THE NUMBER OF LATTICE POINTS IN IRRATIONAL POLYTOPES

BY BENCE BORDA

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Mathematics

Written under the direction of

József Beck

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2016

ABSTRACT OF THE DISSERTATION

The number of lattice points in irrational polytopes

by Bence Borda

Dissertation Director: József Beck

The discrepancy $\left|tP \cap \mathbb{Z}^d\right| - \lambda(P)t^d$ is studied as a function of the real variable $t > 1$, where P is a polytope in $\mathbb{R}^d$ with at least one vertex not in $\mathbb{Z}^d$. For a special class of cross polytopes and orthogonal simplices defined in terms of algebraic numbers the discrepancy is proved to be the sum of an explicit polynomial of t and a randomly fluctuating term of smaller order of magnitude. This phenomenon has not yet been described in the literature for any convex body. A general discrepancy bound is proved for polytopes the coordinates of the vertices of which all belong to a given real quadratic field. As a corollary a new property of the regular dodecahedron is obtained. Finally, answering a question of Beck, a formula is given for the variance of a random fluctuation arising in a lattice point counting problem on the plane. The main methods used are Fourier analysis and the theory of Diophantine approximation.

Acknowledgements

I would like to thank my advisor, József Beck, for the invaluable help and encouragement to write this dissertation. I would also like to express my gratitude to my professors at Eötvös Loránd University, from whom I learned the art of mathematics.

Table of Contents

Chapter 1

Introduction

1.1 Context

Counting the number of elements of a finite set has been a clearly motivated and much studied problem since the beginnings of mathematics. In the present dissertation the sets containing the elements to be counted are geometric objects. Being among the simplest geometric shapes, convex polytopes have played an important role in the history of mathematics. As a tribute to this rich history, a new property of a Platonic solid, the regular dodecahedron is proved.

The objects to be counted, however, could rather be characterized as discrete, or more specifically, number theoretic. Points with integral coordinates, called lattice points, are important in a wide range of fields from algebra, number theory and combinatorics, all the way to operations research. The main method used comes from yet another field. To count discrete objects in geometric sets, analytic methods are used, thereby completing the whole spectrum of mathematics.

There are several classical lattice point counting problems, many of which fit into the following framework. Given a compact convex set $B \subset \mathbb{R}^d$, and a magnifying factor $t > 1$, we wish to estimate the number of lattice points in the set

$$tB = \left\{ tx \in \mathbb{R}^d : x \in B \right\},$$

which we will denote by $\left|tB \cap \mathbb{Z}^d\right|$. It is not difficult to come up with the intuition that the number of lattice points in tB is close to the Lebesgue measure $\lambda(tB) = \lambda(B)t^d$. The difference

$$\left|tB \cap \mathbb{Z}^d\right| - \lambda(B)t^d$$

is called the discrepancy, or lattice rest of tB. In many applications one wishes to find an upper bound to the discrepancy as a function of the real variable t by finding an exponent $\alpha > 0$ such that

$$\left|tB \cap \mathbb{Z}^d\right| - \lambda(B)t^d = O\left(t^\alpha\right). \tag{1.1}$$

A considerable amount of attention has been given to the case when B is a compact convex set containing the origin in its interior, such that the boundary of B is a smooth $d-1$ dimensional submanifold of $\mathbb{R}^d$ with a nonzero and finite Gaussian curvature. It is easy to see that in this case (1.1) is true with $\alpha = d-1$. In [17] Müller conjectures that under these conditions (1.1) in fact holds with $\alpha = d-2+\varepsilon$ for any $\varepsilon > 0$ in dimensions $d = 3$ and $d = 4$, and with $\alpha = d-2$ in dimensions $d \geq 5$. The quest for improving the best exponent α for which (1.1) indeed holds for such general convex bodies is still ongoing.

It should be mentioned, that there are special classes of convex bodies for which the lattice point problem is completely solved. In the 1928 paper [15] Jarník considers the ellipsoid

$$B = \left\{(x_1, \ldots, x_d) \in \mathbb{R}^d : \frac{x_1^2}{a_1} + \cdots + \frac{x_d^2}{a_d} \leq 1\right\},$$

where $a_1, \ldots, a_d > 0$. It is established that if the coefficients $a_1, \ldots, a_d > 0$ are all rational, and $d \geq 5$, then the discrepancy $\left|tB \cap \mathbb{Z}^d\right| - \lambda(B)t^d$ is both $O\left(t^{d-2}\right)$ and $\Omega\left(t^{d-2}\right)$, as $t \to \infty$. Note that this result applies to all spheres of dimension $d \geq 5$ centered at the origin. The discrepancy can be much smaller for irrational coefficients, however. In the same paper [15] Jarník proves that in dimension $d \geq 4$ the discrepancy of tB is $O\left(t^{\frac{d}{2}+\varepsilon}\right)$ for every $\varepsilon > 0$, for almost every $a_1, \ldots, a_d > 0$ in the sense of the Lebesgue measure. To mention a more recent development, in the 1997 paper [3] Bentkus and Götze prove that in dimension $d \geq 9$ the discrepancy of tB, where B is an ellipsoid of arbitrary center and orientation, is $O\left(t^{d-2}\right)$, as $t \to \infty$. Thus the general ellipsoid problem is basically solved in dimensions $d \geq 9$.

In the 1960s Ehrhart studied the general lattice point counting problem in the case when B is the convex hull of finitely many lattice points in $\mathbb{R}^d$, in other words when

B is a lattice polytope. In a series of papers [8], [9] and [10] Ehrhart proves that for every lattice polytope B there exists a polynomial p with rational coefficients such that for any integer $t \geq 1$ we have $\left|tB \cap \mathbb{Z}^d\right| = p(t)$. This polynomial is called the Ehrhart polynomial of the lattice polytope B. Not surprisingly the highest degree term of $p(t)$ is always $\lambda(B)t^d$. The term of degree $d-1$ is known to be one half of the normalized surface area of tB. Here the normalized surface area of a hyperface of tB is defined as the surface area of the hyperface divided by the covolume of the $d-1$ dimensional sublattice of $\mathbb{Z}^d$ on the rational hyperplane containing the given hyperface.

Note that among compact convex sets the class of sets with a smooth boundary, and the class of lattice polytopes are in a sense two extreme points of a spectrum. The general lattice point counting problem in these two classes turned out to be disparate. For a compact convex set B with a smooth boundary the discrepancy $\left|tB \cap \mathbb{Z}^d\right| - \lambda(B)t^d$ is considered a purely random fluctuation, from which we cannot extract any non-oscillating main term. For a lattice polytope B, on the other hand, the discrepancy $\left|tB \cap \mathbb{Z}^d\right| - \lambda(B)t^d$ for integral values of t is simply a polynomial of t, with no random fluctuation present.

In this dissertation the general lattice point counting problem is studied in the case when the compact convex set B is a polytope in $\mathbb{R}^d$ the vertices of which are not all lattice points. There are surprisingly few results in this area, all of which apply only in dimension $d = 2$. The first problem of this type dates back almost a hundred years. In [11] and [12] Hardy and Littlewood considered the closed right triangle B in the plane with vertices $(0,0), (a_1,0), (0,a_2)$, where $a_1, a_2 > 0$, such that the slope $-\frac{a_2}{a_1}$ of the hypotenuse is irrational. For a magnifying factor $t > 1$, it was determined that the number of lattice points in tB has a main term

$$q(t) = \frac{a_1 a_2}{2} t^2 + \frac{a_1 + a_2}{2} t.$$

Notice that the highest degree term of the polynomial $q(t)$ is the area of tB, while the linear term is one half of the total length of the legs of the triangle. Thus $q(t)$ is the analogue of an Ehrhart polynomial, even though B is not a lattice polygon. The difference $\left|tB \cap \mathbb{Z}^d\right| - q(t)$ is considered a purely random fluctuation. It is not

difficult to see that the order of magnitude of this fluctuation is closely related to the classical Diophantine approximation problem of approximating the irrational number $\frac{a_2}{a_1}$ by rational numbers of small denominators. Hardy and Littlewood showed that if $\frac{a_2}{a_1}$ is a quadratic irrational, that is, an irrational number the minimal polynomial over $\mathbb{Q}$ of which is of degree 2, then the fluctuation $\left|tB \cap \mathbb{Z}^d\right| - q(t)$ is both $O(\log t)$ and $\Omega(\log t)$. As a groundbreaking result, in the same papers they showed that if $\frac{a_2}{a_1}$ is algebraic, then the fluctuation $\left|tB \cap \mathbb{Z}^d\right| - q(t)$ is $O(t^\alpha)$ for some $\alpha < 1$ depending on $\frac{a_2}{a_1}$. This is one of the first results related to the approximation of general algebraic numbers by rationals.

1.2 The main results

1.2.1 The polyhedral sphere problem

Consider the polytope

$$P = \left\{(x_1, \ldots, x_d) \in \mathbb{R}^d : \frac{|x_1|}{a_1} + \cdots + \frac{|x_d|}{a_d} \leq 1\right\},$$

where $a_1, \ldots, a_d > 0$, and let $t > 1$ be real. We call the problem of estimating $\left|tP \cap \mathbb{Z}^d\right|$ the polyhedral sphere problem. Notice that P is both a polyhedron, and the unit ball with respect to the norm

$$(x_1, \ldots, x_d) \mapsto \frac{|x_1|}{a_1} + \cdots + \frac{|x_d|}{a_d}$$

on $\mathbb{R}^d$, justifying the terminology. We studied this problem in the case when all of the coefficients $a_1, \ldots, a_d$ are algebraic.

In this generality, the main term of $\left|tP \cap \mathbb{Z}^d\right|$ was identified as a polynomial $p(t)$ of the variable t. The two highest degree terms of $p(t)$ are

$$p(t) = \frac{2^d a_1 \cdots a_d}{d!} t^d + \frac{2^{d-2} a_1 \cdots a_d}{3(d-2)!} \sum_{i=1}^{d} \frac{1}{a_i^2} t^{d-2} + \cdots.$$

It turns out, that in every term of $p(t)$ the exponent of t is congruent to d modulo 2, and every coefficient is a symmetric rational function of $a_1, \ldots, a_d$ with rational coefficients. As expected, the leading coefficient is $\lambda(P)$, but the other coefficients of $p(t)$ do not

seem to have a natural geometric interpretation in terms of the polytope P. For a general formula of $p(t)$ and its main properties see Definition 3.1 and Proposition 3.4 in subsection 3.2.1. While the general formula is somewhat complicated, we want to emphasize that it is explicit.

The difference $\left|tP \cap \mathbb{Z}^d\right| - p(t)$ is considered a purely random fluctuation. We first studied the expected value

$$\frac{1}{T_2 - T_1} \int_{T_1}^{T_2} \left(\left|tP \cap \mathbb{Z}^d\right| - p(t)\right) \, \mathrm{d}t,$$

where $1 \leq T_1 < T_2$. As our first main result we prove that if $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$ are algebraic and linearly independent over $\mathbb{Q}$, and the length of the interval satisfies $T_2 - T_1 \geq 1$, then

$$\frac{1}{T_2 - T_1} \int_{T_1}^{T_2} \left(\left|tP \cap \mathbb{Z}^d\right| - p(t)\right) \, \mathrm{d}t = O(1).$$

The implied constant depends only on $a_1, \ldots, a_d$, but is ineffective. The significance of this result is that it shows that $p(t)$ is indeed the main term. In other words, every coefficient of the polynomial $p(t)$, except for the constant term, has an actual meaning related to the polyhedral sphere problem. It is possible to extend this result to intervals $[T_1, T_2]$ of arbitrary length. If $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$ are algebraic and linearly independent over $\mathbb{Q}$, then for any $1 \leq T_1 < T_2$ and any $\varepsilon > 0$ we have

$$\frac{1}{T_2 - T_1} \int_{T_1}^{T_2} \left(\left|tP \cap \mathbb{Z}^d\right| - p(t)\right) \, \mathrm{d}t = O\left(1 + \left(\frac{1}{T_2 - T_1}\right)^{\frac{d-2}{d-1}+\varepsilon}\right),$$

where the implied constant depends only on $a_1, \ldots, a_d$ and ε. For a proof see Theorem 3.5 in subsection 3.2.2.

We were also able to find a uniform bound on the fluctuation. If $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$ are algebraic and linearly independent over $\mathbb{Q}$, then for any $\varepsilon > 0$ we have

$$\left|tP \cap \mathbb{Z}^d\right| - p(t) = O\left(t^{\frac{(d-1)(d-2)}{2d-3}+\varepsilon}\right),$$

as $t \to \infty$. Note that the exponent $\frac{(d-1)(d-2)}{2d-3} + \varepsilon$ is simply ε in dimension $d = 2$, it is $\frac{2}{3} + \varepsilon$ in dimension $d = 3$, and roughly $\frac{d}{2}$ when d is large. For a proof see Theorem 3.7 in subsection 3.2.3.

Seeing these results the natural question arises whether we can relax the condition on the linear independence of $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$, and still get nontrivial bounds. The answer is yes. The relaxed condition we worked with, is that $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$ are algebraic, and any k of them are linearly independent over $\mathbb{Q}$ for some $2 \leq k \leq d$. Note that $k = d$ means linear independence, while the weakest condition $k = 2$ simply means that the ratios $\frac{a_i}{a_j}$ are irrational for $i \neq j$. Under these conditions we were able to prove the uniform bound

$$\left|tP \cap \mathbb{Z}^d\right| - p(t) = O\left(t^{\frac{(d-1)(2d-k-3)}{2d-4}+\varepsilon}\right),$$

as $t \to \infty$ for any $\varepsilon > 0$, in the case $2 \leq k \leq d-1$. Note that $\frac{(d-1)(2d-k-3)}{2d-4} < d - \frac{k+1}{2}$. This means that even under the weakest condition $k = 2$ we were able to improve the trivial discrepancy bound $O\left(t^{d-1}\right)$.

The relevance of the relaxed condition comes from the fact that it can be difficult to prove the linear independence of a large number of algebraic numbers over $\mathbb{Q}$, but it can be much easier to prove that any k of them are independent. As an illustration for the case $k = 3$, consider d elements of the field $\mathbb{Q}\left(\sqrt[3]{2}\right)$. Since $1, \sqrt[3]{2}, \sqrt[3]{4}$ is a basis of $\mathbb{Q}\left(\sqrt[3]{2}\right)$ over $\mathbb{Q}$, every element can be written in the form $a + b\sqrt[3]{2} + c\sqrt[3]{4}$ for some $a, b, c \in \mathbb{Q}$. To prove that any 3 of the d elements are linearly independent over $\mathbb{Q}$, it is therefore enough to check that a given $3 \times d$ matrix has full rank over $\mathbb{Q}$.

