminus \{0\}$$

Hence,

$$\begin{aligned} x_1^2 + x_2^2 + x_3^2 + x_4^2 &\equiv z_1^2 + z_2^2 + (az_1 + bz_2)^2 + (bz_1 - az_2)^2 \\ &\equiv \underbrace{(1 + a^2 + b^2)}_{\equiv 0 \bmod p}(z_1^2 + z_2^2) \\ &\equiv 0 \bmod p \end{aligned}$$

Since $x \in S$, and $x \neq 0$ we conclude that $x_1^2 + x_2^2 + x_3^2 + x_4^2 = p$. This proves the corollary. □

## 2.3 3 Basis reduction

Let $M$ be a lattice in $\mathbb{R}^n$ and let $\Lambda$ be a sublattice of $M$, i.e., $\Lambda \subset M$ and $\Lambda$ is a lattice in $\mathbb{R}^n$. Hence, there exists a matrix $C \in \mathrm{Mat}_n(\mathbb{Z})$ with $\det C \neq 0$ such that $\Lambda = C \cdot M$.
We define the index of $\Lambda$ in $M$ by

$$I = |\det(C)| = \frac{\det(\Lambda)}{\det(M)}.$$

Note that $I \cdot C^{-1} = \mathrm{adj}(C) \in \mathrm{Mat}_n(\mathbb{Z})$, and thus $I \cdot M = I \cdot C^{-1} \cdot \Lambda \subset \Lambda$, so

$$I \cdot M \cdot C \subset \Lambda \subset M \tag{50}$$

**Theorem 2.3.1.** *Let $\Lambda$ be a sublattice of the lattice $M \subset \mathbb{R}^n$, and let $b_1, \dots, b_n$ be a basis of $M$. Then there exists a basis $a_1, \dots, a_n$ of $\Lambda$ with*

$$\begin{aligned} a_1 &= v_{11} b_1 \\ a_2 &= v_{21} b_1 + v_{22} b_2 \\ &\vdots \\ a_n &= v_n 1 b_1 + \cdots + v_{nn} b_n \end{aligned} \tag{51}$$

*with $v_{ij} \in \mathbb{Z}$ and $v_{ii} \neq 0$ for $(1 \leq j \leq i \leq n)$.*
*Conversely, if $a_1, \dots, a_n$ is a basis of $\Lambda$ then there exists a basis $b_1, \dots, b_n$ of $M$ such that (51) holds.*

*Proof.* By (50) we know that there exist $v_{ij} \in \mathbb{Z}$ with $v_{ii} \neq 0$ and $|v_{ii}|$ minimal such that

$$a_i = v_{i1} b_1 + \cdots + v_{ii} b_i \in \Lambda$$

We will show that $a_1, \dots, a_n$ is a basis for $\Lambda$.
Let $c \in \Lambda$ and suppose $c$ is not a $\mathbb{Z}$-linear combination of the $a_i$'s. As $c \in M$ there exist $t_i \in \mathbb{Z}$ such that

$$c = t_1 b_1 + \cdots + t_k b_k \ (1 \leq k \leq n,\ t_k \neq 0)$$

If there exist several such $c$'s then we choose one where $k$ is minimal. Next we note that $v_{kk} \neq 0$. Hence, there exists a $s \in \mathbb{Z}$ such that

$$|t_k - sv_{kk}| < |v_{kk}|. \tag{52}$$

Thus

$$c - sa_k = (t_1 - sv_{k1})b_1 + \cdots + (t_k - sv_{kk})b_k$$

lies in $\Lambda$ (as $c$ and $a_k$ do) and is <u>not</u> a $\mathbb{Z}$-linear combination of the $a_i$'s as $c$ is not. Thus, by minimality of $k$ we must have $t_k - sv_{kk} \neq 0$. But then (52) contradicts the minimality of $|v_{kk}|$. Hence, $c$ must be a $\mathbb{Z}$-linear combination of the $a_i$'s and thus $a_1, \ldots, a_n$ is a basis. This proves the first part.
For the second part suppose $a_1, \ldots, a_n$ is a basis of $\Lambda$. By the first part and (50) there exists a basis $I \cdot b_1, \ldots, Ib_n$ of $I \cdot M \subset \Lambda$ with

$$\begin{aligned}
Ib_1 &= w_{11}a_1 \\
Ib_2 &= w_{21}a_1 + w_{22}a_2 \\
&\vdots \\
Ib_n &= w_{n1}a_1 + \cdots + w_{nn}a_n
\end{aligned}$$

with $w_{ij} \in \mathbb{Z}$ and $w_{ij} \neq 0$.
Successively solving the above system for $a_i$ we get a system as in (51) but a priori with $v_{ij} \in \mathbb{Q}$. But $b_1, \ldots, b_n$ is a basis of $M$ and the $a_i \in M$. As the representation

$$a = t_1 b_1 + \cdots + t_n b_n \; (t_i \in \mathbb{R})$$

is unique we conclude that $v_{ij} \in \mathbb{Z}$, and this proves the second part. □

**Lemma 2.3.2** (Hadamard's inequality)**.** *Let $a_1, \ldots, a_n \in \mathbb{R}^n$. Then*

$$|\det(a_1, \ldots, a_n)| \leq |a_1| \cdots |a_n|.$$

*Proof.* This is geometrically obvious as the volume of a parallelepiped is no larger than product of the lengths of the spanning vectors. However, here is a formal proof.
If $a_1, \ldots, a_n$ are linearly dependent then the inequality is trivial. Now assume $a_1, \ldots, a_n$ are linearly independent. Put

$$c_i = a_i - \sum_{j<i} a_i c_j |c_j|^{-2} \cdot c_j$$

Then

$$c_i \cdot c_j = 0 \; (i \neq j) \tag{53}$$

and

$$a_i = t_{i1}c_1 + \cdots + t_{ii-1}c_{i-1} + c_i \tag{54}$$

By (53) and (54) we get

$$|a_i|^2 = a_i a_i = (\sum_{j=1}^{i-1} t_{ij}^2 |c_j|^2) + |c_i|^2 \geq |c_i|^2$$

and $\det(a_1, \dots, a_n) = \det(c_1, \dots, c_n)$ (by linearity of determinant in columns). Moreover,

$$(\det(c_1, \dots, c_n))^2 = \det\left(\begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}\begin{bmatrix} c_1 & \cdots & c_n \end{bmatrix}\right) = \prod_{i=1}^{n} |c_i|^2 \le \prod_{i=1}^{n} |a_i|^2.$$

□

**Definition.** *A* <u>*distance function*</u> *$f$ on $\mathbb{R}^n$ is a function $f : \mathbb{R}^n \to \mathbb{R}$ such that*

- $f(x) \ge 0 \ \forall x \in \mathbb{R}^n$
- $f(tx) = |t| f(x) \ \forall x \in \mathbb{R}^n \ \forall t \in \mathbb{R}$
- *$f$ is continuous.*

**Definition.** *We say $C$ is a star body in $\mathbb{R}^n$ if*

- *$C \subset \mathbb{R}^n$ compact*
- *$0 \in \mathrm{Int}(C)$, i.e., origin lies in the interior of $C$*
- $x \in C \implies t \cdot x \in C \ (0 \le t \le 1)$

**Remark.**

- *For a star body $C$ we have $t \cdot C \subset C$ for $0 \le t \le 1$.*
- *Every compact, convex $C \subset \mathbb{R}^n$ with the origin in its interior is a star body in $\mathbb{R}^n$.*
- *To every symmetric star body $C$ in $\mathbb{R}^n$ we can associate a distance function $f_C$ defined by*
$$f_C(x) = \inf\{\lambda : x \in \lambda \cdot C\}.$$
*Note that $f_C(x) = 0 \implies x = 0$.*
*Why? If $x \ne 0$, then there exists a $\lambda > 0$ such that $\lambda x \notin C$. Hence, $f_C(x) \ge \frac{1}{\lambda}$.*
- *If $C$ is symmetric and convex then $f_C$ is actually a norm on $\mathbb{R}^n$ (cf exercise sheet 5). In particular, $f_C$ satisfies the triangle-inequality.*

**Lemma 2.3.3.** *Let $C$ be a convex, symmetric star body in $\mathbb{R}^n$, and let $\Lambda$ be a lattice in $\mathbb{R}^n$ with successive minima $\lambda_1, \dots, \lambda_n$ with respect to $C$. Then there exist linearly independent $a_1, \dots, a_n \in \Lambda$ with $f_C(a_i) = \lambda_i$.*
*Moreover, if $a \in \Lambda$ and $f_C(a) < \lambda_j$ then $a_1, \dots, a_{j-1}, a$ are linearly dependent.*

*Proof.* The set $(\lambda_n + 1) \cdot C$ is compact and by definition of $\lambda_n$ contains $n$ linearly independent lattice points. By the definition of the $\Lambda_i$'s it suffices to consider these points. But by Theorem **??** there are only finitely many of these, and so the claim easily follows. □

**Corollary 2.3.4.** *Let $C$ be a convex, symmetric star body in $\mathbb{R}^n$, and let $\Lambda$ be a lattice in $\mathbb{R}^n$ with successive minima $\lambda_1, \dots, \lambda_n$ with respect to $C$. Then there exists a basis $b_1, \dots, b_n$ of $\Lambda$ such that for $j = 1, \dots, n$:*

$$x \in \Lambda \text{ and } f_C(x) < \lambda_j \implies x = u_1 b_1 + \dots + u_{j-1} b_{j-1}$$

*for some $u_1, \dots, u_{j-1} \in \mathbb{Z}$.*

*Proof.* Let $a_1, \ldots, a_n \in \Lambda$ be as in Lemma 2.3.3. Let $\Lambda' = (a_1, \ldots, a_n)\mathbb{Z}^n$ be the sublattice of $\Lambda$ with basis $a_1, \ldots, a_n$. By Theorem 2.3.1 there exists a basis $b_1, \ldots, b_n$ of $\Lambda$ with (51); so $a_j$ is dependent only on $b_1, \ldots, b_j$. By Lemma 2.3.3 if $f_C(x) < \lambda_j$, then

$$\begin{aligned} x &= s_1 a_1 + \cdots + s_{j-1} a_{j-1} \\ &= u_1 b_1 + \cdots + u_{j-1} b_{j-1} \end{aligned}$$

with $u_i \in \mathbb{Q}$.
As $x \in \Lambda$ and $b_1, \ldots, b_{j-1}$ are linearly independent we conclude that $u_1, \ldots, u_{j-1} \in \mathbb{Z}$. □

**Example** (Exercise)**.** *Let* $C = B_0(1)$ *and* $\Lambda = \begin{pmatrix} 2 & 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \mathbb{Z}^5$.
*There exists* <u>*no*</u> *basis* $b_1, \ldots, b_5$ *such that* $|b_i| = \lambda_i$ $(1 \le i \le 5)$ *with* $\lambda_i = \lambda_i(\Lambda, C)$.

**Lemma 2.3.5.** *Let $C$ be a convex, symmetric star body in $\mathbb{R}^n$, let $\Lambda$ be a lattice in $\mathbb{R}^n$, and let $a_1, \ldots, a_n$ be linearly independent vectors in $\Lambda$. Then there exists a basis $b_1, \ldots, b_n$ of $\Lambda$ such that*

$$f_C(b_j) \le \max\left\{ f_C(a_j), \frac{1}{2}(f_C(a_1) + \cdots + f_C(a_j)) \right\}$$

*Proof.* Consider the sublattice $\Lambda' = (a_1 \ldots a_n)\mathbb{Z}^n \subset \Lambda$. By Theorem 2.3.1 there exists a basis $c_1, \ldots, c_n$ of $\Lambda$ such that

$$\begin{aligned} a_1 &= v_{11} c_1 \\ a_2 &= v_{21} c_1 + v_{22} c_2 \\ &\vdots \\ a_n &= v_{n1} c_1 + \cdots + v_{nn} c_n \end{aligned} \tag{55}$$

with $v_{ij} \in \mathbb{Z}$ and $v_{ii} \ne 0$.
Consider

$$b_j = c_j + t_{jj-1} a_{j-1} + \cdots + t_{j1} a_1 \tag{56}$$

where $t_i \in \mathbb{R}$.
If $b_1, \ldots, b_n$ are in $\Lambda$ then by (55) they form a basis of $\Lambda$.
How do we choose $t_{ji}$? If $v_{jj} = \pm 1$ then we put $b_j = \pm a_j$, which clearly is in the required form (56) and obviously

$$f_C(b_j) = f_C(a_j).$$

Now suppose $|v_{jj}| \ge 2$. Now solving (55) for $c_j$ yields

$$c_j = v_{jj}^{-1} a_j + k_{jj-1} a_{j-1} + \cdots + k_{j1} a_1$$

with $k_{ji} \in \mathbb{Q}$.
We choose $t_{ji} \in \mathbb{Z}$ such that

$$|k_{ji} + t_{ji}| \le \frac{1}{2}.$$

Then $b_j \in \Lambda$ and

$$b_j = l_{jj}a_j + l_{jj-1}a_{j-1} + \cdots + l_{j1}a_1$$

with

$$\begin{aligned} |l_{jj}| &= |v_{jj}| \le \frac{1}{2} \text{ and} \\ |l_{ji}| &= |t_{ji} + k_{ji}| \le \frac{1}{2} \text{ for } i < j \end{aligned}$$

Using that $C$ is a <u>convex</u>, <u>symmetric</u> star body we have the triangle inequality. Hence,

$$\begin{aligned} f_C(b_j) &\le f_C(l_{jj}a_j) + \cdots + f_C(l_{j1}a_1) \\ &= |l_{jj}|f_C(a_1) + \cdots + |l_{j1}|f_C(a_1) \\ &\le \frac{1}{2}(f_C(a_j) + \cdots + f_C(a_1)). \end{aligned}$$

□

**Corollary 2.3.6.** *Let $C$ be a convex, symmetric star body, and let $\Lambda$ be a lattice in $\mathbb{R}^n$ with successive minima $\lambda_1, \ldots, \lambda_n$ with respect to $C$. Then there exists a basis $b_1, \ldots, b_n$ of $\Lambda$ with*

$$f_C(b_j) \le \max\left\{\lambda_j, \frac{1}{2}(\lambda_1 + \cdots + \lambda_j)\right\}$$

*Proof.* Immediate form Lemmas 2.3.3 and 2.3.5. □

## 2.4 Minkowski's Second Theorem

Minkowski's Second Theorem is a refinement of his First Theorem and a central result in Geometry of Numbers. Let's start by rephrasing Minkowski's First Theorem.

First note if $C \subset \mathbb{R}^n$ is convex, symmetric and of positive volume then there exist $\varepsilon > 0$, $x \in C$ such that

$$B_\varepsilon(x) \subset C.$$

But then there exists $\varepsilon' > 0$ such that $B_{\varepsilon'}(0) \subset C$. So the origin lies in the interior of $C$. So we can consider the successive minima $\lambda_1, \ldots, \lambda_n$ of $\Lambda$ with respect to $C$, where $\Lambda$ is a lattice in $\mathbb{R}^n$.

Note that by definition of $\lambda_1$:

$$\forall \varepsilon > 0 : \ (\lambda_1 - \varepsilon)C \text{ contains } \underline{\text{no}} \text{ non-zero lattice point.}$$

Minkowski's First Theorem yields:

$$\lambda_1^n \cdot \mathrm{Vol}(C) = \mathrm{Vol}(\lambda_1 C) \le 2^n \cdot \det \Lambda \tag{57}$$

On the other hand (57) and $\mathrm{Vol}\, C > 2^n \det \Lambda$ implies $\lambda_1 < 1$ and hence $C$ contains a non-zero lattice point. The following theorem is much more precise than (57)!

**Theorem 2.4.1** (Minkowski's Second Theorem)**.** *Let $C$ be a convex, symmetric star body in $\mathbb{R}^n$, and let $\Lambda$ be a lattice in $\mathbb{R}^n$ with successive minima $\lambda_1, \dots, \lambda_n$ with respect to $C$. Then*

$$\frac{2^n}{n!} \det \Lambda \leq \lambda_1 \dots \lambda_n \cdot \operatorname{Vol} C \leq 2^n \det \Lambda$$

**Remark.**

- *Both bounds are sharp. For the upper bound take $\Lambda = \mathbb{Z}^n$ and $C = [-1,1]^n$; so $\lambda_1 = \dots = \lambda_n = 1 = \det \Lambda$, and $\operatorname{Vol} C = 2^n$. For the lower bound take $\Lambda = \mathbb{Z}^n$ and $C$ defined by $|x_1| + \dots + |x_n| \leq 1$. Then $\operatorname{Vol} C = \frac{2^n}{n!}$ and $\lambda_1 = \dots = \lambda_n = 1 = \det \Lambda$.*
- *The upper bound is much harder to prove. We will prove Theorem 2.4.1 only for the ball $C = B_1(0)$.*

*Proof.* (Special case $C = B_1(0)$)
Put

$$\delta(C) = \sup_M \frac{\lambda_1^n(M, C)}{\det M}$$

where the supremum runs over all lattices $M$ in $\mathbb{R}^n$. By Minkowski's First Theorem 2.2.2 and (57), we have

$$\delta(C) \leq \frac{2^n}{\operatorname{Vol} C}.$$

We will show that if $C = B_1(0)$ then

$$\det(\Lambda) \leq \lambda_1 \dots \lambda_n \leq \delta(C) \det(\Lambda). \tag{58}$$