1.2.2 The orthogonal simplex problem

Consider the orthogonal simplex

$$S = \left\{(x_1, \ldots, x_d) \in \mathbb{R}^d : x_1, \ldots, x_d \geq 0, \frac{x_1}{a_1} + \cdots + \frac{x_d}{a_d} \leq 1\right\},$$

where $a_1, \ldots, a_d > 0$, and let $t > 1$ be real. We call the problem of estimating $\left|tS \cap \mathbb{Z}^d\right|$ the orthogonal simplex problem. The orthogonal simplex problem can easily be reduced to the polyhedral sphere problem using an inclusion-exclusion type argument. This reduction yields that the main term of $\left|tS \cap \mathbb{Z}^d\right|$ is a polynomial $q(t)$ of the variable t related to and derived from the polynomial $p(t)$ as in the polyhedral sphere problem. The three highest degree terms of $q(t)$ are

$$q(t) = \frac{a_1 \cdots a_d}{d!} t^d + \frac{a_1 \cdots a_d}{2(d-1)!} \sum_{i=1}^{d} \frac{1}{a_i} t^{d-1} +$$

$$\frac{a_1 \cdots a_d}{(d-2)!} \left(\frac{1}{12} \sum_{1 \le i \le d} \frac{1}{a_i^2} + \frac{1}{4} \sum_{1 \le i < j \le d} \frac{1}{a_i a_j} \right) t^{d-2} + \cdots .$$

Note that the leading coefficient of $q(t)$ is $\lambda(S)$, while the coefficient of t^{d-1} is one half of the surface area of the orthogonal hyperfaces of S. Thus $q(t)$ is the analogue of an Ehrhart polynomial, even though S is not a lattice polytope. In the case when $a_1, \ldots, a_d$ are positive integers, the same simplex S does have an actual Ehrhart polynomial, which is a classical example studied in the literature. Let us now compare the coefficient of t^{d-2} of $q(t)$ and the Ehrhart polynomial of S. It is known that if $a_1, \ldots, a_d > 0$ are pairwise relatively prime integers, then the coefficient of t^{d-2} in the Ehrhart polynomial of S is

$$\frac{a_1 \cdots a_d}{(d-2)!} \left(\frac{1}{12} \sum_{1 \le i \le d} \frac{1}{a_i^2} + \frac{1}{4} \sum_{1 \le i < j \le d} \frac{1}{a_i a_j} \right) +$$

$$\frac{1}{(d-2)!} \left(\frac{d}{4} + \frac{1}{12 a_1 \cdots a_d} - \sum_{i=1}^{d} D\left(\frac{a_1 \cdots a_d}{a_i}, a_i \right) \right),$$

where D denotes the Dedekind sum defined as

$$D(a,b) = \sum_{k=1}^{b-1} \left(\frac{k}{b} - \frac{1}{2} \right) \left(\left\{ \frac{ak}{b} \right\} - \frac{1}{2} \right)$$

for relatively prime positive integers a, b. This coefficient has been computed using different methods: in [20] toric varieties, in [6] and [7] Fourier analysis, while in [2] the residue theorem is used. The other coefficients of the Ehrhart polynomial of S are also known. They involve higher dimensional and more complicated generalizations of the Dedekind sum D. These complicated terms are completely absent from our polynomial $q(t)$ describing the case when $a_1, \ldots, a_d$ are allowed to be irrational.

The results on the polyhedral sphere problem carry over to the orthogonal simplex problem via the reduction. If $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$ are algebraic and linearly independent over $\mathbb{Q}$, then

$$\left|tS \cap \mathbb{Z}^d\right| - q(t) = O\left(t^{\frac{(d-1)(d-2)}{2d-3}+\varepsilon}\right),$$

as $t \to \infty$ for any $\varepsilon > 0$. If we only assume that $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$ are algebraic, and any k of them are linearly independent over $\mathbb{Q}$ for some $2 \leq k \leq d-1$, then

$$\left|tS \cap \mathbb{Z}^d\right| - q(t) = O\left(t^{\frac{(d-1)(2d-k-3)}{2d-4}+\varepsilon}\right),$$

as $t \to \infty$ for any $\varepsilon > 0$. It would also be straightforward to carry over the results on the expected value from the polyhedral sphere problem. See subsection 3.2.4 for the precise reduction of the orthogonal simplex problem to the polyhedral sphere problem, the general formula for $q(t)$ and the proofs.

Note that in dimension $d = 2$ the orthogonal simplex S is the closed right triangle with vertices $(0,0), (a_1,0), (0,a_2)$ studied by Hardy and Littlewood in [11] and [12]. Thus when $\frac{a_2}{a_1}$ is an algebraic irrational we recover their result $\left|tS \cap \mathbb{Z}^d\right| - q(t) = O\left(t^{\alpha}\right)$ for some $\alpha < 1$, and improve it to $\left|tS \cap \mathbb{Z}^d\right| - q(t) = O\left(t^{\varepsilon}\right)$ for any $\varepsilon > 0$.

Altogether we found that for the polytope P as in the polyhedral sphere problem and for the simplex S as in the orthogonal simplex problem for any real $t > 1$ the discrepancies $\left|tP \cap \mathbb{Z}^d\right| - \lambda(P)t^d$ and $\left|tS \cap \mathbb{Z}^d\right| - \lambda(S)t^d$ can be written as the sum of a nontrivial, explicitly computable polynomial of the real variable t, and a purely random fluctuation. To our knowledge this phenomenon has not yet been described in the literature for any convex body.

The closest result of this type is about the torus in $\mathbb{R}^3$. For constants $0 < b < a$ consider the torus

$$B = \left\{(x,y,z) \in \mathbb{R}^3 : \left(\sqrt{x^2+y^2} - a\right)^2 + z^2 \leq b^2\right\}.$$

Nowak in [18] proves that for any real $t > 1$ and any $\varepsilon > 0$ we have

$$\left|tB \cap \mathbb{Z}^3\right| = \lambda(B)t^3 + F_{a,b}(t)t^{\frac{3}{2}} + O\left(t^{\frac{11}{8}+\varepsilon}\right), \tag{1.2}$$

where $F_{a,b}(t)$ is a bounded, periodic function given by the absolutely convergent trigonometric series

$$F_{a,b}(t) = 4a\sqrt{b}\sum_{n=1}^{\infty} n^{-\frac{3}{2}} \sin\left(2\pi nbt - \frac{\pi}{4}\right).$$

Thus for the torus B the discrepancy $\left|tB \cap \mathbb{Z}^3\right| - \lambda(B)t^3$ can be written as the sum of a computable quantity and a purely random fluctuation. The fact that B is not convex is a technicality. The main difference between this and our result is the nature of the computable quantity in the discrepancy.

1.2.3 Polytopes with coordinates in a quadratic field

Let $D > 1$ be a square-free integer, and consider the real quadratic field $\mathbb{Q}\left(\sqrt{D}\right)$. We studied the class of polytopes P with vertices in $\mathbb{Q}\left(\sqrt{D}\right)^d$. From an algebraic point of view, this is the second simplest class of polytopes, the simplest class being the lattice polytopes. A polytope can either be given as the convex hull of finitely many points, or given as finitely many linear inequalities defining a bounded set. It is easy to see that we would get the same class of polytopes by stipulating that the linear inequalities have coefficients in $\mathbb{Q}\left(\sqrt{D}\right)$.

Consider a polytope P with vertices in $\mathbb{Q}\left(\sqrt{D}\right)^d$ which contains the origin in its interior. It is easy to see that the discrepancy $\left|tP \cap \mathbb{Z}^d\right| - \lambda(P)t^d$ of such a polytope is both $O\left(t^{d-1}\right)$ and $\Omega\left(t^{d-2}\right)$, as $t \to \infty$. Indeed, since the coordinates of the normal vector of a hyperface span a vector space of dimension at most 2 over $\mathbb{Q}$, there exists a $d-2$ dimensional sublattice of $\mathbb{Z}^d$ orthogonal to the normal vector, implying that $\left|tP \cap \mathbb{Z}^d\right|$ as a function of the real variable t has jumps of size constant times t^{d-2}. Here the term normal vector simply means a nonzero vector of arbitrary length orthogonal to a hyperface.

If there exists a hyperface of P with a normal vector in $\mathbb{Q}^d$, then $\left|tP \cap \mathbb{Z}^d\right|$ has jumps of size constant times t^{d-1}, implying that the discrepancy is $\Omega\left(t^{d-1}\right)$, as $t \to \infty$. We were able to prove a general discrepancy bound when this trivial reason for the number of lattice points to have large jumps is not present. If P is a polytope with vertices in $\mathbb{Q}\left(\sqrt{D}\right)^d$, and no hyperface of P has a normal vector in $\mathbb{Q}^d$, then

$$\left|tP \cap \mathbb{Z}^d\right| - \lambda(P)t^d = O\left(t^{d-\frac{8}{7}} \log^{\frac{4}{7}} t\right),$$

as $t \to \infty$. The implied constant depends only on the polytope P, and is effective. Thus the possible order of magnitude of the discrepancy of a polytope with vertices in $\mathbb{Q}\left(\sqrt{D}\right)^d$ exhibits a "gap". Moreover, the discrepancy is as large as possible only when a trivial condition is satisfied. See Theorem 3.8 in section 3.3 for a proof.

This result is far from being best possible in dimension $d = 2$. A general discrepancy bound in [13] implies that if P is a polygon in $\mathbb{R}^2$ with vertices in $\mathbb{Q}\left(\sqrt{D}\right)^2$, such that no side of P has a normal vector in $\mathbb{Q}^2$, then

$$\left|tP \cap \mathbb{Z}^2\right| - \lambda(P)t^2 = O\left(\log t\right),$$

as $t \to \infty$. Since every such polygon can be decomposed into right triangles with axis parallel legs and a hypotenuse with slope in $\mathbb{Q}\left(\sqrt{D}\right)$, this basically follows from the results of Hardy and Littlewood on the number of lattice points in a right triangle in [11] and [12]. In higher dimensions, however, a polytope might not be decomposable into orthogonal simplices. Therefore the proof of our result in arbitrary dimension is not a simple reduction to the orthogonal simplex problem.

The case $d = 3$ already yields nontrivial results. As an illustration, the lattice point counting problem for the regular dodecahedron is studied. First note that the regular dodecahedron cannot be embedded into $\mathbb{R}^3$ in such a way that all of its vertices are lattice points. Since the faces of the regular dodecahedron are regular pentagons, this easily follows from the fact that there does not exist a regular pentagon in $\mathbb{Z}^3$. For a detailed proof see [14]. This means that the regular dodecahedron does not have an Ehrhart polynomial giving the precise number of lattice points in its integral dilates.

It is possible, however, to embed the regular dodecahedron into $\mathbb{R}^3$ so that all of its vertices are in $\mathbb{Q}\left(\sqrt{5}\right)^3$. Indeed, the standard regular dodecahedron D as given in [5] has the 20 vertices

$$\left(0, \pm\varphi^{-1}, \pm\varphi\right), \quad \left(\pm\varphi, 0, \pm\varphi^{-1}\right), \quad \left(\pm\varphi^{-1}, \pm\varphi, 0\right), \quad \left(\pm 1, \pm 1, \pm 1\right),$$

where $\varphi = \frac{1+\sqrt{5}}{2}$ is the golden ratio. The equations of the 12 faces of D are

$$\varphi x \pm y = \pm\varphi^{-1}, \quad \varphi y \pm z = \pm\varphi^{-1}, \quad \varphi z \pm x = \pm\varphi^{-1},$$

showing that no face of D has a normal vector in $\mathbb{Q}^3$. Therefore Theorem 3.8 applies yielding

$$\left|tD \cap \mathbb{Z}^3\right| - \lambda(D)t^3 = O\left(t^{\frac{13}{7}} \log^{\frac{4}{7}} t\right),$$

as $t \to \infty$, with an effective implied constant.

1.2.4 The methods used

The main methods used in the proofs are Fourier analysis and the theory of Diophantine approximation. For a polytope P in $\mathbb{R}^d$ and a real number $t > 1$ consider the Poisson summation formula

$$\left|tP \cap \mathbb{Z}^d\right| = \sum_{n \in \mathbb{Z}^d} \chi_{tP}(n) \sim \sum_{m \in \mathbb{Z}^d} \hat{\chi}_{tP}(m), \tag{1.3}$$

where χ_{tP} is the characteristic function of tP, and $\hat{f}$ denotes the Fourier transform of the function f, as in Definition 2.1 in section 2.1. The right hand side of (1.3) is considered a formal series, and the symbol $\sim$ means we might not have equality. For any integer $N > 0$ consider the Nth Cesàro mean $C(tP, N)$ of this formal series, defined as

$$C(tP, N) = \frac{1}{N^d} \sum_{M \in [0, N-1]^d} \sum_{m \in [-M_1, M_1] \times \cdots \times [-M_d, M_d]} \hat{\chi}_{tP}(m),$$

where $M = (M_1, \ldots, M_d)$. In Theorem 2.5 in section 2.3 we prove that if the vertices of P have algebraic coordinates, then $C(tP, N)$ approximates the number of lattice points in tP with an explicit error term, up to an ineffective constant factor. The main ingredients of the proof are the properties of the Fejér kernel, and Theorem 2.4 of Schmidt on simultaneous Diophantine approximation, which is used to bound the number of lattice points close to the boundary of tP. For a special class of polytopes

we were able to avoid using the theorem of Schmidt resulting in an effective error term: see Theorem 2.7 in section 2.3.

Recall that if P is as in the polyhedral sphere problem, then the main term of $\left|tP \cap \mathbb{Z}^d\right|$ is a polynomial $p(t)$ defined in Definition 3.1 in subsection 3.2.1. Let us offer an intuitive understanding based on Fourier analysis of the curious fact, that in every term of $p(t)$ the exponent of t is congruent to d modulo 2. Applying the integral transformation $x \mapsto tx$ in the definition of the Fourier transform, we get that for any $t > 0$

$$\hat{\chi}_{tP}(m) = \int_{tP} e^{-2\pi i\langle m,x\rangle}\,\mathrm{d}x = t^d \int_P e^{-2\pi i\langle m,x\rangle t}\,\mathrm{d}x.$$

Note that this is not true for negative values of t. We also have

$$\hat{\chi}_{tP}(m) + \hat{\chi}_{tP}(-m) = 2t^d \int_P \cos\left(2\pi\langle m,x\rangle t\right)\,\mathrm{d}x.$$

The right hand side is clearly an entire function of the variable t, depending on m, which satisfies the functional equation $f(-t) = (-1)^d f(t)$. Since the Cesàro means $C(tP,N)$ are weighted sums of $\hat{\chi}_{tP}(m)$ with equal weights given to $\hat{\chi}_{tP}(m)$ and $\hat{\chi}_{tP}(-m)$, for any integer $N > 0$ the Cesàro mean $C(tP,N)$ also has an analytic continuation as an entire function satisfying the functional equation $f(-t) = (-1)^d f(t)$. It is therefore not surprising, that the polynomial $p(t)$, which is the main term of $C(tP,N)$, also satisfies $p(-t) = (-1)^d p(t)$, resulting in a zero coefficient for every power of t the exponent of which is not congruent to d modulo 2.