In particular, as $\operatorname{Vol}(C) \geq \frac{2^n}{n!}$, we get

$$\frac{2^n}{n!} \lambda_1 \dots \lambda_n \leq \operatorname{Vol}(C) \lambda_1 \dots \lambda_n \leq 2^n \det(\Lambda).$$

For the lower bound in (58) take linearly independent $a_1, \dots, a_n \in \Lambda$ with $|a_i| = \lambda_i$ using Lemma 2.3.3. For the sublattice $\Lambda' = (a_1 \dots a_n)\mathbb{Z}^n \subset \Lambda$ we have

$$\det(\Lambda') = I \det(\Lambda),$$

where $I$ is the index of $\Lambda'$ in $\Lambda$. By Hadamard's inequality we have

$$|a_1| \cdot \dots \cdot |a_n| \geq \det(\Lambda') \geq \det(\Lambda).$$

Thus

$$\det(\Lambda) \leq \lambda_1 \dots \lambda_n.$$

Now we prove the upper bound in (58). Let $b_1, \dots, b_n$ be a basis of $\Lambda$ as in Corollary 2.3.4. As in the prove of Lemma 2.3.2 we can find mutually orthogonal vectors $c_1, \dots, c_n$ such that

$$b_j = t_{j1} c_1 + \dots + t_{jj} c_j \ (t_{ji} \in \mathbb{R})$$

By scaling we can assume $|c_j|^2 = 1$ $(1 \leq j \leq n)$. Now

$$\sum_{j=1}^n u_j b_j = \sum_{i=1}^n \left( \sum_{j \geq i} u_j t_{ji} \right) c_i$$

thus

$$\left|\sum_{j=1}^{n} u_j b_j\right|^2 = \sum_{i=1}^{n} \left(\sum_{j \geq i} u_j t_{ji}\right)^2 \tag{59}$$

Next we show that

$$\sum_{i=1}^{n} \lambda_i^{-2} \left(\sum_{j \geq i} u_j t_{ji}\right)^2 \geq 1 \tag{60}$$

where $u = (u_1, \ldots, u_n) \in \mathbb{Z}^n \smallsetminus 0$.
Let $u \in \mathbb{Z}^n \smallsetminus 0$ with

$$u_J \neq 0 \text{ and } u_j = 0 \text{ for } j > J. \tag{61}$$

Then $u_1 b_1 + \cdots + u_n b_n,\ b_1, \ldots, b_{J-1}$ are linearly independent and by Corollary 2.3.4 we have

$$\left|\sum_{j=1}^{n} u_j b_j\right|^2 \geq \lambda_J^2 \tag{62}$$

Moreover, (61) implies that summands with $j > J$ in (59) and (60) are zero. Thus, the left hand side in (60) is equal to

$$\sum_{i \leq J} \lambda_i^{-2} \left(\sum_{i \geq j} u_j t_{ji}\right)^2 \geq \sum_{i \leq J} \lambda_J^{-2} \left(\sum_{i \geq j} u_j t_{ji}\right)^2 = \lambda_J^{-2} \left|\sum_{j=1}^{n} u_j b_j\right|^2 \underbrace{\geq}_{\text{by (62) and (59)}} 1 \tag{63}$$

So if $\Lambda'$ is the lattice with basis

$$b'_j = t_{j1} \lambda_1^{-1} c_1 + \cdots + t_{jj} \lambda_j^{-1} c_j \ (1 \leq j \leq n) \tag{64}$$

Then

$$\left|\sum_{j=1}^{n} u_j b'_j\right| \geq 1$$

for every point $\sum_{j=1}^{n} u_j b'_j \in \Lambda' \smallsetminus 0$. Hence,

$$\lambda_1(\Lambda', C) \geq 1 \tag{65}$$

But

$$\det(\Lambda') = \lambda_1^{-1} \cdots \lambda_n^{-1} \det(\Lambda) \tag{66}$$

$\lambda_i = \lambda_i(\Lambda, C)$
Moreover, by definition

$$\frac{\lambda_1^n(\Lambda', C)}{\det \Lambda'} \leq \sup_M \frac{\lambda_1^n(M, C)}{\det M} = \delta(C) \tag{67}$$

Combining (65), (66) and (67) we conclude

$$\lambda_1 \cdots \lambda_n = \det(\Lambda) \frac{1}{\det(\Lambda')} \leq \det(\Lambda) \frac{\lambda_1^n(\Lambda', C)}{\det(\Lambda')} \leq \det(\Lambda) \cdot \delta(C) \tag{68}$$

□

## 2.5 Counting lattice points

How many integer pairs $(x, y)$ solve the Diophantine inequality

$$x^2 + y^2 \leq T?$$

What about

$$3x^2 + 5y^2 + 7z^2 \leq T,$$

or more generally

$$F(\underline{x}) < T,$$

where $F$ is a positive definite quadratic form in $n$ variables? Even more generally, let $S \subset \mathbb{R}^n$ and $\Lambda$ a lattice in $\mathbb{R}^n$, we would like to get a (non-trivial) estimate for $(\Lambda \cap S)$.

Suppose $S$ is measurable and "nicely shaped", and let

$$F_v = [0,1)b_1 + \cdots + [0,1)b_n + v \ (v \in \Lambda)$$

be a fundamental cell (with respect to basis $b_1, \ldots, b_n$).

The idea is as follows: $|\Lambda \cap S| \approx$ number of $F_v$'s that lie in $S \approx \frac{\mathrm{Vol}(S)}{\mathrm{Vol}(F_v)} = \frac{\mathrm{Vol}(S)}{\det(\Lambda)}$

"Nice set"

"Bad set"

To characterize nice sets we use the following definition.

**Definition.** *Let $n \geq 2$, $M$ be in $\mathbb{N}$ and $L \geq 0$ real. We say the set $Z$ lies in* $\underline{\mathrm{Lip}(n, M, L)}$ *if*

- $Z \subset \mathbb{R}^n$
- *there exist $M$ maps $\phi_i : [0,1]^{n-1} \to \mathbb{R}^n$*

*satisfying a Lipschitz condition with constant $L$, i.e.,*

$$|\phi_i(x) - \phi_i(y)| \leq L|x - y| \ \forall x, y \in [0,1]^{n-1}$$

*and such that the union of their images covers $Z$, i.e.,*

$$Z \subset \cup_{i=1}^{M} \phi_i([0,1]^{n-1}).$$

**Example.** *The sphere $S' \subset \mathbb{R}^2$ lies in* $\mathrm{Lip}(2, 1, 2\pi)$.

$$\phi(x) = (\cos(2\pi x), \sin(2\pi x)) \ \textit{for } 0 \leq x \leq 1$$

We can now state the main result of Section 5. Recall that the boundary $\partial S$ of $S \subset \mathbb{R}^n$ is defined by the topological closure $\bar{S}$ minus the interior $\mathrm{Int}(S)$ of $S$.

$$\partial S = \bar{S} \setminus \mathrm{Int}(S)$$

We follow an approach of Masser and Vasler.

**Theorem 2.5.1.** *Let $S \subset \mathbb{R}^n$ be bounded and suppose that $\partial S \in \mathrm{Lip}(n, M, L)$. Let $\Lambda$ be a lattice in $\mathbb{R}^n$ and $\lambda_1$ its first successive minimum with respect to the unit ball. Then, $S$ is measurable and*

$$\left| |\Lambda \cap S| - \frac{\mathrm{Vol}(S)}{\det \Lambda} \right| \leq c \cdot M \cdot \left( \frac{L}{\lambda_1} + 1 \right)^{n-1}$$

*where $c$ is a constant depending only on $n$.*

For the proof we need the following lemma.

**Lemma 2.5.2.** *Let $S \subset \mathbb{R}^n$ be bounded and measurable and let $\Lambda$ be a lattice in $\mathbb{R}^n$. Let $b_1, \ldots, b_n$ be a basis of $\Lambda$, $F_v = [0,1)b_1 + \cdots + [0,1)b_n + v$ the corresponding fundamental cells and write*

$$\mathcal{T} = |\{v \in \Lambda : F_v \cap \partial S \neq \varnothing\}|,$$

*the number of cells that intersect the boundary $\partial S$ of $S$. Then*

$$\left| |\Lambda \cap S| - \frac{\operatorname{Vol} S}{\det \Lambda} \right| \leq \mathcal{T}$$

*Proof.* The cells $F_v$ $(v \in \Lambda)$ define a partition of $\mathbb{R}^n$. Every $F_v$ contains exactly one lattice point, namely $v$. Let $m = |\{v \in \Lambda : F_v \subset S\}|$. Then $m \leq |S \cap \Lambda|$.
Now suppose $v \in S$. Then either $F_v \subset S$ or $F_v$ contains a point from $S$ and from its complement $S^C$. The connecting line segment of these points lies in $F_v$ as $F_v$ is convex and it must contain a point of the boundary $\partial S$. Hence,

$$|\Lambda \cap S| \leq m + \mathcal{T}.$$

Now $\operatorname{Vol} F_v = \det \Lambda$, and the union of all cells $F_v$ that lie in $S$ has volume $m \cdot \det \Lambda$. The volume of the union of cells $F_v$ that have non-empty intersection with $S$ is at most $(m + \mathcal{T}) \det \Lambda$. So we have proved the inequalities

- $m \leq |\Lambda \cap S| \leq (m + \mathcal{T})$
- $m \cdot \det \Lambda \leq \operatorname{Vol} S \leq (m + \mathcal{T}) \det \Lambda$

Thus

$$\left| |\Lambda \cap S| - \frac{\operatorname{Vol} S}{\det \Lambda} \right| \leq \mathcal{T}$$

□

We can now prove Theorem 2.5.1.

*Proof of Theorem 2.5.1.* We use $c_1, c_2, c_3, c_4, c_5, c_6, c_7$ to denote constants that depend only on $n$.
First let's assume $\Lambda = \mathbb{Z}^n$, so $\lambda_1 = 1$. We take the standard basis $b_i = c_i$ $(1 \leq i \leq n)$ and apply Lemma 2.5.2; so $\mathcal{T} = \mathcal{T}(e_1, \ldots, e_n)$. We split $[0,1]^{n-1}$ into $L_1^{n-1}$ subcubes of side length $\frac{1}{L_1}$ where $L_1 = [L] + 1$. The images of these subcubes under the parametrizing maps $\phi$ have diameter at most $L \cdot \frac{c_1}{L_1} \leq c_1$. Thus, no more than $c_2 = (c_1 + 2)^n$ cells $F_v$ can meet such a single image (= the image of a single subcube). Hence,

$$\mathcal{T} \leq M \cdot c_2 \cdot L_1^{n-1} \leq M \cdot c_2 \cdot (L+1)^{n-1} \tag{69}$$

As $\lambda_1(\mathbb{Z}^n) = 1$ this proves the claim for $\Lambda = \mathbb{Z}^n$ thanks to Lemma 2.5.1.

Now let $\Lambda$ be an arbitrary lattice. By Corollary 2.3.6 there exists a basis $b_1, \ldots, b_n$ of $\Lambda$ such that

$$|b_i| \leq c_3 \lambda_i \ (1 \leq i \leq n)$$

where $\lambda_1, \ldots, \lambda_n$ are the successive minima with respect to the unit ball. Let

$$A^{-1} = (b_1 \ldots b_n) \in \mathrm{GL}_n(\mathbb{R}),$$

so that

$$A(\Lambda) = \mathbb{Z}^n.$$

Now

$$|S \cap \Lambda| = |A(S \cap \Lambda)| = |A(S) \cap A(\Lambda)| = |A(S) \cap \mathbb{Z}^n|.$$

So we can apply the case $\Lambda = \mathbb{Z}^n$ to the set $A(S)$. The boundary $\partial A(S)$ can be parametrized by the $M$ maps $\psi(x) = A(\phi(x))$ which satisfy

$$|\psi(x) - \psi(y)| \le \|A\| \cdot L|x - y|,$$

where $\|A\|$ denotes the Euclidean operator norm of $A$.
By Cramer's rule the entries of row $i$ of $A$ are of the form $\frac{\mu}{\det \Lambda}$, where $\mu$ is a minor of the matrix with columns $b_1, \dots, b_n$ omitting $b_i$. Using Hadamard's inequality we conclude

$$|\mu| \le \frac{|b_1| \cdots |b_n|}{|b_i|} \le c_4 \cdot \frac{\lambda_1 \cdots \lambda_n}{\lambda_i}.$$

By Minkowski's Second Theorem we have

$$\lambda_1, \dots, \lambda_n \le c_5 \cdot \det \Lambda.$$

Hence, each entry of $A$ has absolute value at most $\frac{c_6}{\lambda_1}$. It follows

$$\|A\| \le \frac{c_7}{\lambda_1}.$$

Replacing $L$ in (69) by $\frac{c_7 L}{\lambda_1}$ proves the theorem. □

**Remark.** *We have not shown that $S$ is measurable. One could do that by showing that $\partial S$ has measure zero, and noting that every closed set is measurable. Why is* $\mathrm{Vol}(\partial S) = 0$*? Take $\Lambda = k^{-1}\mathbb{Z}^n$ where $k \in \mathbb{N}$, and $\mathcal{T}$ be associated to the basis $k^{-1}e_1, \dots, k^{-1}e_n$. The proof of Theorem 2.5.1 yields*

$$\mathcal{T} \le c_n M \left( \frac{L}{k^{-1}} + 1 \right)^{n-1} \le c_n M (L+1)^{n-1} k^{n-1}.$$

*Since*

$$\mathrm{Vol}(\partial S) \le \mathcal{T} \,\mathrm{Vol}\, F_v \le c_n M (L+1)^{n-1} k^{n-1} k^{-n} \to 0 \text{ as } k \to \infty.$$

*Hence* $\mathrm{Vol}(\partial S) = 0$.

In some applications a more precise error term is needed that involves also the higher successive minima. With a bit more effort the following result could be proved.

**Theorem 2.5.3.** *Same hypothesis as in Theorem 2.5.1 and $\lambda_1, \dots, \lambda_n$ successive minima with respect to the unit ball. Then*

$$\left| |\Lambda \cap S| - \frac{\mathrm{Vol}\, S}{\det \Lambda} \right| \le cM \underbrace{\max_{0 \le i < n} \frac{L}{\lambda_1 \cdot \lambda_i}}_{:=1 \text{ for } i=0}$$

*and* $c = n^{3n^2}$.

# 3 Algebraic Number Theory

**Remark** (References).

- *Daniel Marcus "Number fields" Springer*
- *Course Notes "Algebraic Number Theory" Math 2803(?) by Matt Baker (available online on his webpage)*
- *Serge Lang: Algebraic Number Theory Addison & Wesley*

## 3.1 Introduction

Algebraic number theory is concerned with finite field extensions of $\mathbb{Q}$ and their "ring of integers", e.g., $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$ and $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$. These extensions of $\mathbb{Q}$ and $\mathbb{Z}$ are often needed; even when studying questions that initially involve only integers.

Let's consider some examples. We start with the very simple Diophantine equation

$$x^2 - x = y^3$$

to be solved with $x, y \in \mathbb{Z}$. We can factor the left hand side, and note that the factors $x, x-1$ are coprime. The unique prime factorization in $\mathbb{Z}$ tells us that

$$x = \pm u'^3 = u^3 \text{ with } u = \pm u'$$
$$x - 1 = \pm v'^3 = v^3 \text{ with } v = \pm v'$$

So $u^3 - v^3 = 1$. So $(u, v) = (1, 0), (0, -1)$ and thus $(x, y) = (1, 0), (0, 0)$.

So here $\mathbb{Z}$ itself was sufficient. Next let's consider

$$x^2 + 2 = y^3$$

Now the polynomial $x^2+2$ does not factor over $\mathbb{Z}$, but it does over $\mathbb{Z}[\sqrt{-2}] = \{a + b\sqrt{-2} : a, b \in \mathbb{Z}\}$

$$x^2 + 2 = (x + \sqrt{-2})(x - \sqrt{-2})$$

If $x^2 + 2 = y^3$ then $x + \sqrt{-2}$ and $x - \sqrt{-2}$ are coprime in $\mathbb{Z}[\sqrt{-2}]$.
Why? Suppose $r \in \mathbb{Z}[\sqrt{-2}]$ and

$$r \mid x + \sqrt{-2} \text{ and}$$
$$r \mid x - \sqrt{-2}.$$

Thus $r \mid 2\sqrt{-2}$.
Let $\bar{r}$ be the complex conjugate of $r$. Then

$$\bar{r} \mid \overline{x + \sqrt{-2}} = x - \sqrt{-2}.$$

Thus

$$r\bar{r} \mid (x + \sqrt{-2})(x - \sqrt{-2}) = x^2 + 2 \text{ and}$$
$$r\bar{r} \mid (2\sqrt{-2})(-2\sqrt{-2}) = 8.$$

As $r\bar{r} \in \mathbb{Z}$ we conclude that $sr\bar{r} = 8$ with $s \in \mathbb{Z}[\sqrt{-2}]$ implies $s \in \mathbb{Z}$. So either

$$r = \pm 1 \text{ or } 2 \mid r\bar{r}.$$

If

$$2 \mid r\bar{r} \mid x^2 + 2 = y^3$$

then

$$\begin{aligned} & 2 \mid y^1 \\ \Longrightarrow\ & 8 \mid y^3 \\ \Longrightarrow\ & 8 \mid x^2 + 2 \end{aligned}$$

which is impossible since $x^2 \in \{\bar{0}, \bar{1}\}$ mod 4.
So we have $r = \pm 1$ and so

$$x + \sqrt{-2} \text{ and } x - \sqrt{-2}$$

are coprime in $\mathbb{Z}[\sqrt{-2}]$.
We conclude also that $\pm 1$ are the only units in $\mathbb{Z}[\sqrt{-2}]$. Suppose we have a unique prime factorization in $\mathbb{Z}[\sqrt{-2}]$. Then we could conclude as before that there exist $u, v \in \mathbb{Z}[\sqrt{-2}]$ such that

$$\begin{aligned} u^3 &= x + \sqrt{-2} \\ v^3 &= x - \sqrt{-2} \end{aligned}$$

With $u = a + b\sqrt{-2}$ $(a, b \in \mathbb{Z})$ we get

$$u^3 = (a^3 - 6ab^2) + (3a^2b - 2b^3)\sqrt{-2} = x + \sqrt{-2}$$

Hence,

$$\begin{aligned} a(a^2 - 6b^2) &= x \\ b(3a^2 - 2b^2) &= 1 \end{aligned}$$

So $b = \pm 1$. If $b = -1$ then $3a^2 - 2 = -1$ which is impossible. So $b = 1$ and $a^2 = 1$.
Hence $(x, y) \in \{(5, 3), (-5, 3)\}$.
As we shall see later $\mathbb{Z}[\sqrt{-2}]$ really has a unique prime factorization.