It might be worth mentioning, that for any lattice polytope P the polynomial

$$f(t) = \left|tP \cap \mathbb{Z}^d\right| - \frac{1}{2}\left|t\left(\partial P\right) \cap \mathbb{Z}^d\right|$$

is known to satisfy $f(-t) = (-1)^d f(t)$ for integral values of t: this is one of the equivalent forms of the famous Ehrhart–Macdonald reciprocity first proved in [16]. For P as in the polyhedral sphere problem the lattice points on the boundary are treated as an error term. Therefore the fact that the main term $p(t)$ satisfies $p(-t) = (-1)^d p(t)$ means that $p(t)$ is the analogue of an Ehrhart polynomial, even though P is not a lattice polytope.

To study the Cesàro means $C(tP, N)$ we needed to find the Fourier transform $\hat{\chi}_{tP}$ for an arbitrary polytope P. We found a representation of this Fourier transform in the special case when P is a simplex, which to our knowledge has not yet appeared in the literature. In Theorem 3.1 in section 3.1 we prove that

$$\hat{\chi}_{tS}(m) = \frac{(-1)^d d!}{(2\pi i)^{d+1}} \lambda(S) \int_{|z|=R} \frac{e^{-2\pi i z t}}{(z - \langle m, v_1 \rangle) \cdots (z - \langle m, v_{d+1} \rangle)} \, \mathrm{d}z, \tag{1.4}$$

where S is an arbitrary simplex in $\mathbb{R}^d$ with vertices $v_1, \ldots, v_{d+1}$, and $R > 0$ is large enough so that all the singularities of the integrand are inside the circle $|z| = R$. Since the variable t appears only in the complex exponential function, (1.4) is some kind of Fourier expansion of $\hat{\chi}_{tS}(m)$ in the variable t. The Fourier transform $\hat{\chi}_{tP}(m)$ for a general polytope P can be found by first decomposing P into simplices, then using (1.4) on all the simplices in the triangulation of P. The complex line integral in (1.4) is evaluated using the residue theorem.

In the special case, when P is as in the polyhedral sphere problem, we applied the representation (1.4) on the orthogonal simplices used to triangulate P, to find $\hat{\chi}_{tP}(m)$. The Cesàro means $C(tP, N)$ are certain weighted sums of $\hat{\chi}_{tP}(m)$. In subsection 3.2.1 we were able to find the contribution of all the residues at $z = 0$ in $C(tP, N)$ up to a small error term: this became the polynomial $p(t)$ playing the role of the main term of $\left|tP \cap \mathbb{Z}^d\right|$. The contribution of all the other residues in $C(tP, N)$ was treated as a random fluctuation. It might be worth mentioning that the high degree terms of $p(t)$ come from high order poles of the integrand in (1.4). The proof thus offers an intuitive understanding of the fact that the discrepancy $\left|tP \cap \mathbb{Z}^d\right| - \lambda(P)t^d$ is the sum of a polynomial and a random fluctuation.

When we tried to bound the random fluctuation in the polyhedral sphere problem, that is, the contribution of all residues in (1.4) other than the residue at $z = 0$, we encountered a simultaneous Diophantine approximation problem. The solution of that problem seems to be interesting in its own right. We proved that if $\alpha_1, \ldots, \alpha_d$ are algebraic reals such that $1, \alpha_1, \ldots, \alpha_d$ are linearly independent over $\mathbb{Q}$, then for any integer $M > 0$ and any real $\varepsilon > 0$ we have

$$\sum_{m=1}^{M} \frac{1}{\|m\alpha_1\| \cdots \|m\alpha_d\|} = O\left(M^{2-\frac{1}{d}+\varepsilon}\right),$$

where the implied constant is ineffective, and where $\|x\|$ denotes the distance of a real number x from the nearest integer. This result is well known for $d = 1$, moreover it is known that the exponent $1 + \varepsilon$ in the error term is best possible. We were unable to find the case $d \geq 2$ in the literature, thus it seems to be a new result. Unfortunately we do not know if the exponent $2 - \frac{1}{d} + \varepsilon$ is best possible in general. The proof is the combination of Theorem 2.3 of Schmidt on simultaneous Diophantine approximation and the pigeonhole principle. For a slightly more general form and the proof see Proposition 3.6 in subsection 3.2.2.

Finally, let us consider a polytope P with vertices in $\mathbb{Q}\left(\sqrt{D}\right)^d$, where $D > 1$ is a square-free integer, such that no hyperface of P has a normal vector in $\mathbb{Q}^d$. We were able to show that every such polytope can be triangulated into simplices satisfying the same conditions. Thus we can use the representation (1.4) to bound the discrepancy of such simplices. We found a general way of bounding the contribution of all the residues in (1.4) for every lattice point $m \neq 0$, which is related to a Diophantine approximation problem. The reason we worked with the quadratic field $\mathbb{Q}\left(\sqrt{D}\right)$, is that this Diophantine approximation problem has an effective solution for quadratic irrationals. See Theorem 3.8 in section 3.3 for a proof.

1.2.5 Lattice points in a right triangle

Consider the closed right triangle S with vertices $(0,0), (1,0), (0,\alpha)$, where $\alpha > 0$ is irrational. The problem of estimating $\left|tS \cap \mathbb{Z}^2\right|$ for real numbers $t > 0$ has been studied by Hardy and Littlewood in [11] and [12], and by Beck in section 4.5 of [1]. It is not difficult to come up with the intuition that this problem is related to the problem of approximating α by rationals of small denominators. Since the continued fraction representation of α provides an effective solution to this Diophantine approximation problem, instead of assuming that α is algebraic, we expressed our results in terms of the partial quotients of α.

The main term of $\left|tS \cap \mathbb{Z}^2\right|$ was identified as

$$g(t) = \frac{\alpha}{2}t^2 + \frac{\alpha+1}{2}t + \frac{(\alpha\{t\}+1)(1-\{t\})}{2}.$$

The difference $\left|tS \cap \mathbb{Z}^2\right| - g(t)$ is considered a random fluctuation. Consider the variance

$$\frac{1}{T}\int_0^T \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right)^2 \, \mathrm{d}t \tag{1.5}$$

of this fluctuation, where $T \geq 1$ is real. In [1] Beck evaluates (1.5) in the special case when α is a quadratic irrational in terms of arithmetic quantities of the real quadratic field $\mathbb{Q}(\alpha)$. In section 4.5 of [1] Beck raises the question whether given the continued fraction representation $\alpha = [a_0; a_1, a_2, \ldots]$ of an arbitrary irrational $\alpha > 0$, the variance (1.5) can be evaluated in terms of the partial quotients a_k. We were able to prove that if the partial quotients satisfy $a_k = O\left(k^d\right)$ for some real number $d \geq 0$, then using the notation $\frac{p_k}{q_k} = [a_0; a_1, a_2, \ldots, a_{k-1}]$ for the convergents to α, we have

$$\frac{1}{T}\int_0^T \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right)^2 \, \mathrm{d}t = \frac{1}{360}\sum_{q_k<T} a_k^2 + O\left(\log^{d+1} T + \log^{2d} T \log\log T\right), \tag{1.6}$$

as $T \to \infty$. The sum is over all positive integers k such that the kth convergent denominator q_k is less than T. The implied constant depends only on α and is effective.

In the special case when α is a quadratic irrational, the partial quotients satisfy the condition $a_k = O\left(k^d\right)$ with $d = 0$. In this case the order of magnitude of both the main term and the error term of (1.6) is constant times $\log T$, which makes our result not applicable. If, on the other hand α is Euler's number e, then we have the continued fraction representation

$$e = [2; 1, 2, 1, 1, 4, 1, \ldots, 1, 2n, 1, \ldots], \tag{1.7}$$

which means that the condition $a_k = O\left(k^d\right)$ is satisfied with $d = 1$. Note that there is a whole class of irrational numbers related to Euler's number e with a continued fraction representation similar to (1.7), satisfying $a_k = O\left(k^d\right)$ with $d = 1$, including $e^{\frac{2}{n}}$ for any positive integer n. For this class of irrationals the order of magnitude of the

main term of (1.6) is constant times $\left(\frac{\log T}{\log\log T}\right)^3$, which is larger than the error term $O\left(\log^2 T \log\log T\right)$. Thus our result (1.6) complements, rather than generalizes the results of Beck on the variance (1.5).

In section 4.1 we collected the facts about continued fractions and Diophantine approximation used in the proofs. Section 4.2 is dedicated to the proof of (1.6).

1.3 Trivial discrepancy bound for polytopes

Consider an arbitrary polytope P in $\mathbb{R}^d$. We conclude chapter 1 by discussing a trivial way to bound the discrepancy $\left|P \cap \mathbb{Z}^d\right| - \lambda(P)$. The idea we use is quite simple. For every lattice point $n \in P \cap \mathbb{Z}^d$ consider the axis parallel cube of unit side length $\left[-\frac{1}{2}, \frac{1}{2}\right] + n$ centered at n. On the one hand, the total Lebesgue measure of these cubes is $\left|P \cap \mathbb{Z}^d\right|$. On the other hand, the total Lebesgue measure of these cubes is close to $\lambda(P)$, the error being related to the number lattice points close to the boundary of P.

One could carelessly think that this simple idea shows

$$\left|P \cap \mathbb{Z}^d\right| - \lambda(P) = O\left(\mathrm{Surf}(P)\right), \tag{1.8}$$

where $\mathrm{Surf}(P)$ denotes the surface area of P. The general discrepancy bound (1.8), however, is false in dimensions $d \geq 3$. Perhaps the simplest counterexample is the polytope

$$P = [1, N] \times [-a, a] \times \cdots \times [-a, a],$$

where $N > 1$ is an integer and $0 < a < 1$. For this particular polytope we have $\left|P \cap \mathbb{Z}^d\right| = N$, but the Lebesgue measure and the surface area of P can be made arbitrarily small by choosing a small enough. This shows that there does not exist a universal implied constant for the class of all polytopes which would make (1.8) true.

Nevertheless it is possible to turn our simple idea into a precise proof, yielding a universal trivial discrepancy bound for the class of all polytopes in terms of simple geometric quantities. Moreover the trivial bound we prove below is invariant under isometries.

Proposition 1.1 (Trivial discrepancy bound for polytopes) *Let $d \geq 2$ be an integer, and let P be a polytope in $\mathbb{R}^d$. Let $R(P) > 0$ be the radius of a closed ball which covers P, and let $H(P)$ denote the number of hyperfaces of P. Then*

$$\left|\left|P \cap \mathbb{Z}^d\right| - \lambda(P)\right| \leq \omega(d-1)\sqrt{d}H(P)\left(R(P) + \frac{\sqrt{d}}{2}\right)^{d-1},$$

where $\omega(d-1)$ denotes the Lebesgue measure of the $d-1$ dimensional unit ball.

Proof: For a lattice point $n \in \mathbb{Z}^d$ let $C_n = \left[-\frac{1}{2}, \frac{1}{2}\right]^d + n$ denote the axis parallel cube with unit side length centered at n. Let

$$A = \left\{x \in \mathbb{R}^d : \operatorname{dist}(x, \partial P) \leq \frac{\sqrt{d}}{2}\right\},$$

where $\operatorname{dist}(x, S)$ denotes the distance of the point x from the set S, and ∂P denotes the boundary of P. Then we have

$$P \backslash A \subseteq \bigcup_{n \in P \cap \mathbb{Z}^d} C_n \subseteq P \cup A. \tag{1.9}$$

Indeed, let $x \in P \backslash A$ be arbitrary. Let $n \in \mathbb{Z}^d$ be such that $x \in C_n$. (Note that this n might not be unique.) The distance of any point of C_n from n is at most $\frac{\sqrt{d}}{2}$, therefore $|x - n| \leq \frac{\sqrt{d}}{2}$. Since x is in P, and its distance from ∂P is more than $\frac{\sqrt{d}}{2}$, we have that $n \in P$. Therefore this particular C_n shows up in the union in (1.9), hence

$$x \in \bigcup_{n \in P \cap \mathbb{Z}^d} C_n.$$

To see the second containment in (1.9), let $n \in P \cap \mathbb{Z}^d$ be arbitrary, and let $x \in C_n$. Then $|x - n| \leq \frac{\sqrt{d}}{2}$, therefore either $x \in P$ or $\operatorname{dist}(x, \partial P) \leq \frac{\sqrt{d}}{2}$, showing $x \in P \cup A$.

By taking the Lebesgue measures of the sets in (1.9) we obtain

$$\lambda\left(P \backslash A\right) \leq \lambda\left(\bigcup_{n \in P \cap \mathbb{Z}^d} C_n\right) \leq \lambda(P \cup A).$$

Since the Lebesgue measure of the union is simply $\left|P \cap \mathbb{Z}^d\right|$, it follows that

$$\lambda(P) - \lambda(A) \leq \left|P \cap \mathbb{Z}^d\right| \leq \lambda(P) + \lambda(A),$$

$$\left|\left|P \cap \mathbb{Z}^d\right| - \lambda(P)\right| \leq \lambda(A). \tag{1.10}$$

It is therefore enough to find an upper bound to $\lambda(A)$.

For a hyperface H of P let

$$A_H = \left\{ x \in \mathbb{R}^d : \operatorname{dist}(x, H) \leq \frac{\sqrt{d}}{2} \right\}.$$

Since the hyperfaces cover ∂P, we have that $A = \cup_H A_H$, where the union is taken over all hyperfaces of P. Consider now a closed ball B of radius $R(P)$ which covers P, and an arbitrary hyperface H. The affine hyperplane containing H intersects B in a $d-1$ dimensional closed ball B_H of radius at most $R(P)$. We have $H \subseteq B_H$. It is now easy to see, that A_H can be covered by a d dimensional cylinder the base of which is a $d-1$ dimensional ball of radius at most $R(P) + \frac{\sqrt{d}}{2}$, and the height of which is $\sqrt{d}$. Therefore $\lambda(A_H)$ is at most the Lebesgue measure of this cylinder:

$$\lambda(A_H) \leq \omega(d-1) \left(R(P) + \frac{\sqrt{d}}{2} \right)^{d-1} \cdot \sqrt{d}.$$

Using this estimate together with $A = \cup_H A_H$ in (1.10) finishes the proof.

□

Corollary 1.2 *Let $d \geq 2$ be an integer, and let P be a polytope in $\mathbb{R}^d$. Then for every $t > 1$ we have*

$$\left|tP \cap \mathbb{Z}^d\right| - \lambda(P)t^d = O\left(t^{d-1}\right).$$

The implied constant depends only on P, and is effective.

Proof: We can apply Proposition 1.1 on the polytope tP with $\lambda(tP) = \lambda(P)t^d$, $H(tP) = H(P)$, and $R(tP) = R(P)t$. The upper bound to the discrepancy we obtain is a polynomial of degree $d-1$ in the variable t, the coefficients of which are explicitly expressed in terms of d, $H(P)$ and $R(P)$.

□

Chapter 2

Poisson summation formula for polytopes

2.1 The general approach

The main tool for estimating the number of lattice points in a given polytope will be Fourier analysis, more specifically the Poisson summation formula. Let us fix some terminology and notation first.

Definition 2.1 *Let $d \geq 1$ be an integer, and $f : \mathbb{R}^d \to \mathbb{R}$ be Lebesgue integrable on $\mathbb{R}^d$. The Fourier transform of f is the function $\hat{f} : \mathbb{R}^d \to \mathbb{C}$ defined as*

$$\hat{f}(y) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i \langle x,y \rangle} \, \mathrm{d}x \qquad \left(y \in \mathbb{R}^d \right),$$

where $\langle x, y \rangle$ denotes the scalar product of the vectors $x, y \in \mathbb{R}^d$, and the integral is a Lebesgue integral.

The Poisson summation formula is a celebrated result in Fourier analysis which connects a function f and its Fourier transform $\hat{f}$ by considering the sum of their values over all lattice points. We shall say, that a Lebesgue integrable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the Poisson summation formula, if

$$\sum_{n \in \mathbb{Z}^d} f(n) = \sum_{m \in \mathbb{Z}^d} \hat{f}(m). \tag{2.1}$$

Note that this is a somewhat vague definition, as we did not specify a mode of convergence for the two series. There are several sufficient conditions known which imply that a given function satisfies the Poisson summation formula. For example, if $f : \mathbb{R}^d \to \mathbb{R}$ is arbitrarily many times differentiable, and has a compact support, then it satisfies (2.1) (see e.g. [19]).

Our main observation is that the number of lattice points in a given polytope can

be expressed in the same form as the left hand side of (2.1). Indeed, by introducing the characteristic function χ_P of a polytope P, we have that

$$\left|P \cap \mathbb{Z}^d\right| = \sum_{n \in \mathbb{Z}^d} \chi_P(n). \tag{2.2}$$

Note that since every polytope is bounded, the series on the right hand side of (2.2) has finitely many nonzero terms, therefore we do not encounter any convergence issues. The expression (2.2) gives the idea to try to use Poisson summation on the function χ_P to study the number of lattice points in P. This, however, raises the following question. Is it true that for any polytope P, the characteristic function χ_P satisfies the Poisson summation formula?

The answer is unfortunately no. Even though the function χ_P has a compact support, it is not differentiable (not even continuous), which makes all the known theorems on the Poisson summation formula not applicable. The real reason why the formula does not hold in general, however, is that

$$\sum_{m \in \mathbb{Z}^d} \hat{\chi}_P(m) \tag{2.3}$$

is additive in P, while $\left|P \cap \mathbb{Z}^d\right|$ is not. Indeed, let P_1 and P_2 be the two polytopes obtained by cutting a polytope P into two pieces with an affine hyperplane. Then we have $\hat{\chi}_P = \hat{\chi}_{P_1} + \hat{\chi}_{P_2}$, and therefore

$$\sum_{m \in \mathbb{Z}^d} \hat{\chi}_P(m) = \sum_{m \in \mathbb{Z}^d} \hat{\chi}_{P_1}(m) + \sum_{m \in \mathbb{Z}^d} \hat{\chi}_{P_2}(m),$$

provided that both series on the right hand side converge. On the other hand $\left|P \cap \mathbb{Z}^d\right|$ clearly does not enjoy such an additivity property:

$$\left|P \cap \mathbb{Z}^d\right| \neq \left|P_1 \cap \mathbb{Z}^d\right| + \left|P_2 \cap \mathbb{Z}^d\right|, \tag{2.4}$$

since the lattice points in P lying on the affine hyperplane with which we cut P are counted once on the left hand side, but twice on the right hand side of (2.4).

The rest of the chapter is devoted to studying the relationship between the formal series (2.3), and the number of lattice points in P. More specifically, we will be interested in the following problem. Given a polytope P, which is defined in terms of algebraic numbers, and a magnifying factor $t > 1$, how can we use the formal series (2.3) to estimate the number of lattice points in tP? The observation above implies, that an error term will inevitably appear. It is not difficult to come up with the intuition, that this error term will be related to the lattice points on the boundary of tP, since those points are the reason why the additivity breaks down.

2.2 Cesàro means and the Fejér kernel

We begin by defining the two main quantities associated with the formal series (2.3) which will play a role in estimating the number of lattice points.

Definition 2.2 *Let P be a polytope in $\mathbb{R}^d$, and let $\hat{\chi}_P$ denote the Fourier transform of its characteristic function, as defined in Definition 2.1. Let $M_1, \dots, M_d \geq 0$ be integers, and let $M = (M_1, \dots, M_d)$. We define $S(P, M)$ as*

$$S(P, M) = \sum_{m_1=-M_1}^{M_1} \cdots \sum_{m_d=-M_d}^{M_d} \hat{\chi}_P(m_1, \dots, m_d).$$

Definition 2.3 *Let P be a polytope in $\mathbb{R}^d$, and let $N > 0$ be an integer. We define $C(P, N)$ as*

$$C(P, N) = \frac{1}{N^d} \sum_{M \in [0, N-1]^d} S(P, M).$$

Notice that $S(P, M)$ plays the role of the partial sums, while $C(P, N)$ plays the role of the Cesàro means of the formal series (2.3).

Let us now introduce the two most important kernels in the theory of Fourier series.

Definition 2.4 *Let $M_1, \dots, M_d \geq 0$ be integers, and let $M = (M_1, \dots, M_d)$. We define the Dirichlet kernel $D_M : \mathbb{R}^d \to \mathbb{R}$ as*

$$D_M(x_1, \dots, x_d) = \sum_{m_1=-M_1}^{M_1} \cdots \sum_{m_d=-M_d}^{M_d} e^{2\pi i(m_1 x_1 + \cdots + m_d x_d)} \qquad (x_1, \dots, x_d \in \mathbb{R}).$$

Definition 2.5 *Let $N > 0$ be an integer. We define the Fejér kernel $F_N : \mathbb{R}^d \to \mathbb{R}$ as*

$$F_N(x) = \frac{1}{N^d} \sum_{M \in [0,N-1]^d} D_M(x) \qquad \left(x \in \mathbb{R}^d\right).$$

These kernels are best known in the case $d = 1$. Although there might be several natural ways to generalize them to higher dimensions, the definitions chosen above are going to be very easy to work with. Notice for example, that both these kernels can easily be factored into one dimensional kernels of the same type:

$$D_{(M_1,\ldots,M_d)}(x_1, \ldots, x_d) = D_{M_1}(x_1) \cdots D_{M_d}(x_d),$$

$$F_N(x_1, \ldots, x_d) = F_N(x_1) \cdots F_N(x_d). \tag{2.5}$$

The most important properties of the kernels are listed in the following proposition.

Proposition 2.1 *Let $M_1, \ldots, M_d \geq 0$ be integers, and $M = (M_1, \ldots, M_d)$. Let $N > 0$ be an integer.*

(i) *D_M and F_N are periodic in each coordinate with period 1.*

(ii) *D_M and F_N are even functions.*

(iii) $\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d} D_M(x)\,\mathrm{d}x = 1$ *and* $\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d} F_N(x)\,\mathrm{d}x = 1.$

(iv) *$F_N(x) \geq 0$ for every $x \in \mathbb{R}^d$.*

Proof:

(i)-(iii) Trivial.

(iv) Using the observation (2.5), we can reduce the claim to the special case of $d = 1$. The claim is well known in $d = 1$, see e.g. section 1.4.3 in [19].

□

It is a basic fact in the theory of Fourier series, that the Dirichlet kernel is associated with the partial sums, while the Fejér kernel is associated with the Cesàro means of a Fourier series. We shall use the kernels to study the partial sums $S(P, M)$ and

the Cesàro means $C(P,N)$. The relationship between them is stated in the following proposition.

Proposition 2.2 *Let P be a polytope in $\mathbb{R}^d$, let $M_1,\dots,M_d \geq 0$ be integers, and let $M=(M_1,\dots,M_d)$. Let $N>0$ be an integer, and let $L\in\mathbb{R}$ be arbitrary. Define the function $f:\mathbb{R}^d\to\mathbb{R}$ as*

$$f(x)=\sum_{n\in\mathbb{Z}^d}\chi_P(n+x) \qquad \left(x\in\mathbb{R}^d\right).$$

(i) $S(P,M)-L=\displaystyle\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d}(f(x)-L)\,D_M(x)\,\mathrm{d}x$

(ii) $C(P,N)-L=\displaystyle\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d}(f(x)-L)\,F_N(x)\,\mathrm{d}x$

The proof of Proposition 2.2 is trivial and can be found in any introductory textbook on Fourier analysis, e.g. in [19], therefore will be omitted. Notice that in the definition of $f(x)$ in terms of an infinite series only finitely many terms are nonzero, hence we do not encounter any convergence issues.

An important property of the Fejér kernel is that the main contribution of its integral on $\left[-\frac{1}{2},\frac{1}{2}\right]^d$ comes from a small neighborhood of the origin. We can turn this into a quantitative statement as follows.

Proposition 2.3 *Let $d\geq 1$ be an integer. For every integer $N>0$ and every $0<h<\frac{1}{2}$ we have*

$$0\leq\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d\setminus[-h,h]^d}F_N(x)\,\mathrm{d}x\leq\frac{2d}{\pi}\cdot\frac{1+\log N}{hN}.$$

Even though this proposition is probably well-known, we shall give a proof for the sake of completeness.

Proof: Using Proposition 2.1 (iii) and (iv), we know that the integral is in the interval $[0,1]$. Therefore we may assume, that

$$\frac{2d}{\pi}\cdot\frac{1+\log N}{hN}<1. \tag{2.6}$$

We will first focus on the case $d=1$. Fix an integer $M\geq 0$, and consider

$$\int_{-h}^{h} D_M(x)\,\mathrm{d}x = \int_{-h}^{h} \sum_{m=-M}^{M} e^{2\pi i m x}\,\mathrm{d}x.$$

Since the contribution of the $m = 0$ term is $2h$, we get

$$\int_{-h}^{h} D_M(x)\,\mathrm{d}x = 2h + \sum_{\substack{m=-M \\ m\neq 0}}^{M} \left[\frac{e^{2\pi i m x}}{2\pi i m}\right]_{-h}^{h} = 2h + \sum_{\substack{m=-M \\ m\neq 0}}^{M} \frac{e^{2\pi i m h} - e^{-2\pi i m h}}{2\pi i m} =$$

$$2h + \sum_{\substack{m=-M \\ m\neq 0}}^{M} \frac{\sin(2\pi m h)}{\pi m} = 2h + 2\sum_{m=1}^{M} \frac{\sin(2\pi m h)}{\pi m}. \tag{2.7}$$

In the last step we used the fact, that the terms indexed by m and $-m$ in the sum are equal.

We can identify this last sum as the partial sum of a classical Fourier series. Recall that the series

$$\sum_{m=1}^{\infty} \frac{\sin(2\pi m x)}{\pi m}$$

is convergent for any real number x, and its sum is $\frac{1}{2} - \{x\}$ for every $x \in \mathbb{R}\backslash\mathbb{Z}$. Since $0 < h < \frac{1}{2}$, we can rewrite (2.7) as

$$\int_{-h}^{h} D_M(x)\,\mathrm{d}x = 2h + 2\sum_{m=1}^{\infty} \frac{\sin(2\pi m h)}{\pi m} - 2\sum_{m=M+1}^{\infty} \frac{\sin(2\pi m h)}{\pi m} =$$

$$2h + 2\left(\frac{1}{2} - \{h\}\right) - 2\sum_{m=M+1}^{\infty} \frac{\sin(2\pi m h)}{\pi m} = 1 - 2\sum_{m=M+1}^{\infty} \frac{\sin(2\pi m h)}{\pi m}. \tag{2.8}$$

We will apply summation by parts on the last series in (2.8). To this end, let

$$a_n = \sum_{k=1}^{n} \sin(2\pi k h) = \frac{\cos(\pi h) - \cos((2n+1)\pi h)}{2\sin(\pi h)}.$$

Notice that we have

$$|a_n| \leq \frac{1}{|\sin(\pi h)|}$$

for every n. Since $\sin x$ is concave on $\left[0, \frac{\pi}{2}\right]$, we have that $\sin x \geq \frac{2}{\pi}x$ on the same interval, where $\frac{2}{\pi}x$ is the chord connecting the endpoints of the graph of $\sin x$. We have $\pi h \in \left[0, \frac{\pi}{2}\right]$, therefore $\sin(\pi h) \geq \frac{2}{\pi}\pi h$, which yields

$$|a_n| \leq \frac{1}{2h}$$

for every n. To apply summation by parts, fix an integer $T > M+1$ and consider

$$\left| \sum_{m=M+1}^{T} \frac{\sin(2\pi mh)}{\pi m} \right| = \left| \sum_{m=M+1}^{T} (a_m - a_{m-1}) \frac{1}{\pi m} \right| =$$

$$\left| -a_M \frac{1}{\pi(M+1)} + \sum_{m=M+1}^{T-1} a_m \left(\frac{1}{\pi m} - \frac{1}{\pi(m+1)} \right) + a_T \frac{1}{\pi T} \right| \leq$$

$$|a_M| \frac{1}{\pi(M+1)} + \sum_{m=M+1}^{T-1} |a_m| \left(\frac{1}{\pi m} - \frac{1}{\pi(m+1)} \right) + |a_T| \frac{1}{\pi T} \leq$$

$$\frac{1}{2\pi h(M+1)} + \frac{1}{2h} \sum_{m=M+1}^{T-1} \left(\frac{1}{\pi m} - \frac{1}{\pi(m+1)} \right) + \frac{1}{2\pi hT} =$$

$$\frac{1}{2\pi h(M+1)} + \frac{1}{2h} \left(\frac{1}{\pi(M+1)} - \frac{1}{\pi T} \right) + \frac{1}{2\pi hT} = \frac{1}{\pi h(M+1)}.$$

Taking the limit, as $T \to \infty$, we get

$$\left| \sum_{m=M+1}^{\infty} \frac{\sin(2\pi mh)}{\pi m} \right| \leq \frac{1}{\pi h(M+1)}.$$