Now let's consider the Fermat equation

$$x^n + y^n = z^n \ (n \geq 3)$$

We could try to apply the same strategy to show that at least one of the coordinates equals 0. It suffices to consider prime exponents. Let's assume $p > 2$. We can also assume $\gcd(x, y, z) = 1$. Now take $\mathbb{Z}[\zeta]$ where $\zeta = e^{-\frac{2\pi\imath}{p}}$. Then

$$t^p - 1 = (t - 1)(t - \zeta) \ldots (t - \zeta^{p-1}).$$

Replacing $t$ by $-\frac{x}{y}$ we conclude

$$x^p + y^p = (x + y)(x + \zeta y) \cdots (x + \zeta^{p-1} y)$$

We split the solutions in two classes:

1. $(x,y,z)$ with $p \nmid xyz$
2. $(x,y,z)$ with $p$ divides exactly one of the coordinates.

We consider only solutions as in 1). For $p = 3$ we note that $x^3 + y^3 = z^3$ is impossible as each of these cubes is $\pm 1 \bmod 9$. So assume $p > 3$. Suppose that there exists a unique prime factorization in $\mathbb{Z}[\zeta]$. Then one can show that

$$x + \zeta y = \varepsilon \alpha^p$$

where $\varepsilon$ is a unit and $\alpha \in \mathbb{Z}[\zeta]$. Then one can show that if

$$\begin{gathered} x + \zeta y = \varepsilon \alpha^p \text{ and} \\ p \nmid xy \end{gathered}$$

then

$$x \equiv y \bmod p.$$

As

$$x^p + (-z)^p = (-y)^p$$

we also conclude $x \equiv -z \bmod p$. So

$$2x^p \equiv x^p + y^p = z^p \equiv (-x)^p \pmod{p}.$$

So

$$p \mid 3x^p.$$

As $p > 3$ and $p \nmid x$ we get a contradiction; so no solutions of class 1), provided $\mathbb{Z}[\zeta]$ has a unique prime factorization.

The latter holds for $p < 23$ but it "usually" fails. To solve Fermat completely Wiles and Wiles-Taylor used the theory of elliptic curves. On the other hand the Catalan equation

$$x^n - y^m = 1 \ (n, m > 1,\ x, y > 0)$$

was solved completely by Mihăilescu using algebraic number theory.

## 3.2 2. Basic notions

Let $R$ be a commutative ring with 1, and denote by $R^*$ the subset of its units. An element $x \in R$ is called *irreducible* if

- $x \neq 0$, $x \notin R^*$ and
- $x = a \cdot b$ with $a, b \in R \implies a \in R^*$ or $b \in R^*$

An element $\pi \in R$ is called *prime* if

- $\pi \notin R^*$, $\pi \neq 0$ and
- $\pi \mid x \cdot y$ with $x, y \in R \implies \pi \mid x$ or $\pi \mid y$

Two elements $x, y \in R$ are called *associate* if $\exists u \in R^*$ such that $y = ux$.

A ring $R$ is called a unique factorisation domain (UFD) if

1. $R$ is an integral domain

2. Every non-zero non-unit $x \in R$ can be written as a product $x = q_1 \cdots q_r$ with finitely many irreducible elements $q_1, \ldots, q_r \in R$.

3. This decomposition is unique up to unites and the order of the factors.

**Example.** • *$R = \mathbb{Z}$, $R^* = \{\pm 1\}$.*
*$\pi$ is prime if and only if $\pi$ is irreducible.*
*And $R$ is a UFD.*

- *$R = \mathbb{Z}[\sqrt{-5}]$.*
*If $x = a + b\sqrt{-5} \neq 0$ in $R$ then $x^{-1} = \frac{a - b\sqrt{-5}}{a^2 + 5b^2}$. So*

$$x^{-1} \in R \implies a^2 + 5b^2 \mid a.$$

*Thus $R^\times = \{\pm 1\}$.*
*Consider the norm map $N : R \to \mathbb{Z}$ defined by*

$$N(a + b\sqrt{-5}) = a^2 + 5b^2.$$

*Then $N(x \cdot y) = N(x) \cdot N(y)$ for all $x, y \in \mathbb{Z}[\sqrt{-5}]$ and $N(x) = 1 \iff x \in R^*$. Consider the decompositions*

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}) \tag{70}$$

*All factors are irreducible.*
*Why? If $xy = 2 \implies 4 = N(2) = N(x)N(y)$. Now $N(x) = 2$ is impossible.*
*Hence, $N(x) = 1$ or $N(y) = 1$. So either $x \in R^*$ or $y \in R^*$.*
*The same argument applies for the other factors.*
*Clearly none of these are associate so (70) are two essentially different decompositions in irreducable factors. So $R$ is not a UFD. None of the factors in (70) is prime.*
*Indeed, e.g., $2 \mid (1 + \sqrt{-5})(1 - \sqrt{-5})$. But $2 \nmid 1 + \sqrt{-5}$ and $2 \nmid 1 - \sqrt{-5}$ otherwise*

$$4 = N(2) \mid N(1 \pm \sqrt{-5}) = 6 \lightning$$

Recall that an ideal $I$ of $R$ is an additive subgroup of $R$ that satisfies

$$r \in R \text{ and } x \in I \implies r \cdot x \in I$$

An ideal $\mathfrak{p}$ of $R$ is called a *prime ideal* if $a, b \in R$ and $a \cdot b \in \mathfrak{p} \implies a \in \mathfrak{p}$ or $b \in \mathfrak{p}$. An ideal $I$ of $R$ is called *maximal* if the only ideals of $R$ containing $I$ are $R$ and $I$ itself. If $I, J$ are ideals of $R$ then we define

- $I + J = \{x + y : x \in I, y \in J\}$
- $I \cdot J = \{\sum_{i=1}^n x_i y_i : n \in \mathbb{N}, x_i \in I, y_i \in J (1 \leq i \leq n)\}$

These are both ideals of $R$.

An ideal $I$ of $R$ is called *principal* if there exists an $x \in R$ such that

$$I = \{r \cdot x : r \in R\} = \langle x \rangle.$$

A ring in which every ideal is principle is called a *principal ideal domain* (PID).

A ring is called *Euclidean* if there exists a map $\phi : R \to \mathbb{Z}$ such that

- $\phi(x) \geq 0$
- $\phi(0) = 0$
- $\forall x, y \in R$, $y \neq 0$ there exist $r, q \in R$ such that $x = q \cdot y + r$ and either $r = 0$ or $\phi(r) < \phi(y)$.

**Example.**

- *$R = \mathbb{Z}$, $\phi(x) = |x|$*
- *$R = K[t]$, where $K$ is a field.*
  *$\phi(x) = \deg_t(x)$ if $x \neq 0$ and $\phi(0) = 0$.*

**Theorem 3.2.1.** *Every Euclidean ring is a PID.*

*Proof.* Let $I$ be an ideal of $R$. If $I = (0)$ then we are done.
Suppose $I \neq \langle 0 \rangle$. Then let $y \in I$ be non-zero with $\phi(y)$ minimal. Then $I = \langle y \rangle$.
Why? Suppose $x \in I$. Then there exist $q, r \in R$ such that $x = q \cdot y + r$. As $I$ is an ideal we have $q \cdot y \in I$ and thus $r = x - qy \in I$. By the minimality of $y$ we have $r = 0$. This shows that $x \in \langle y \rangle$. □

So in particular, $\mathbb{Z}$ and $K[t]$ are PIDs.