Using this estimate in (2.8) we get

$$\int_{-h}^{h} D_M(x)\,\mathrm{d}x \geq 1 - \frac{2}{\pi h(M+1)}.$$

Let us now take the average of this inequality over all integral values of M in $[0, N-1]$ to obtain

$$\int_{-h}^{h} F_N(x)\,\mathrm{d}x \geq 1 - \frac{2}{\pi h} \cdot \frac{1}{N} \sum_{M=0}^{N-1} \frac{1}{M+1} \geq 1 - \frac{2(1+\log N)}{\pi hN}. \tag{2.9}$$

Using the factorization (2.5) together with Fubini's theorem, it is easy to express the integral of the d dimensional Fejér kernel on the cube $[-h, h]^d$ as

$$\int_{[-h,h]^d} F_N(x)\,\mathrm{d}x = \left(\int_{-h}^{h} F_N(x)\,\mathrm{d}x\right)^d.$$

The assumption (2.6) implies that the right hand side of (2.9) is positive, therefore we can raise (2.9) to the dth power to get

$$\int_{[-h,h]^d} F_N(x)\,\mathrm{d}x \geq \left(1 - \frac{2(1+\log N)}{\pi h N}\right)^d. \tag{2.10}$$

Let us consider the general inequality $(1-x)^d \geq 1-dx$, which holds for every $x \in [0,1]$. Indeed, the function $(1-x)^d$ is convex on $[0,1]$, and $1-dx$ is its tangent line at the point $x=0$. The assumption (2.6) implies, that we can apply this general inequality with $x = \frac{2(1+\log N)}{\pi h N}$ in (2.11) to get

$$\int_{[-h,h]^d} F_N(x)\,\mathrm{d}x \geq 1 - \frac{2d}{\pi}\cdot\frac{1+\log N}{hN}.$$

Finally, using Proposition 2.1 (iii):

$$\int_{\left[-\frac{1}{2},\frac{1}{2}\right]\setminus[-h,h]^d} F_N(x)\,\mathrm{d}x \leq \frac{2d}{\pi}\cdot\frac{1+\log N}{hN}.$$

□

2.3 Poisson summation formula with explicit error term

Given a polytope P and a magnifying factor $t > 1$, we want to use the Cesàro means $C(tP, N)$ defined in Definition 2.3 to approximate the number of lattice points in tP. The main results of this chapter are explicit error bounds for this approximation, which hold for different classes of polytopes: see Theorem 2.5 and Theorem 2.7. We will only consider polytopes the vertices of which have algebraic coordinates.

A crucial step in the proofs will be to estimate the number of lattice points which are close to the boundary of tP. The intuitive reason why this quantity is of interest is quite simple. By changing the values of the function χ_{tP} in a small neighborhood of the boundary, we can turn it into an arbitrarily many times differentiable function with a compact support, which therefore will satisfy the Poisson summation formula.

The error we make by replacing χ_{tP} by this new function on the left hand side of (2.1) is at most the number of lattice points in a small neighborhood of the boundary. Even though we will not formally introduce this new, arbitrarily many times differentiable function in the proofs, the intuitive reasoning above clearly motivates the study of the lattice points close to the boundary.

To study the lattice points close to the boundary of our polytope, we will need certain facts from the theory of simultaneous Diophantine approximation. Let us recall two important and deep theorems from that field. Here and from now on $\|x\|$ will denote the distance of the real number x from the nearest integer.

Theorem 2.3 (Schmidt, [21]) *Let $\alpha_1, \ldots, \alpha_d$ be real algebraic numbers, such that $1, \alpha_1, \ldots, \alpha_d$ are linearly independent over $\mathbb{Q}$. Then for every $\varepsilon > 0$, the inequality*

$$\|m\alpha_1\| \cdots \|m\alpha_d\| \leq \frac{1}{m^{1+\varepsilon}}$$

has finitely many integral solutions $m \in \mathbb{N}$.

Theorem 2.4 (Schmidt, [21]) *Let $\alpha_1, \ldots, \alpha_d$ be real algebraic numbers, such that $1, \alpha_1, \ldots, \alpha_d$ are linearly independent over $\mathbb{Q}$. Then for every $\varepsilon > 0$, the inequality*

$$\|m_1\alpha_1 + \cdots + m_d\alpha_d\| \leq \frac{1}{|m|^{d+\varepsilon}}$$

has finitely many integral solutions $m = (m_1, \ldots, m_d) \in \mathbb{Z}^d$.

The proof of these theorems is quite long and complicated. It is possible, however, to show their equivalence using a much simpler technique, called Khintchine's transference principle (section V.3 Theorem IV in [4]). Because of this reason Theorems 2.3 and 2.4 are sometimes called dual versions of each other. An important observation to make about these results is that they provide no upper bound to the absolute value of the solutions, which phenomenon is called ineffectiveness. Also note that Theorem 2.4 is most commonly stated with the maximum norm of m on the right hand side, instead of the Euclidean norm $|m|$ used here. Since every two norms on $\mathbb{R}^d$ are equivalent, it does not matter which norm we use.

We are ready to formulate and prove the first Poisson summation formula with explicit error term.

Theorem 2.5 *Let $2 \leq k \leq d$ be integers, and let P be a polytope in $\mathbb{R}^d$. Suppose that every hyperface of P has a normal vector $v = (v_1, \ldots, v_d)$, such that $v_1, \ldots, v_d$ are all algebraic and span a vector space of dimension at least k over $\mathbb{Q}$. Then for every $t > 1$, every $\varepsilon > 0$ and every integer $N > 1$ we have*

$$C(tP, N) - \left|tP \cap \mathbb{Z}^d\right| = O\left(t^{d-k} + t^{d-1+\varepsilon}\sqrt{\frac{\log N}{N}}\right),$$

where $C(tP, N)$ is as in Definition 2.3. The implied constant depends only on P and ε, and is ineffective.

Notice that an error term of t^{d-k} is basically inevitable. Indeed, if the coordinates of a normal vector v span a vector space of dimension k over $\mathbb{Q}$, then there exist $d-k$ linearly independent rational vectors orthogonal to v. Thus if $t > 1$ is such that tP has at least one lattice point on the given hyperface, then it also contains every lattice point from a $d-k$ dimensional sublattice within a $d-1$ dimensional ball of radius constant times t. In other words $\left|tP \cap \mathbb{Z}^d\right|$ as a function of the real variable t has jumps of size constant times t^{d-k}. If we are to approximate this with a continuous function of t, an error of t^{d-k} is inevitable.

Intuitively, k can be thought of as a measure of how irrational the polytope P is. The case $k = d$ means that the coordinates $v_1, \ldots, v_d$ of the normal vector are linearly independent over $\mathbb{Q}$. In this case the first error term is simply $O(1)$, as there can be at most one lattice point on every hyperface.

Proof of Theorem 2.5: We start by proving a lemma which will help bound the number of lattice points close to the boundary of the polytope tP.

Lemma 2.6 *Let $2 \leq k \leq d$ be integers, $R > 1$ and $a > 0$. Consider a closed ball B in $\mathbb{R}^d$ of radius R, and two parallel affine hyperplanes at distance a from each other. Suppose the normal vector of the affine hyperplanes has algebraic coordinates which span a vector space of dimension k over $\mathbb{Q}$. Then for every $\varepsilon > 0$ the number of lattice points in B which fall between the two affine hyperplanes is at most $O\left(R^{d-k} + aR^{d-1+\varepsilon}\right)$. The implied constant depends only on the normal vector and ε, and is ineffective.*

Proof of Lemma 2.6: Let $v = (v_1, \ldots, v_d)$ denote the common normal vector of the affine hyperplanes. We may assume that $v_d = 1$. Indeed, v has a nonzero coordinate, so

we can first assume $v_d \neq 0$. Multiplying by a nonzero real number is a linear bijection from $\mathbb{R}$ to $\mathbb{R}$ as a vector space over $\mathbb{Q}$, and linear bijections preserve the dimension of a span. Therefore the numbers $\frac{v_1}{v_d}, \ldots, \frac{v_{d-1}}{v_d}, 1$ are all algebraic and span a vector space of dimension k over $\mathbb{Q}$. From now on we assume $v_d = 1$. Then the equations of the affine hyperplanes are of the form $\left\langle \frac{v}{|v|}, x \right\rangle = b$ and $\left\langle \frac{v}{|v|}, x \right\rangle = b + a$ for some real number b, and the region we are interested in is

$$A = \left\{ x \in B : b \leq \left\langle \frac{v}{|v|}, x \right\rangle \leq b + a \right\}.$$

Extend the linearly independent set $\{1\}$ into a basis of the vector space spanned by $\{v_1, \ldots, v_{d-1}, 1\}$ over $\mathbb{Q}$. Let $\{\alpha_1, \ldots, \alpha_{k-1}, \alpha_k\}$ be the basis obtained, where $\alpha_k = 1$. Since every element of the vector space is algebraic, so are $\alpha_1, \ldots, \alpha_{k-1}$. Therefore we may apply Theorem 2.4 on the numbers $\alpha_1, \ldots, \alpha_{k-1}$. Theorem 2.4 implies that there exists a constant $K > 0$ depending only on $\varepsilon > 0$ and $\alpha_1, \ldots, \alpha_{k-1}$ such that

$$\|m_1\alpha_1 + \cdots + m_{k-1}\alpha_{k-1}\| \geq \frac{K}{|m|^{k-1+\varepsilon}} \tag{2.11}$$

for every $m \in \mathbb{Z}^{k-1} \backslash \{0\}$. Note that the ineffectiveness of Theorem 2.4 means that we cannot find an explicit value for K.

Since $\{\alpha_1, \ldots, \alpha_{k-1}, \alpha_k\}$ is a basis over $\mathbb{Q}$, we can express $v_1, \ldots, v_d$ in the form

$$v_i = \sum_{j=1}^{k} A_{i,j}\alpha_j$$

for some rational coefficients $A_{i,j} \in \mathbb{Q}$. Let $Q > 0$ be an integer for which $QA_{i,j} \in \mathbb{Z}$ for every i, j. Consider the map $g : A \cap \mathbb{Z}^d \to \mathbb{R}$ defined as $g(n) = \left\langle \frac{v}{|v|}, n \right\rangle$. We will first bound the size of the range of g. To this end, let $n, n' \in A \cap \mathbb{Z}^d$ be lattice points such that $g(n) \neq g(n')$. Then we have

$$\left|g(n) - g(n')\right| = \left|\left\langle n - n', \frac{v}{|v|} \right\rangle\right| = \frac{1}{|v|}\left|\sum_{i=1}^{d}(n_i - n_i')v_i\right| =$$

$$\frac{1}{Q|v|}\left|\sum_{j=1}^{k}\sum_{i=1}^{d}(n_i - n_i')QA_{i,j}\alpha_j\right|.$$

Using the facts that $QA_{i,j} \in \mathbb{Z}$ and that $\alpha_k = 1$, we have that the $j = k$ term is an integer. Therefore

$$\left|g(n) - g(n')\right| \geq \frac{1}{Q|v|} \left\| \sum_{j=1}^{k-1} \sum_{i=1}^{d} (n_i - n_i') Q A_{i,j} \alpha_j \right\|.$$

By letting

$$m_j = \sum_{i=1}^{d} (n_i - n_i') Q A_{i,j}$$

we obtain an integral vector $m = (m_1, \ldots, m_{k-1}) \in \mathbb{Z}^{k-1}$. Suppose first that $m \neq 0$. Then (2.11) implies that

$$\left|g(n) - g(n')\right| \geq \frac{1}{Q|v|} \left\| \sum_{j=1}^{k-1} m_j \alpha_j \right\| \geq \frac{K}{Q|v||m|^{k-1+\varepsilon}}.$$

Since m is a linear function of $n - n'$, we have $|m| = O(|n - n'|)$. Since n, n' are both in the ball B of radius R, we have $|n - n'| \leq 2R$. Thus we got

$$\left|g(n) - g(n')\right| = \Omega\left(\frac{1}{R^{k-1+\varepsilon}}\right), \tag{2.12}$$

if the vector $m \neq 0$. If $m = 0$ then $|g(n) - g(n')|$ is the absolute value of an integer. Since we assumed $g(n) \neq g(n')$, and since $R > 1$, we get (2.12) in the case $m = 0$ as well. From the definitions of A and g we can see that the range $g\left(A \cap \mathbb{Z}^d\right)$ is a subset of the interval $[b, b + a]$, which has length a. On the other hand (2.12) shows that the points of the range $g\left(A \cap \mathbb{Z}^d\right)$ have a minimum distance of $\Omega\left(\frac{1}{R^{k-1+\varepsilon}}\right)$. Therefore the size of the range is at most

$$\left|g\left(A \cap \mathbb{Z}^d\right)\right| = O\left(\lceil aR^{k-1+\varepsilon} \rceil\right). \tag{2.13}$$

The geometric meaning of (2.12) is the following. If we draw an affine hyperplane through every lattice point in A parallel to the ones given in the statement of the lemma, then these hyperplanes cannot be too close to each other. To bound the number of lattice points in A, we now have to study how many lattice points there can be on a particular member of this family of parallel hyperplanes.

Consider an affine hyperplane H perpendicular to v, which contains a lattice point $n \in A \cap \mathbb{Z}^d$. Then for any other lattice point $n' \in H \cap \mathbb{Z}^d$ on H we have $\langle n - n', v \rangle = 0$. Thus the set $H \cap \mathbb{Z}^d$ is contained in a rational affine subspace orthogonal to v. Since the coordinates of v span a vector space of dimension k over $\mathbb{Q}$, this rational affine subspace has dimension $d - k$. It is easy to see that the number of lattice points in a given rational affine subspace of dimension $d - k$ which lie in a closed ball of radius $R > 1$ is $O\left(R^{d-k}\right)$. Therefore $\left|H \cap A \cap \mathbb{Z}^d\right| = O\left(R^{d-k}\right)$.

In terms of the function $g : A \cap \mathbb{Z}^d \to \mathbb{R}$ we have thus proved, that its range has size $O\left(\lceil aR^{k-1+\varepsilon} \rceil\right)$, and that every value in its range is attained at most $O\left(R^{d-k}\right)$ times. Therefore the size of its domain satisfies

$$\left|A \cap \mathbb{Z}^d\right| = O\left(\lceil aR^{k-1+\varepsilon} \rceil \cdot R^{d-k}\right).$$

Using $\lceil aR^{k-1+\varepsilon} \rceil \leq aR^{k-1+\varepsilon} + 1$ finishes the proof of Lemma 2.6.

We are now ready to prove the theorem. Let us fix real numbers $\varepsilon > 0$, $t > 1$ and $0 < h < \frac{1}{2}$, and an integer $N > 1$. Let us introduce the function $f : \mathbb{R}^d \to \mathbb{R}$,

$$f(x) = \sum_{n \in \mathbb{Z}^d} \chi_{tP}(n + x) \qquad \left(x \in \mathbb{R}^d\right).$$

Since tP is bounded, the series defining f has finitely many nonzero terms, therefore we do not encounter any convergence issues. Note that f is periodic in each coordinate with period one, and that $f(x)$ is the number of lattice points in the translated polytope $tP - x$. We can apply Proposition 2.2 (ii) on the polytope tP and $L = \left|tP \cap \mathbb{Z}^d\right|$ to obtain

$$C(tP, N) - \left|tP \cap \mathbb{Z}^d\right| = \int_{\left[-\frac{1}{2}, \frac{1}{2}\right]^d} \left(f(x) - \left|tP \cap \mathbb{Z}^d\right|\right) F_N(x) \, \mathrm{d}x,$$

where $C(tP, N)$ is as in Definition 2.3 and $F_N(x)$ is as in Definition 2.5. Let us consider the integral on $[-h, h]^d$ and on $\left[-\frac{1}{2}, \frac{1}{2}\right]^d \setminus [-h, h]^d$ separately, and use the triangle inequality together with the fact that $F_N(x) \geq 0$ from Proposition 2.1 (iv) to obtain

$$\left|C(tP, N) - \left|tP \cap \mathbb{Z}^d\right|\right| \leq \int_{[-h,h]^d} \left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| F_N(x) \, \mathrm{d}x +$$

$$\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d \setminus [-h,h]^d} \left| f(x) - \left| tP \cap \mathbb{Z}^d \right| \right| F_N(x)\,\mathrm{d}x. \tag{2.14}$$

To get an upper bound for the first integral, rewrite the first factor as

$$\left| f(x) - \left| tP \cap \mathbb{Z}^d \right| \right| = \left| \sum_{n \in \mathbb{Z}^d} \left(\chi_{tP}(n+x) - \chi_{tP}(n) \right) \right|.$$

We are only interested in this quantity in the case $x \in [-h,h]^d$, which implies $|x| \le \sqrt{d}h$. Note that if $n \in \mathbb{Z}^d$ is a lattice point such that its distance from the boundary of tP satisfies $\mathrm{dist}(n, \partial(tP)) > \sqrt{d}h$, then $\chi_{tP}(n+x) - \chi_{tP}(n) = 0$. Indeed, since $|x| \le \sqrt{d}h$, the distance of n and $n+x$ is at most $\sqrt{d}h$, therefore either both n and $n+x$, or neither n nor $n+x$ belong to tP. Hence

$$\left| f(x) - \left| tP \cap \mathbb{Z}^d \right| \right| \le \left| \left\{ n \in \mathbb{Z}^d : \mathrm{dist}(n, \partial(tP)) \le \sqrt{d}h \right\} \right| \tag{2.15}$$

for every $x \in [-h,h]^d$. Consider a closed ball B of radius $R(P)$ which covers P, and a hyperface H of P. Let B_H denote the intersection of B and the affine hyperplane containing H. Since the hyperfaces cover the boundary of a polytope, we have $\partial P \subseteq \bigcup_H B_H$, where the union is taken over all hyperfaces H of P. Applying the magnifying factor t we get $\partial(tP) \subseteq \bigcup_H tB_H$, therefore

$$\left\{ y \in \mathbb{R}^d : \mathrm{dist}(y, \partial(tP)) \le \sqrt{d}h \right\} \subseteq \bigcup_H \left\{ y \in \mathbb{R}^d : \mathrm{dist}(y, tB_H) \le \sqrt{d}h \right\}, \tag{2.16}$$

where the union is taken over all hyperfaces H of P. To cover a particular member of this union, first increase the radius of tB by $\sqrt{d}h$, then intersect the obtained ball by two affine hyperplanes parallel to H, both at distance $\sqrt{d}h$ from H. Thus $\left| \left\{ n \in \mathbb{Z}^d : \mathrm{dist}(n, tB_H) \le \sqrt{d}h \right\} \right|$ is at most the number of lattice points in a closed ball of radius $R = tR(P) + \sqrt{d}h = O(t)$ which lie between two parallel affine hyperplanes at distance $a = 2\sqrt{d}h = O(h)$ from each other. Moreover, the normal vector of the affine hyperplanes can be chosen to be the same as the normal vector of the hyperface H. Since the coordinates of the normal vector of H are algebraic and span a vector space of dimension $k_H \ge k$ over $\mathbb{Q}$, Lemma 2.6 implies that

$$\left|\left\{n \in \mathbb{Z}^d : \operatorname{dist}(n, tB_H) \leq \sqrt{d}h\right\}\right| = O\left(R^{d-k_H} + aR^{d-1+\varepsilon}\right) = O(t^{d-k} + ht^{d-1+\varepsilon}).$$

Since there is a constant number of hyperfaces, we obtain

$$\left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| \leq \left|\left\{n \in \mathbb{Z}^d : \operatorname{dist}(n, \partial(tP)) \leq \sqrt{d}h\right\}\right| = O\left(t^{d-k} + ht^{d-1+\varepsilon}\right)$$

uniformly in $x \in [-h, h]^d$. Using Proposition 2.1 (iii)-(iv), the first integral in (2.14) can be bounded by

$$\int_{[-h,h]^d} \left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| F_N(x)\mathrm{d}x \leq$$

$$\sup_{x\in[-h,h]^d} \left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| \cdot \int_{[-\frac{1}{2},\frac{1}{2}]^d} F_N(x)\,\mathrm{d}x = O\left(t^{d-k} + ht^{d-1+\varepsilon}\right). \tag{2.17}$$

To bound the second integral in (2.14) rewrite the first factor as

$$\left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| = \left|\left|(tP - x) \cap \mathbb{Z}^d\right| - \left|tP \cap \mathbb{Z}^d\right|\right|.$$

Since $\lambda(tP - x) = \lambda(tP)$, we can use the triangle inequality to get

$$\left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| \leq \left|\left|(tP - x) \cap \mathbb{Z}^d\right| - \lambda(tP - x)\right| + \left|\left|tP \cap \mathbb{Z}^d\right| - \lambda(tP)\right|.$$

Now we apply the trivial discrepancy bound from Proposition 1.1 on the polytopes $tP - x$ and tP. The upper bound we get is a polynomial in t of degree $d - 1$, therefore we have

$$\left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| = O\left(t^{d-1}\right).$$

Using the fact that $F_N(x) \geq 0$ from Proposition 2.1 (iv) and Proposition 2.3 we can find an upper bound to the second integral in (2.14).

$$\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d \setminus [-h,h]^d} \left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| F_N(x)\,\mathrm{d}x = O\left(t^{d-1}\frac{\log N}{hN}\right). \tag{2.18}$$

Using (2.17) and (2.18) in (2.14) we obtain

$$C(tP,N) - \left|tP \cap \mathbb{Z}^d\right| = O\left(t^{d-k} + ht^{d-1+\varepsilon} + t^{d-1}\frac{\log N}{hN}\right).$$

Since $0 < h < \frac{1}{2}$ was arbitrary, choosing $h = \sqrt{\frac{\log N}{N}}$ to make the second and the third error term have similar orders of magnitude finishes the proof of Theorem 2.5.

□

We noted in Theorem 2.5 that the implied constant in the error term is ineffective. This means that even though we were able to prove the existence of an implied constant which depends only on P and ε, the proof does not provide a way to actually find such a constant. The reason for this is that we used Theorem 2.4 of Schmidt on simultaneous Diophantine approximation, which is ineffective. Our last goal of the chapter is to find an effective version of Theorem 2.5 for a special class of polytopes.

Notice that in the proof of Theorem 2.5 we encountered a simultaneous Diophantine approximation problem with $k-1$ irrational numbers, where k is as in the statement of the theorem. Since the only effective methods of Diophantine approximation are about the approximation of a single irrational number, the best we can hope for is to find an effective version of Theorem 2.5 in the special case, when $k = 2$. In this case every hyperface H of the polytope has a normal vector the coordinates of which are linear combinations of 1 and a single irrational number α_H with rational coefficients. Unfortunately the only class of algebraic numbers for which simple effective methods are known in the theory of Diophantine approximation is the quadratic irrationals. Therefore we shall assume that the normal vectors of our polytope have coordinates in a real quadratic field.

Since we are in the case $k = 2$ of Theorem 2.5, an error of $O\left(t^{d-2}\right)$ will be inevitable. This error is quite large, very close to the trivial discrepancy bound $O\left(t^{d-1}\right)$ of Corollary 1.2. Nevertheless there might still be natural lattice point counting problems where an error of $O\left(t^{d-2}\right)$ is acceptable. For example, we might want to see if a certain polytope satisfies the conjecture of Müller in [17] about the discrepancy of smooth convex bodies being $O\left(t^{d-2+\varepsilon}\right)$ for $d = 3, 4$, and $O\left(t^{d-2}\right)$ for $d \geq 5$.

The following theorem is thus an effective version of Theorem 2.5 for a special class of polytopes.

Theorem 2.7 *Let $d \geq 2$, and let P be a polytope in $\mathbb{R}^d$. Suppose that every hyperface H of P has a normal vector of the form $p_H + \sqrt{D_H} q_H$, where $D_H > 1$ is a square-free integer, and $p_H, q_H \in \mathbb{Z}^d$ are linearly independent over $\mathbb{Q}$. Then for every $t > 1$ and every integer $N > 0$ we have*

$$\left| C(tP, N) - \left| tP \cap \mathbb{Z}^d \right| \right| \leq Ct^{d-2} + Dt^{d-1} \sqrt{\frac{1 + \log N}{N}},$$

with

$$C = 4^{d-2} H(P) R(P)^{d-2},$$

$$D = d 6^{\frac{d}{2}} \sqrt{\omega(d-1)} H(P) R(P)^{d-1} \sqrt{\frac{1}{H(P)} \sum_H \left(|p_H| + \sqrt{D_H}\, |q_H| \right)^2}.$$

Here $C(tP, N)$ is as in Definition 2.3, $\omega(d-1)$ is the Lebesgue measure of the $d-1$ dimensional unit ball, $H(P)$ is the number of hyperfaces of P, $R(P) > \sqrt{d}$ is the radius of a closed ball which covers P, and the summation is over all hyperfaces H of P.

The actual form of the constants C and D is basically irrelevant. The only reason formulas for them are provided is to emphasize that they are effectively computable. Our goal was not to find the best possible constant factors, but rather to find easily computable and simple looking ones. Note that using the well-known explicit formula

$$\omega(d) = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(1 + \frac{d}{2}\right)}, \tag{2.19}$$

we have that the factor $d 6^{\frac{d}{2}} \sqrt{\omega(d-1)}$ has limit zero, as $d \to \infty$. The quantities $H(P)$ and $R(P)$ are simple geometric quantities associated with P, which are invariant under isometries. Note that the assumption $R(P) > \sqrt{d}$ is not very restrictive. The radius of the circumscribed sphere of the unit cube $[0,1]^d$ is as large as $\frac{\sqrt{d}}{2}$, and most polytopes of interest will probably have an $R(P)$ value larger than that. Since the theorem does not require us to choose the smallest possible covering ball, even in the case when P could be covered by a smaller ball, we can choose an $R(P)$ value larger than $\sqrt{d}$.

The quantity

$$\sqrt{\frac{1}{H(P)}\sum_{H}\left(|p_H|+\sqrt{D_H}\,|q_H|\right)^2} \tag{2.20}$$

is more complicated, however. It is the quadratic mean of values associated with the hyperfaces, which intuitively measure how irrational the hyperfaces are. Clearly (2.20) is not invariant under isometries. By rotating P we might even lose the property that the hyperfaces have normal vectors with quadratic irrational coordinates.

Proof of Theorem 2.7: We start by formulating and proving an analogue of Lemma 2.6.

Lemma 2.8 *Let $d \geq 2$ be an integer, $R > 1$ and $a > 0$. Consider a closed ball B in $\mathbb{R}^d$ of radius R, and two parallel affine hyperplanes at distance a from each other. Suppose the normal vector of the affine hyperplanes is of the form $p + \sqrt{D}q$, where $D > 1$ is a square-free integer, and $p, q \in \mathbb{Z}^d$ are linearly independent over $\mathbb{Q}$. Then the number of lattice points in B which fall between the two affine hyperplanes is at most*

$$(2R+1)^{d-2} + \left(|p| + \sqrt{D}|q|\right)^2 a(2R+1)^{d-1}.$$

Proof of Lemma 2.8: The affine hyperplanes have equations of the form

$$\left\langle \frac{p+\sqrt{D}q}{\left|p+\sqrt{D}q\right|}, x \right\rangle = b,$$

$$\left\langle \frac{p+\sqrt{D}q}{\left|p+\sqrt{D}q\right|}, x \right\rangle = b + a$$

for some real number b. The region we are interested in is therefore

$$A = \left\{ x \in B : b \leq \left\langle \frac{p+\sqrt{D}q}{\left|p+\sqrt{D}q\right|}, x \right\rangle \leq b + a \right\}.$$

Consider the map $g : A \cap \mathbb{Z}^d \to \mathbb{R}$ defined as

$$g(n) = \left\langle \frac{p+\sqrt{D}q}{\left|p+\sqrt{D}q\right|}, n \right\rangle.$$

We start by bounding the size of the range of g. To this end, let $n, n' \in A \cap \mathbb{Z}^d$ be two lattice points such that $g(n) \neq g(n')$. Then we have

$$\left|g(n) - g(n')\right| = \frac{1}{\left|p + \sqrt{D}q\right|} \left|\langle p, n - n'\rangle + \sqrt{D}\,\langle q, n - n'\rangle\right|.$$

Here both $\langle p, n - n'\rangle$ and $\langle q, n - n'\rangle$ are integers. Estimating $|g(n) - g(n')|$ is thus equivalent to the classical problem of approximating the quadratic irrational $\sqrt{D}$ by rational numbers. We proceed with the standard trick of multiplying by the conjugate to get

$$\left|g(n) - g(n')\right| = \frac{1}{\left|p + \sqrt{D}q\right|} \cdot \frac{\left|\langle p, n - n'\rangle^2 - D\,\langle q, n - n'\rangle^2\right|}{\left|\langle p, n - n'\rangle - \sqrt{D}\,\langle q, n - n'\rangle\right|}.$$

Since we assumed $g(n) \neq g(n')$, the numerator of the second factor is the absolute value of a nonzero integer, therefore it is at least 1. Using the triangle inequality and the Cauchy–Schwarz inequality we obtain

$$\left|g(n) - g(n')\right| \geq \frac{1}{\left(|p| + \sqrt{D}|q|\right)^2 |n - n'|}.$$