**Corollary 3.2.2.** *The rings $\mathbb{Z}[\sqrt{-1}]$ and $\mathbb{Z}[\sqrt{-2}]$ are Euclidean and thus PIDs.*

*Proof.* Identify $\mathbb{C}$ with $\mathbb{R}^2$ then $\mathbb{Z}[\sqrt{-2}]$ can be seen as a lattice in $\mathbb{R}^2$ with fundamental cells $[0,1)\begin{pmatrix}1\\0\end{pmatrix} + [0,1)\begin{pmatrix}0\\\sqrt{2}\end{pmatrix} + v$ $(v \in \Lambda = \mathbb{Z}[\sqrt{-2}])$
We take $\phi(\cdot) = N(\cdot)$ the norm map, so $\phi(a + b\sqrt{-2}) = a^2 + 2b^2$. So $\phi(z) \in \mathbb{Z}$, add pic
$\phi(z) \geq 0$, $\phi(0) = 0$.
Now let $x, y \in \mathbb{Z}[\sqrt{-2}]$ and $y \neq 0$. Let $q \in \mathbb{Z}[\sqrt{-2}]$ be a closest lattice point to the complex number $\frac{x}{y}$. Then

$$\left|\frac{x}{y} - q\right| \leq \frac{\sqrt{3}}{2}.$$

Put $r = x - qy$. Hence,

$$\begin{aligned}
\phi(v) = \phi(x - qy) &= |x - qy|^2 \\
&= |y|^2 \left|\frac{x}{y} - q\right|^2 \\
&\leq \frac{3}{4}|x|^2 = \frac{3}{4}\phi(y) \\
&< \phi(y).
\end{aligned}$$

This shows that $\mathbb{Z}[\sqrt{-2}]$ is Euclidean. The same argument applies for $\mathbb{Z}[\sqrt{-1}]$. □

The argument fails already for $\mathbb{Z}[\sqrt{-3}]$ as $\frac{\sqrt{1+3}}{2} \nless 1$. And indeed $\mathbb{Z}[\sqrt{-3}]$ is not Euclidean.

## 3.3 Integrality

Let $A$ be a subring of $B$. We say $b \in B$ is *integral* over $A$ if $b$ is the root of a *monic* polynomial with coefficients in $A$. Clearly every $a \in A$ is integral over $A$. We say $B$ is integral over $A$ if every $b \in B$ is integral over $A$.
Note that $x = \frac{r}{s} \in \mathbb{Q}$ with $\gcd(r,s) = 1$ is integral over $\mathbb{Z}$ if and only if $x \in \mathbb{Z}$.
Why? Indeed,

$$\left(\frac{r}{s}\right)^n + a_1\left(\frac{r}{s}\right)^{n-1} + \cdots + a_{n-1}\left(\frac{r}{s}\right) + a_n = 0 \; (a_i \in \mathbb{Z})$$

then

$$r^n + sa_1r^{n-1} + \cdots + s^{n-1}a_{n-1}r + s^n a_n = 0$$

Hence, $s \mid r^n$ and thus $s = \pm 1$. So $x \in \mathbb{Z}$.

Let $A_B = \{b \in B : b \text{ is integral over } A\}$. We call this the *integral closure* of A in B.

We will show that $A_B$ is a ring. In particular, if $x, y \in B$ are integral over $A$ then so are $x \cdot y$ and $x + y$.

Recall that an $A$-module $M$ is a generalisation of the concept of a vector space over a field, where the field is replaced by a ring $A$. We say that $M$ is finitely generated as an $A$-module if there exist $m_1, \dots, m_r \in M$ such that every $m \in M$ can be written as

$$m = a_1m_1 + \cdots + a_rm_r$$

where $a_1, \dots, a_r \in A$. We say that $m_1, \dots, m_r$ generate $M$ as an $A$-module.

**Lemma 3.3.1.** *Let $A \subset B$ be rings and let $M$ be a $B$-module. Suppose that $M$ is finitely generated as a $B$-module and that $B$ is finitely generated as an $A$-module. Then $M$ is finitely generated as an $A$-module.*

*Proof.* Let $x_1, \dots, x_m$, and $y_1, \dots, y_n$ be generators for $M$ as a $B$-module and $B$ as an $A$-module respectively. Then $x_iy_j$ ($1 \le i \le m$, $1 \le j \le n$) are generators for $M$ as an $A$-module.
Why? Let $x \in M$ and write

$$x = \sum_{i=1}^{m} b_i x_i$$

with $b_i \in B$. Moreover, for each $i$ we can find $a_{ij} \in A$ such that

$$b_i = \sum_{j=1}^{n} a_{ij} y_j.$$

Thus

$$x = \sum_i \left(\sum_j a_{ij} y_j\right) \cdot x_i = \sum_{i,j} a_{ij} x_i y_j$$

□

Recall that all rings in this Chapter 3 are integral domains with 1 (unless specified otherwise). For rings $A \subset B$ and $x \in B$ we write $A[x]$ for the smallest ring contains $A$ and $x$.

**Theorem 3.3.2.** *Let $A \subset B$ be rings and $x \in B$. The following statements are equivalent:*

*i)* *$x$ is integral over $A$*

*ii)* *$A[x]$ is finitely generated as an $A$-module.*

*iii)* *$A[x]$ is contained in a subring of $B$ which is finitely generated as an $A$-module.*

*Proof.* *i)* $\Rightarrow$ *ii)* If $x$ is integral over $A$ then

$$x^n + a_1 x^{n-1} + \cdots + a_n = 0 \ (a_i \in A)$$

Thus $x^n = -(a_1 x^{n-1} + \ldots a_n)$ and so $A[x]$ is generated by $1, x, \ldots, x^{n-1}$ as an $A$-module.
*ii)* $\Rightarrow$ *iii)* Trivial
*iii)* $\Rightarrow$ *i)* Suppose $A[x] \subset C$ for a subring $C$ of $B$ that is finitely generated as an $A$-module. As $C$ is a ring and $x \in C$ we have

$$x \cdot C \subset C,$$

i.e., $y \in C \implies x \cdot y \in C$. Let $y_1, \ldots, y_n$ be generators for $C$ and express

$$x \cdot y_i = \sum_j a_{ij} y_j$$

with $a_{ij} \in A$ $(1 \le i \le n)$. We get a matrix equation

$$\begin{pmatrix} xy_1 \\ \vdots \\ xy_n \end{pmatrix} = T \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

with $T = [a_{ij}]$. As $1 \in A \subset C$ the vector $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \neq 0$. Now

$$(xI - T) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = 0.$$

Hence $\det(xI - T) = 0$.
Now

$$\det(xI - T) = x^n + Q(x)$$

where $Q(x) \in A[x]$ and $\deg Q \le n - 1$.

This proves that $x$ is integral over $A$. □

**Corollary 3.3.3.** *Let $A \subset B \subset C$ be rings and suppose $C$ is integral over $B$ and $B$ is integral over $A$. Then $C$ is integral over $A$.*

*Proof.* Let $x \in C$. We want to show that $x$ is integral over $A$. Now

$$x^n + b_1 x^{n-1} + \cdots + b_n = 0 \ (b_i \in B)$$

Let

$$\tilde{B}_i = A[b_1, \ldots, b_i] \ (0 \le i \le n).$$

Then $b_i$ is integral over $\tilde{B}_{i-1}$ and so by Theorem 3.3.2 $\tilde{B}_i$ is finitely generated over $\tilde{B}_{i-1}$. By Lemma 3.3.1 we conclude that $\tilde{B}_n$ is finitely generated over $A$. As $b_1, \ldots, b_n \in \tilde{B}_n$ $x$ is integral over $\tilde{B}_n$. Thus by Theorem 3.3.2 $\tilde{B}_n[x]$ is finitely generated over $\tilde{B}_n$. Again by Lemma 3.3.1 we get that $\tilde{B}_n[x]$ is finitely generated over $A$ and thus by Theorem 3.3.2 also integral over $A$. □

**Corollary 3.3.4.** *Let $A \subset B$ be rings. Then*

$$A_B = \{b \in B : b \text{ integral over } A\}$$

*is a ring.*

*Proof.* It suffices to show that

$$x, y \in A_B \implies x \cdot y, x + y \in A_B.$$

So let $x, y \in A_B$. By Theorem 3.3.2 $A[x]$ is finitely generated as an $A$-module. As $y$ is integral over $A$ it is also integral over $A[x]$ and thus

$$(A[x])[y] = A[x, y]$$

is finitely generated as an $A[x]$-module. By Lemma 3.3.1 $A[x, y]$ is finitely generated over $A$. Thus by Theorem 3.3.2 every element in $A[x, y]$ is integral over $A$; in particular $x \cdot y$ and $x + y$ □