Finally, since n, n' are in a ball of radius R, we have $|n - n'| \leq 2R$, and hence

$$\left|g(n) - g(n')\right| \geq \frac{1}{\left(|p| + \sqrt{D}|q|\right)^2 2R}. \tag{2.21}$$

We can see from the definition of A and g, that the range $g\left(A \cap \mathbb{Z}^d\right)$ is a subset of the interval $[b, b + a]$ of length a. The inequality (2.21) provides a minimum distance between the points of the range. Therefore the size of the range satisfies

$$\left|g\left(A \cap \mathbb{Z}^d\right)\right| \leq \left(|p| + \sqrt{D}|q|\right)^2 2Ra + 1. \tag{2.22}$$

Following the same steps as in the proof of Lemma 2.6, we now have to study how many times g can attain a given value. Let $c \in g\left(A \cap \mathbb{Z}^d\right)$ be an arbitrary element of the range, and consider its set of preimages $g^{-1}(c)$. For any $n, n' \in g^{-1}(c)$ we have $g(n) - g(n') = 0$ and therefore

$$\langle p, n - n' \rangle + \sqrt{D} \langle q, n - n' \rangle = 0. \tag{2.23}$$

Here $\langle p, n - n' \rangle$ and $\langle q, n - n' \rangle$ are integers, and $D > 1$ is square-free. Therefore (2.23) can only be satisfied, if both $\langle p, n - n' \rangle = 0$ and $\langle q, n - n' \rangle = 0$. Thus $n - n'$ is orthogonal to two integral vectors, which are linearly independent over $\mathbb{Q}$. This means, that there exists a rational affine subspace V of dimension $d-2$ such that $g^{-1}(c) \subseteq V$. Let us identify V with $\mathbb{R}^{d-2}$ via an Euclidean isometry. Consider a $d-2$ dimensional open ball B_n within V around each point $n \in g^{-1}(c)$ of radius $\frac{1}{2}$. The set $g^{-1}(c)$ contains only lattice points, therefore their distance from each other is at least 1, making the balls B_n disjoint. The set $V \cap B$ is a $d-2$ dimensional ball of radius at most R. By increasing the radius of $V \cap B$ by $\frac{1}{2}$, we can ensure that it will cover B_n for every $n \in g^{-1}(c)$. Comparing the $d-2$ dimensional Lebesgue measure of

$$\bigcup_{n \in g^{-1}(c)} B_n$$

and $V \cap B$ with its radius increased, we obtain

$$\left|g^{-1}(c)\right| \omega(d-2) \left(\frac{1}{2}\right)^{d-2} \leq \omega(d-2) \left(R + \frac{1}{2}\right)^{d-2},$$

$$\left|g^{-1}(c)\right| \leq (2R+1)^{d-2}. \tag{2.24}$$

The estimates (2.22) and (2.24) together prove that the domain $A \cap \mathbb{Z}^d$ of the map g satisfies

$$\left|A \cap \mathbb{Z}^d\right| \leq \left(\left(|p| + \sqrt{D}|q| \right)^2 2Ra + 1 \right) \cdot (2R+1)^{d-2}.$$

We can simply use $2R < 2R + 1$ to finish the proof of Lemma 2.8.

We are now ready to prove the theorem. Fix real numbers $t > 1$, $0 < h < \frac{1}{2}$, and an integer $N > 0$. Let us introduce the function $f : \mathbb{R}^d \to \mathbb{R}$,

$$f(x) = \sum_{n \in \mathbb{Z}^d} \chi_{tP}(n + x) \qquad \left(x \in \mathbb{R}^d\right).$$

We can apply Proposition 2.2 (ii) on the polytope tP and $L = \left|tP \cap \mathbb{Z}^d\right|$. Similarly to the proof of Theorem 2.5, we have

$$\left|C(tP,N) - \left|tP \cap \mathbb{Z}^d\right|\right| \leq \int_{[-h,h]^d} \left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| F_N(x)\,\mathrm{d}x +$$

$$\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d \setminus [-h,h]^d} \left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| F_N(x)\,\mathrm{d}x, \tag{2.25}$$

where $C(tP,N)$ is as in Definition 2.3 and $F_N(x)$ is as in Definition 2.5.

We first find an upper bound for the first integral in (2.25). Consider a closed ball B in $\mathbb{R}^d$ of radius $R(P) > \sqrt{d}$ which covers P. For every hyperface H of P, let B_H denote the intersection of B and the affine hyperplane containing H. Recall (2.15) and (2.16) from the proof of Theorem 2.5. Thus we get

$$\left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| \leq \sum_H \left|\left\{n \in \mathbb{Z}^d : \mathrm{dist}\,(n, tB_H) \leq \sqrt{d}h\right\}\right|$$

for every $x \in [-h,h]^d$. Also recall, that the region

$$\left\{y \in \mathbb{R}^d : \mathrm{dist}(y, tB_H) \leq \sqrt{d}h\right\}$$

can be covered by the region within a ball of radius $R = tR(P) + \sqrt{d}h$ which lies between two affine hyperplanes parallel to H at distance $a = 2\sqrt{d}h$ from each other. For this R value we have

$$2R + 1 = 2R(P)t + 2\sqrt{d}h + 1 < 4R(P)t,$$

where we used $R(P) > \sqrt{d}$, $t > 1$ and $0 < h < \frac{1}{2}$. Lemma 2.8 thus implies, that

$$\left|\left\{n \in \mathbb{Z}^d : \mathrm{dist}\,(n, tB_H) \leq \sqrt{d}h\right\}\right| \leq$$

$$4^{d-2}R(P)^{d-2}t^{d-2} + \left(|p_H| + \sqrt{D_H}|q_H|\right)^2 2\sqrt{d}h\,4^{d-1}R(P)^{d-1}t^{d-1}.$$

Summing this inequality over every hyperface H of P we get

$$\left|f(x) - \left|tP \cap \mathbb{Z}^d\right|\right| \leq$$

$$4^{d-2}H(P)R(P)^{d-2}t^{d-2}+2\sqrt{d}4^{d-1}R(P)^{d-1}\left(\sum_H\left(|p_H|+\sqrt{D_H}|q_H|\right)^2\right)ht^{d-1}$$

for every $x\in[-h,h]^d$. In order to make our formulas shorter and easier to read, let us define the constants

$$C=4^{d-2}H(P)R(P)^{d-2}$$

as in the statement of the theorem, and

$$A=2\sqrt{d}4^{d-1}R(P)^{d-1}\left(\sum_H\left(|p_H|+\sqrt{D_H}|q_H|\right)^2\right). \tag{2.26}$$

With this notation we have

$$\left|f(x)-\left|tP\cap\mathbb{Z}^d\right|\right|\le Ct^{d-2}+Aht^{d-1}$$

for every $x\in[-h,h]^d$. From this inequality we obtain that the first integral in (2.25) satisfies

$$\int_{[-h,h]^d}\left|f(x)-\left|tP\cap\mathbb{Z}^d\right|\right|F_N(x)\,\mathrm{d}x\le Ct^{d-2}+Aht^{d-1}. \tag{2.27}$$

We can bound the second integral in (2.25) in exactly the same way as in the proof of Theorem 2.5. We have

$$\left|f(x)-\left|tP\cap\mathbb{Z}^d\right|\right|=\left|\left|(tP-x)\cap\mathbb{Z}^d\right|-\left|tP\cap\mathbb{Z}^d\right|\right|\le$$

$$\left|\left|(tP-x)\cap\mathbb{Z}^d\right|-\lambda(tP-x)\right|+\left|\left|tP\cap\mathbb{Z}^d\right|-\lambda(tP)\right|.$$

Applying the trivial discrepancy bound from Proposition 1.1 with $H(tP-x)=H(tP)=H(P)$ and $R(tP-x)=R(tP)=tR(P)$ we get

$$\left|f(x)-\left|tP\cap\mathbb{Z}^d\right|\right|\le 2H(P)\omega(d-1)\sqrt{d}\left(tR(P)+\frac{\sqrt{d}}{2}\right)^{d-1}$$

for every $x\in\left[-\frac{1}{2},\frac{1}{2}\right]^d$. Using the fact that the main contribution of the integral of the Fejér kernel $F_N(x)$ on $\left[-\frac{1}{2},\frac{1}{2}\right]^d$ comes from a small neighborhood of the origin from Proposition 2.3, we can bound the second integral in (2.25) as follows:

$$\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d\setminus[-h,h]^d} \left|f(x)-\left|tP\cap\mathbb{Z}^d\right|\right| F_N(x)\,\mathrm{d}x \leq$$

$$2H(P)\omega(d-1)\sqrt{d}\left(tR(P)+\frac{\sqrt{d}}{2}\right)^{d-1}\cdot\frac{2d}{\pi}\cdot\frac{1+\log N}{hN}.$$

Let us use the estimate $tR(P)+\frac{\sqrt{d}}{2}<\frac{3}{2}tR(P)$, which can easily be seen from $t>1$ and $R(P)>\sqrt{d}$. By introducing the constant

$$B=\frac{4}{\pi}d^{\frac{3}{2}}\omega(d-1)\left(\frac{3}{2}\right)^{d-1}H(P)R(P)^{d-1}, \tag{2.28}$$

the bound simplifies to

$$\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d\setminus[-h,h]^d} \left|f(x)-\left|tP\cap\mathbb{Z}^d\right|\right| F_N(x)\,\mathrm{d}x \leq B\frac{1+\log N}{hN}t^{d-1}. \tag{2.29}$$

Using the estimates (2.27) and (2.29) in (2.25), we get

$$\left|C(tP,N)-\left|tP\cap\mathbb{Z}^d\right|\right| \leq Ct^{d-2}+Aht^{d-1}+B\frac{1+\log N}{hN}t^{d-1}. \tag{2.30}$$

The last step is to choose an optimal value for $0<h<\frac{1}{2}$. To make the second and third error terms equal, we have to choose

$$h=\sqrt{\frac{B}{A}\cdot\frac{1+\log N}{N}}.$$

We have to check, however, that this choice for h is indeed less, than $\frac{1}{2}$. We can prove this fact as follows. First, elementary calculations give, that for every integer $N>0$ we have $\frac{1+\log N}{N}\leq 1$. Thus it is enough to see that $\frac{B}{A}<\frac{1}{4}$. From the definitions of the constants A and B from (2.26) and (2.28) we can simplify their ratio as

$$\frac{B}{A}=\frac{2\omega(d-1)d3^{d-1}}{\pi 8^{d-1}}\cdot\frac{1}{\frac{1}{H(P)}\cdot\sum_H\left(|p_H|+\sqrt{D_H}\,|q_H|\right)^2}.$$

For every hyperface H we have $|p_H|\geq 1$ and $|q_H|\geq 1$, since p_H and q_H are nonzero integral vectors. Therefore

$$\left(|p_H| + \sqrt{D_H}\,|q_H|\right)^2 \geq \left(1+\sqrt{2}\right)^2 > 4.$$

Taking the average of this estimate over all hyperfaces H yields

$$\frac{1}{\frac{1}{H(P)} \cdot \sum_H \left(|p_H| + \sqrt{D_H}\,|q_H|\right)^2} < \frac{1}{4}.$$

Therefore it is enough to prove that

$$a_d = \frac{2\omega(d-1)d3^{d-1}}{\pi 8^{d-1}} < 1$$

for every $d \geq 2$. We can see this fact in two steps. First, direct evaluation gives $a_2 < 1$ and $a_3 < 1$. Second, consider the ratio

$$\frac{a_{d+2}}{a_d} = \frac{\omega(d+1)}{\omega(d-1)} \cdot \frac{d+2}{d} \cdot \left(\frac{3}{8}\right)^2.$$

Using the explicit formula (2.19) and the functional equation of the Gamma function, we can simplify this as

$$\frac{a_{d+2}}{a_d} = \pi \frac{\Gamma\left(1+\frac{d-1}{2}\right)}{\Gamma\left(1+\frac{d+1}{2}\right)} \cdot \frac{d+2}{d} \cdot \left(\frac{3}{8}\right)^2 = \pi \frac{1}{1+\frac{d-1}{2}} \cdot \frac{d+2}{d} \cdot \left(\frac{3}{8}\right)^2 = \frac{18\pi}{64} \cdot \frac{d+2}{d(d+1)}.$$

Since $\frac{18\pi}{64} < 1$ and $\frac{d+2}{d(d+1)} < 1$ for every $d \geq 2$, we obtain $\frac{a_{d+2}}{a_d} < 1$. The facts $a_2 < 1$, $a_3 < 1$ and $a_{d+2} < a_d$ clearly imply that $a_d < 1$ for every $d \geq 2$.

We thus proved that the choice

$$h = \sqrt{\frac{B}{A} \cdot \frac{1+\log N}{N}}$$

indeed satisfies $0 < h < \frac{1}{2}$. This choice in (2.30) gives

$$\left|C(tP,N) - \left|tP \cap \mathbb{Z}^d\right|\right| \leq Ct^{d-2} + \sqrt{AB} \cdot \sqrt{\frac{1+\log N}{N}} t^{d-1}.$$

Finally, use the definitions of A and B from (2.26) and (2.28) to simplify $\sqrt{AB}$ as

$$\sqrt{AB} = \sqrt{\frac{8}{\pi}\omega(d-1)6^{d-1}}\, dH(P)R(P)^{d-1} \sqrt{\frac{1}{H(P)} \sum_H \left(|p_H| + \sqrt{D_H}\,|q_H|\right)^2}.$$

To get the constant D as in the statement of the theorem, simply use the estimate $\sqrt{\frac{8}{\pi}6^{d-1}} < 6^{\frac{d}{2}}$.

□

Chapter 3

Lattice point counting problems in high dimension

3.1 The Fourier transform of the characteristic function of a simplex

Let us return to the main problem we study. Given a polytope P in $\mathbb{R}^d$ and a magnifying factor $t > 1$, we want to estimate the number of lattice points in tP. In chapter 2 we saw that the Cesàro means of the formal series

$$\sum_{m \in \mathbb{Z}^d} \hat{\chi}_{tP}(m)$$

approximate $\left|tP \cap \mathbb{Z}^d\right|$. To actually carry out this approximation, we need to compute the Fourier transform $\hat{\chi}_{tP}$.

As noted before, the Fourier transform is additive in the polytope P. Indeed, if we cut a polytope P with an affine hyperplane into two polytopes P_1 and P_2, then we have $\chi_P = \chi_{P_1} + \chi_{P_2}$ almost everywhere, and therefore $\hat{\chi}_P = \hat{\chi}_{P_1} + \hat{\chi}_{P_2}$. Since every polytope can be decomposed into simplices, it will be enough to compute the Fourier transform of the characteristic function of a simplex. The result is stated in the following theorem.