**Remark.**

- *Note that $\overline{\mathbb{Q}} = \mathbb{Q}_{\mathbb{C}}$, so $\overline{\mathbb{Q}}$ is a ring by Corollary 3.3.4.*
  *But $\overline{\mathbb{Q}}$ is even a field; indeed if $x \in \bar{\mathbb{Q}}$, $x \neq 0$ then*

$$\begin{aligned} & x^n + a_1 x^{n-1} + \cdots + a_n = 0 \ (a_i \in \mathbb{Q}) \\ & \qquad \implies (x^{-1})^n + \frac{a_{n-1}}{a_n}(x^{-1})^{n-1} + \cdots + \frac{1}{a_n} = 0 \end{aligned}$$

- *The ring $\mathbb{Z}_{\mathbb{C}} = \{\alpha \in \mathbb{C} : \alpha \text{ integral over } \mathbb{Z}\}$ is called the ring of algebraic integers.*

- *Let $A$ be a field and $\alpha$ a root of a non-zero polynomial with coefficients in $A$, i.e., $\alpha$ is algebraic over $A$.*
  *We write $f_\alpha(x)$ for the monic minimal polynomial of $\alpha$ over $A$, i.e., the monic polynomial in $A[x]$ of minimal degree that vanishes at $\alpha$.*
  *If $h(x) \in A[x]$, $h \neq 0$ and $h(\alpha) = 0$ then $f_\alpha \mid h$ in $A[x]$ as follows from the Euclidean division algorithm.*

**Lemma 3.3.5.** *Let $\alpha \in \bar{\mathbb{Q}}$ be an algebraic number and $f_\alpha(x)$ the monic minimal polynomial over $\mathbb{Q}$. Then:*

$$\alpha \in \mathbb{Z}_{\mathbb{C}} \iff f_\alpha(x) \in \mathbb{Z}[x].$$

*Proof.* "$\Leftarrow$" trivial
"$\Rightarrow$" $\exists h \in \mathbb{Z}[x]$ monic with $h(\alpha) = 0$. Then $f_\alpha \mid h$ in $\mathbb{Q}[x]$. Hence all roots of $f_\alpha$ vanish at $h$. Hence, all roots of $f_\alpha$ are algebraic integers. But the coefficients of $f_\alpha$ are symmetric functions in the roots ($f_\alpha$ is monic!) thus the coefficients are also algebraic integers, and they are also in $\mathbb{Q}$. We already know that $\mathbb{Z}_{\mathbb{C}} \cap \mathbb{Q} = \mathbb{Z}$ thus $f_\alpha \in \mathbb{Z}[x]$. □

A *number field* $K$ is a subfield of $\bar{\mathbb{Q}}$ which as a $\mathbb{Q}$-vector space has finite dimension. The latter is called the degree of $K$ over $\mathbb{Q}$ and denoted by $[K:\mathbb{Q}]$. By the "primitive element theorem" there exist $\alpha \in K$ such that

$$K = \mathbb{Q}(\alpha) = \{\frac{P(\alpha)}{Q(\alpha)} : P, Q \in \mathbb{Q}[x], Q(\alpha) \neq 0\}.$$

In fact $\mathbb{Q}(\alpha) = \mathbb{Q}[\alpha]$ and $1, \alpha, \ldots, \alpha^{\deg(f_\alpha)-1}$ is a $\mathbb{Q}$-basis for $K$, thus

$$[K:\mathbb{Q}] = \deg(f_\alpha) \text{ (see exercise sheet 6).}$$

The integral closure $\mathbb{Z}_K$ of $\mathbb{Z}$ in $K$ is usually dnoted by $O_K$.

The following result is central in algebraic number theory.

**Theorem 3.3.6.** *If $K$ is a number field then $O_K$ has a unique prime factorization of ideals, i.e., if $\mathfrak{a} \neq (1), (0)$ is an ideal in $O_K$ then there exist prime ideals $\mathfrak{p}_1, \ldots, \mathfrak{p}_s$ such that*

$$\mathfrak{a} = \mathfrak{p}_1 \ldots \mathfrak{p}_s$$

*and this decomposition is, up to the order of the factors, unique.*

Additional references: J. Neukirch, "Algebraic number Theory", Springer

We will not prove Theorem 3.3.6.

**Example.** *Consider $K = \mathbb{Q}[\sqrt{-5}]$ then $O_K = \mathbb{Z}[\sqrt{-5}]$ (see sheet 6).*
*Consider the following ideals*

$$\mathfrak{p}_1 = (2, 1+\sqrt{-5}),$$
$$\mathfrak{p}_2 = (2, 1-\sqrt{-5})$$

*generated as $O_K$-modules by $2, 1 \pm \sqrt{-5}$. Then*

$$\begin{aligned}\mathfrak{p}_1\mathfrak{p}_2 &= (4, 2(1-\sqrt{-5}), 2(1+\sqrt{-5}), 6)\\ &= (2)\underbrace{(2, 1+\sqrt{-5}, 1-\sqrt{-5}, 3)}_{=(1)} = (2)\end{aligned}$$

*With*

$$\mathfrak{p}_3 = (3, 1+\sqrt{-5}),$$
$$\mathfrak{p}_4 = (3, 1-\sqrt{-5})$$

*we find*

$$p_3p_4 = (3).$$

*Then the non-unique factorization into irreducable elements*

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$$

*in $O_K$ becomes the unique prime factorization into ideals in $O_K$*

$$\begin{aligned}(6) = (2) \cdot (3) &= (\mathfrak{p}_1\mathfrak{p}_2)(\mathfrak{p}_3\mathfrak{p}_4)\\ &= (\mathfrak{p}_1\mathfrak{p}_3)(\mathfrak{p}_2\mathfrak{p}_4)\\ &= (1 + \sqrt{-5})(1 - \sqrt{-5}).\end{aligned}$$

**Corollary 3.3.7.** *If $O_K$ is a principal ideal domain (PID) then $O_K$ is a unique factorization domain (UFD).*

*Proof.* Exercise (on your own) □

**Example.** *With $K = \mathbb{Q}(\sqrt{-1})$ or $\mathbb{Q}(\sqrt{-2})$ then $O_K = \mathbb{Z}[\sqrt{-1}]$ or $\mathbb{Z}[\sqrt{-2}]$ respectively (see sheet 6). We know that the above rings are Euclidean and hence PIDs, thus UFD.*

## 3.4 The ideal class group

Throughout this subsection $K$ denotes a number field.

A *fractional ideal* $I$ is an additive subgroup of $K$ such that there exists $a \in O_K$, $a \neq 0$ with

$$aI = \{a \cdot r : r \in I\}$$

is an ideal in $O_K$.
Note that the product of two fractional ideals $I$, $J$

$$I \cdot J = \{x \cdot y : x \in I, y \in J\}$$

is again a fractional ideal. For an ideal $J \neq 0$ in $O_K$ we denote

$$J^{-1} = \{x \in K : x \cdot J \subset O_K\}.$$

As $a \cdot J^{-1} \subset O_K$ for any $a \in J$ we easily see that $J^{-1}$ is a fractional ideal of $O_K$.

**Lemma 3.4.1** ("to divide is to contain")**.** *Let $\mathfrak{a}$, $\mathfrak{b}$ be ideals in $O_K$. Then*

$$\mathfrak{a} \mid \mathfrak{b} \iff \mathfrak{b} \subset \mathfrak{a}.$$

*Proof.* If $\mathfrak{b} \subset \mathfrak{a}$ then $\mathfrak{c} := \mathfrak{b} \cdot \mathfrak{a}^{-1} \subset \mathfrak{a}\mathfrak{a}^{-1} = O_K$. Thus $\mathfrak{c}$ is an ideal in $O_K$ and $\mathfrak{b} = \mathfrak{c}\mathfrak{a}$. Conversely if $\mathfrak{b} = \mathfrak{a} \cdot \mathfrak{c}$ with $\mathfrak{c}$ in $O_K$ then $\mathfrak{b} = \mathfrak{a} \cdot \mathfrak{c} \subset \mathfrak{a}$. □

**Lemma 3.4.2.** *The set $I_K$ of non-zero fractional ideal of $O_K$ forms a group under multiplication.*

*Proof.* It suffices to check that we have inverses.
Let $J \in I_K$. Then there exists an $a \in O_K$, $a \neq 0$ such that

$$I := aJ \subset O_K.$$

Then also

$$a \cdot I^{-1} \in I_K.$$

Moreover,

$$J \cdot aI^{-1} = I \cdot I^{-1} = O_K.$$

□

A fractional ideal $I$ is called *principal* if there exists an $x \in K$ such that

$$I = (x) = \{x \cdot r : r \in O_K\}.$$

Write $P_K$ for the subset of $I_K$ of non-zero principal fractional ideals. $P_K$ is a subgroup of $I_K$.

The *ideal class group* $CL_K$ is defined as the quotient group

$$CL_K = I_K / P_K.$$

We have the following exact sequence

$$1 \to O_K^* \to K^* \to I_K \to CL_K \to 1$$

(all maps are homomorphisms).
The map $I_K \to CL_K$ is clearly surjective. The expansion when passing from numbers (in $K^*$) to ideals (in $I_K$) is measured by the class group ($CL_K$) and $O_K^*$ measures the contraction in the same process.

**Theorem 3.4.3.** *$CL_K$ is finite.*

Let $\hom_{\mathbb{Q}}(K)$ be the set of $\mathbb{Q}$-homomorphisms from $K$ into $\mathbb{C}$. If $K = \mathbb{Q}[\alpha]$ and $\sigma \in \hom_{\mathbb{Q}}(K)$ then

$$0 = \sigma(f_\alpha(\alpha)) = f_\alpha(\sigma(\alpha))$$

so $\sigma(\alpha)$ is a root of $f_\alpha$. Denote these roots by $\alpha_1, \dots, \alpha_d$ so $d = [K : \mathbb{Q}]$.
Indeed each $\sigma(\alpha) = \alpha_i$ extends to a $\mathbb{Q}$-homomorphism of $K$. After relabelling let

$$\alpha_1, \dots, \alpha_r$$

be the real and

$$\alpha_{r+1}, \alpha_{r+1+s}, \dots, \alpha_{r+s}, \alpha_{r+2s}$$

be the $s$ pairs of complex conjugate roots of $f_\alpha$.
Then

$$\sigma_1, \dots, \sigma_r$$

are the real embeddings and

$$\sigma_{r+1}, \sigma_{r+1+s}, \dots, \sigma_{r+s}, \sigma_{r+2s}$$

are the $s$ pairs of complex conjugate embeddings.
We consider the *Minkowski-embedding*:

$$\begin{aligned}\sigma : K &\to \mathbb{R}^r \times \mathbb{C}^s \\ \alpha &\mapsto (\sigma_1(\alpha), \dots, \sigma_r(\alpha), \sigma_{r+1}(\alpha), \dots, \sigma_{r+s}(\alpha))\end{aligned}$$

Let $\mathfrak{a} \neq (0)$ be an ideal in $O_K$ and let $N(\mathfrak{a}) = [O_K : \mathfrak{a}]$ be the group index. We call $N(\mathfrak{a})$ the *norm* of $\mathfrak{a}$.

We make use of the following lemma which we won't prove.

**Lemma 3.4.4** (3.4.4)**.**

- *$N(\mathfrak{a})$ is finite for all $\mathfrak{a} \neq (0)$ ideals in $O_K$*
- *$N(\mathfrak{a} \cdot \mathfrak{b}) = N(\mathfrak{a}) \cdot N(\mathfrak{b})$ for $\mathfrak{a}, \mathfrak{b} \neq (0)$ ideals in $O_K$*
- *If $\alpha \in O_K$, $\alpha \neq 0$ $N((\alpha)) = \prod_{\sigma \in \hom_{\mathbb{Q}}(K)} |\sigma(\alpha)|$*
*Moreover, if $(0) \neq \mathfrak{a}$ is an ideal in $O_K$ then $\sigma\mathfrak{a}$ is a lattice in*

$$\mathbb{R}^r \times \mathbb{C}^s \simeq \mathbb{R}^{r+2s} = R^d$$

*($d = [K : \mathbb{Q}]$) with* $\frac{1}{2}$ or $-\frac{1}{2}$?

$$\det(\sigma\mathfrak{a}) = 2^{-s} N(\mathfrak{a}) \cdot |\Delta_K|^{-\frac{1}{2}}$$

*where $\Delta_K \in \mathbb{Z} \setminus 0$ is a certain invariant of $K$ called the discriminant of $K$.*
*Any $I \in I_K$ has the form $I = \mathfrak{a}\mathfrak{b}^{-1}$ with $\mathfrak{a}$ and $\mathfrak{b}$ ideals in $O_K$. By the multiplicity of the norm we can extend $N(\cdot)$ to $I_K$;*

$$N(I) = N(\mathfrak{a})/N(\mathfrak{b}).$$

**Lemma 3.4.5.** *Let $\mathfrak{a} \neq (0)$ be an ideal in $O_K$. Then there exists $0 \neq \alpha \in \mathfrak{a}$ such that*

$$N((\alpha)) \leq \left(\frac{2}{\pi}\right)^s \cdot \sqrt{|\Delta_K|} \cdot N(\mathfrak{a}).$$

*Proof.* Choose $c_i > 0$ $(1 \leq i \leq r+s)$ with

$$\prod_{i=1}^{r+s} c_i^{d_i} > \left(\frac{2}{\pi}\right)^s N(\mathfrak{a}) \sqrt{|\Delta_K|},$$

where $d_i = \begin{cases} 1 & : 1 \leq i \leq r \\ 2 & : r+1 \leq i \leq r+s \end{cases}$.

Consider

$$S = \{x \in \mathbb{R}^r \times \mathbb{C}^s : |x_i| < c_i (1 \leq i \leq r+s)\}.$$

Now $S$ is convex, symmetric in $\mathbb{R}^r \times \mathbb{C}^s \simeq \mathbb{R}^d$ with

$$\begin{aligned}\operatorname{Vol} S &= (2 \cdot c_1) \dots (2c_r)(\pi c_{r+1}^2) \dots (\pi c_{r+s}^2) \\ &> 2^d \cdot \det \sigma(\mathfrak{a}).\end{aligned}$$

By Minkowski's First Theorem there exists an $\alpha \in \mathfrak{a}$, $\alpha \neq 0$ such that $\sigma\alpha \in S$. Thus $|\sigma_i\alpha| < c_i$ $(1 \leq i \leq r+s)$, and hence

$$\begin{aligned} N((\alpha)) &= \prod_{i=1}^{r+s} |\sigma_i(\alpha)|^{d_i} \\ &< \prod_{i=1}^{r+s} c_i^{d_i}. \end{aligned}$$

As $\prod_{i=1}^{r+s} c_i^{d_i}$ can be chosen arbitrarily close to $\left(\frac{2}{\pi}\right)^s N(\mathfrak{a})\sqrt{|\Delta_K|}$ the claim follows. □

**Lemma 3.4.6.** *There are finitely many ideals in $O_K$ with bounded norm, i.e.,*

$$|\{\mathfrak{a} \subset O_K : a \neq 0, N(\mathfrak{a}) < M\}| < \infty \ \forall M > 0.$$

*Proof.* Let $\mathfrak{p}$ be a prime ideal in $O_K$. Then

$$\mathfrak{p} \cap \mathbb{Z} = p\mathbb{Z}$$

with a prime number $p \in \mathbb{Z}$.
By Lemma 3.4.1 $\mathfrak{p} \mid (p)$, hence

$$N(\mathfrak{p}) \mid N((p)) = p^d \ (d = [K : \mathbb{Q}]).$$

As there are only finitely many prime ideals $\mathfrak{p}$ that divide $(p)$ we conclude that there are only finitely many prime ideals of bounded norm. This implies that there are only finitely many ideals in $O_K$ of bounded norm. □

*Proof of Theorem 3.4.3.* Let $c \in CL_K$ and let $I \subset O_K$ be an ideal in $c^{-1}$. We write $[I] = c^{-1}$. By Lemma 3.4.5 we can choose $\alpha \in I$, $\alpha \neq 0$ such that

$$N((\alpha)) \leq \left(\frac{2}{\pi}\right)^s \cdot |\Delta_K|^{\frac{1}{2}} N(I).$$

By Lemma 3.4.1 we have

$$(\alpha) \subset I \implies I \mid (\alpha)$$

so

$$(\alpha) = I \cdot J$$

with $J \subset O_K$. Now $(\alpha) \in P_K$. So

$$[J] = [I]^{-1} = c.$$

Now

$$N(J) = \frac{N((\alpha))}{N(I)} \leq \left(\frac{2}{\pi}\right)^s |\Delta_K|^{\frac{1}{2}}.$$

Hence, any ideal class $c$ has an integral representative $J$ of bounded norm. But by Lemma 3.4.6 there are only finitely many of these. □

## 3.5 Dirichlet's Unit Theorem

Using geometry of numbers for a "multiplicative version" of the Minkowski-embedding one can prove the following fundamental result.

**Theorem 3.5.1** (Dirichlet's Unit Theorem)**.** *Let $K$ be a number field with $r$ real and $s$ pairs of complex conjugate embeddings. The group $O_K^*$ is the direct product of a finite cyclic group and of an abelian free group of rank $r+s-1$. So there exist*

$$\varepsilon_1, \dots, \varepsilon_{r+s-1} \text{ in } O_K^*$$

*such that*

$$\forall \varepsilon \in O_K^* \text{ exists a unique root of unity } \xi \text{ and a vector } (a_1, \dots, a_{r+s-1}) \in \mathbb{Z}^{r+s-1}$$

*such that*

$$\varepsilon = \xi \varepsilon_1^{a_1} \cdots \varepsilon_{r+s-1}^{a_{r+s-1}}.$$