Theorem 3.1 *Let S be a simplex in $\mathbb{R}^d$ with vertices $v_1, \ldots, v_{d+1}$, and let $t > 0$ and $y \in \mathbb{R}^d$ be arbitrary. Then for any $R > \max_k |\langle y, v_k \rangle|$ we have*

$$\hat{\chi}_{tS}(y) = \frac{(-1)^d d!}{(2\pi i)^{d+1}} \lambda(S) \int_{|z|=R} \frac{e^{-2\pi i z t}}{(z - \langle y, v_1 \rangle) \cdots (z - \langle y, v_{d+1} \rangle)} \, \mathrm{d}z.$$

In the theorem above the integral is a complex line integral. Even though the notation $|z| = R$ is slightly ambiguous, it is often used in the literature. It means that we integrate along the positively oriented circle centered at the origin, with radius R. The condition $R > \max_k |\langle y, v_k \rangle|$ means that every singularity $\langle y, v_k \rangle$ of the integrand is inside the circle.

Note that the magnifying factor t shows up only in the exponential function. This means that the theorem is basically some kind of Fourier expansion of $\hat{\chi}_{tS}(y)$ in the variable t. The "frequencies" in this Fourier expansion are the points of the circle $|z| = R$, while the "coefficients" of the expansion are expressed in terms of the vertices of S, and y.

Even though the Fourier transform of the characteristic function of a simplex is probably well-known, the representation given in Theorem 3.1 seems to be a new result. The main advantage of this representation, as opposed to other representations computed directly from the definition of the Fourier transform, is that it holds for any $y \in \mathbb{R}^d$. Why is that important? Eventually we will want to evaluate $\hat{\chi}_{tS}$ at lattice points $y = m \in \mathbb{Z}^d$ to find the Cesàro means. This means that we have to be able to handle cases when several $\langle y, v_k \rangle$ coincide. It is not completely trivial, but any formula computed directly from the definition of the Fourier transform only holds when the values of $\langle y, v_k \rangle$ are all distinct. For a special case of this phenomenon see the conditions of Lemma 3.2 below. Even if we found a formula from the definition which holds for any y for which $\langle y, v_k \rangle$ are all distinct, it is not at all trivial to take the limit of the formula as y approaches a special value for which several $\langle y, v_k \rangle$ coincide. Again, for a special case we refer to the formula stated in Lemma 3.2. The significance of Theorem 3.1 is that it provides a comprehensive way of handling these special values of y. It should be mentioned that neither in the proof, nor in the application of Theorem 3.1 are deep facts from complex analysis used. Complex analysis, more specifically the residue theorem will only be used as a technical way of carrying out computations.

Proof of Theorem 3.1: We start by computing the Fourier transform directly from the definition in a very special case.

Lemma 3.2 *Let S_0 denote the simplex*

$$S_0 = \left\{ x \in \mathbb{R}^d : x_1, \ldots, x_d \geq 0, x_1 + \cdots + x_d \leq 1 \right\},$$

and let $t > 0$. If $y \in \mathbb{R}^d$ is such that $y_k \neq 0$ for every k, and $y_k \neq y_j$ for every $k \neq j$, then

$$\hat{\chi}_{tS_0}(y) = \frac{(-1)^{d+1}}{(2\pi i)^d} \sum_{k=1}^{d} \frac{1 - e^{-2\pi i y_k t}}{y_k \prod_{j \neq k}(y_k - y_j)}.$$

Proof of Lemma 3.2: We prove the lemma by induction on d. When $d = 1$ the simplex S_0 is simply the interval $[0, 1]$, thus $tS_0 = [0, t]$. The Fourier transform of the characteristic function by definition is

$$\hat{\chi}_{tS_0}(y_1) = \int_0^t e^{-2\pi i x_1 y_1} \, \mathrm{d}x_1 = \frac{e^{-2\pi i y_1 t} - 1}{-2\pi i y_1}$$

for every $y_1 \neq 0$, which matches the general formula for $d = 1$. Note that the empty product is 1 by definition.

Suppose now that the lemma is true for $d-1$, and let us prove it for d. We can use Fubini's theorem to integrate over the set

$$tS_0 = \left\{ x \in \mathbb{R}^d : x_1, \ldots, x_d \geq 0, x_1 + \cdots + x_d \leq t \right\}.$$

The last variable x_d runs in the interval $[0, t]$. For a fixed value of $x_d \in [0, t]$ the cross section of tS_0 is

$$(tS_0)_{x_d} = \left\{ (x_1, \ldots, x_{d-1}) \in \mathbb{R}^{d-1} : x_1, \ldots, x_{d-1} \geq 0, x_1 + \cdots + x_{d-1} \leq t - x_d \right\},$$

which is the $d-1$ dimensional version of S_0 magnified by a factor of $t - x_d$. Since the integrand factors as

$$e^{-2\pi i \langle x, y \rangle} = e^{-2\pi i x_d y_d} e^{-2\pi i (x_1 y_1 + \cdots + x_{d-1} y_{d-1})},$$

we can use the inductive hypothesis to get

$$\hat{\chi}_{tS_0}(y) = \int_0^t e^{-2\pi i x_d y_d} \int_{(tS_0)_{x_d}} e^{-2\pi i (x_1 y_1 + \cdots + x_{d-1} y_{d-1})} \, \mathrm{d}x_1 \ldots \mathrm{d}x_{d-1} \, \mathrm{d}x_d =$$

$$\int_0^t e^{-2\pi i x_d y_d} \frac{(-1)^d}{(2\pi i)^{d-1}} \sum_{k=1}^{d-1} \frac{1 - e^{-2\pi i y_k (t - x_d)}}{y_k \prod_{j \neq k, d}(y_k - y_j)} \, \mathrm{d}x_d =$$

$$\frac{(-1)^d}{(2\pi i)^{d-1}} \sum_{k=1}^{d-1} \frac{1}{y_k \prod_{j \neq k, d}(y_k - y_j)} \cdot \left(\frac{e^{-2\pi i y_d t} - 1}{-2\pi i y_d} - \frac{e^{-2\pi i y_d t} - e^{-2\pi i y_k t}}{-2\pi i (y_d - y_k)} \right) =$$

$$\frac{(-1)^{d+1}}{(2\pi i)^d}\sum_{k=1}^{d-1}\frac{1-e^{-2\pi i y_k t}}{y_k\prod_{j\neq k}(y_k-y_j)}+\frac{(-1)^{d+1}}{(2\pi i)^d}\left(\sum_{k=1}^{d-1}\frac{-1}{y_d\prod_{j\neq k}(y_k-y_j)}\right)\cdot\left(1-e^{-2\pi i y_d t}\right).$$

To finish the proof of the lemma, we need to show

$$\sum_{k=1}^{d-1}\frac{-1}{y_d\prod_{j\neq k}(y_k-y_j)}=\frac{1}{y_d\prod_{j\neq d}(y_d-y_j)}.$$

To see this, consider the partial fraction decomposition

$$\frac{1}{\prod_{j=1}^{d-1}(x-y_j)}=\sum_{k=1}^{d-1}\frac{A_k}{x-y_k}, \tag{3.1}$$

where the constant A_k is

$$A_k=\frac{1}{\prod_{j\neq k,d}(y_k-y_j)}.$$

Substituting $x=y_d$ into (3.1) we get the identity

$$\frac{1}{\prod_{j\neq d}(y_d-y_j)}=\sum_{k=1}^{d-1}\frac{-1}{\prod_{j\neq k}(y_k-y_j)}.$$

Thus the proof of Lemma 3.2 is complete.

Now we prove the theorem. Let S_0 be as in Lemma 3.2, let $t>0$ and $y\in\mathbb{R}^d$ be such that $y_k\neq 0$ for every k, and $y_k\neq y_j$ for every $k\neq j$. We can identify the formula we found in Lemma 3.2 as the sum of residues of a meromorphic function, enabling us to rewrite the formula as the following complex line integral:

$$\frac{(-1)^{d+1}}{(2\pi i)^d}\sum_{k=1}^{d}\frac{1-e^{-2\pi i y_k t}}{y_k\prod_{j\neq k}(y_k-y_j)}=\frac{(-1)^{d+1}}{(2\pi i)^{d+1}}\int_{|z|=R}\frac{1-e^{-2\pi i z t}}{z(z-y_1)\cdots(z-y_d)}\,\mathrm{d}z,$$

where $R>\max_k|y_k|$ so that every singularity of the integrand is inside the circle. Indeed, first note that the conditions on y imply that the integrand has $d+1$ distinct isolated singularities. The singularity at $z=0$ is removable, since the numerator has a zero at that point. The singularity at $z=y_k$ is a simple pole, the residue of which is exactly the kth term of the left hand side.

We claim that

$$\hat{\chi}_{tS_0}(y) = \frac{(-1)^{d+1}}{(2\pi i)^{d+1}} \int_{|z|=R} \frac{1 - e^{-2\pi i z t}}{z(z-y_1)\cdots(z-y_d)}\,\mathrm{d}z \tag{3.2}$$

holds for any $t > 0$ and any $y \in \mathbb{R}^d$, as long as $R > \max_k |y_k|$. To see this, fix the value of t, and fix an arbitrary positive constant r. It is enough to prove (3.2) for $|y| < r$, because r was arbitrary. The left hand side by definition is the parametric integral of a bounded function over the bounded set tS_0, therefore Lebesgue's dominated convergence theorem implies that it is a continuous function of y. By fixing $R > r$ it is easy to see that the right hand side of (3.2) is also a continuous function of y on the open ball $|y| < r$. Lemma 3.2 implies that these two continuous functions are equal on a dense subset of the open ball $|y| < r$, therefore they are equal everywhere on $|y| < r$.

Note that the contribution of 1 in the numerator of (3.2) is zero, i.e.

$$\int_{|z|=R} \frac{1}{z(z-y_1)\cdots(z-y_d)}\,\mathrm{d}z = 0$$

provided that $R > \max_k |y_k|$. Indeed, the residue theorem implies that the value of this integral does not depend on R, as long as R is large enough so that all the singularities are within the circle. On the other hand, the trivial estimate implies that the integral is $O\left(\frac{1}{R^d}\right)$, making its limit zero, as $R \to \infty$. We have thus proved, that

$$\hat{\chi}_{tS_0}(y) = \frac{(-1)^{d}}{(2\pi i)^{d+1}} \int_{|z|=R} \frac{e^{-2\pi i z t}}{z(z-y_1)\cdots(z-y_d)}\,\mathrm{d}z \tag{3.3}$$

holds for all $t > 0$ and all $y \in \mathbb{R}^d$, if $R > \max_k |y_k|$.

Now we generalize (3.3) to an arbitrary simplex. Let S be an arbitrary simplex in $\mathbb{R}^d$ with vertices $v_1, \ldots, v_{d+1} \in \mathbb{R}^d$, as in the theorem. It is easy to see that S is the image of S_0 under an affine transformation of $\mathbb{R}^d$. Indeed, let M be the $n \times n$ matrix the columns of which are the vectors $v_1 - v_{d+1}, v_2 - v_{d+1}, \ldots, v_d - v_{d+1}$, and let the affine transformation $g : \mathbb{R}^d \to \mathbb{R}^d$ be defined as $g(x) = Mx + v_{d+1}$, where Mx denotes the product of the matrix M and the column vector x as in linear algebra. Since the vertices of S_0 are the zero vector and the standard basis vectors in $\mathbb{R}^d$, it is easy to check that g maps the vertices of S_0 to the vertices of S. Affine transformations map convex sets to convex sets, therefore we have $g(S_0) = S$. Clearly $g'(x) = M$, so applying the integral transformation formula with the transformation g we get

$$\hat{\chi}_S(y) = \int_S e^{-2\pi i\langle x,y\rangle}\,\mathrm{d}x = \int_{S_0} e^{-2\pi i\langle Mx+v_{d+1},y\rangle}\,|\det M|\,\mathrm{d}x =$$

$$e^{-2\pi i\langle v_{d+1},y\rangle}\,|\det M| \int_{S_0} e^{-2\pi i\langle x,M^T y\rangle}\,\mathrm{d}x = e^{-2\pi i\langle v_{d+1},y\rangle}\,|\det M|\,\hat{\chi}_{S_0}(M^T y), \tag{3.4}$$

where M^T denotes the transpose of the matrix M. To introduce the magnifying factor t, we can use the integral transformation formula again with the simpler transformation $x \mapsto tx$. The Jacobian of this transformation is t^d which gives us

$$\hat{\chi}_{tS}(y) = \int_{tS} e^{-2\pi i\langle x,y\rangle}\,\mathrm{d}x = \int_S e^{-2\pi i\langle tx,y\rangle} t^d\,\mathrm{d}x = t^d\hat{\chi}_S(ty). \tag{3.5}$$

Using (3.4) and (3.5) we can express the quantity we are interested in as

$$\hat{\chi}_{tS}(y) = e^{-2\pi i\langle v_{d+1},ty\rangle}\,|\det M|\,t^d\hat{\chi}_{S_0}(tM^T y).$$

By applying (3.5) on S_0 instead of S and $M^T y$ instead of y, we have $t^d\hat{\chi}_{S_0}(tM^T y) = \hat{\chi}_{tS_0}(M^T y)$, thus

$$\hat{\chi}_{tS}(y) = e^{-2\pi i\langle v_{d+1},ty\rangle}\,|\det M|\,\hat{\chi}_{tS_0}(M^T y).$$

The factor $|\det M|$ has a clear geometric meaning. To express it in terms of S, substitute $y = 0$ in (3.4) to get $\lambda(S) = |\det M|\,\lambda(S_0)$. It is well-known that $\lambda(S_0) = \frac{1}{d!}$, thus $|\det M| = d!\lambda(S)$, which yields

$$\hat{\chi}_{tS}(y) = e^{-2\pi i\langle v_{d+1},ty\rangle} d!\lambda(S)\hat{\chi}_{tS_0}(M^T y).$$

We now want to use (3.3) to replace $\hat{\chi}_{tS_0}(M^T y)$. To do so, we need to find the coordinates of its argument $M^T y$. Recall that the columns of M are the vectors $v_1 - v_{d+1}, \ldots, v_d - v_{d+1}$. Therefore the coordinates of $M^T y$ are

$$\langle v_1 - v_{d+1}, y\rangle, \langle v_2 - v_{d+1}, y\rangle, \ldots, \langle v_d - v_{d+1}, y\rangle.$$

Thus we obtain

$$\hat{\chi}_{tS}(y) = \frac{(-1)^d d!}{(2\pi i)^{d+1}} \lambda(S) \int_{|z|=R} \frac{e^{-2\pi i(z+\langle v_{d+1}, y\rangle)t}}{z(z - \langle v_1 - v_{d+1}, y\rangle) \cdots (z - \langle v_d - v_{d+1}, y\rangle)} \, \mathrm{d}z,$$

where $R > \max_k |\langle v_k - v_{d+1}, y\rangle|$. To obtain the final form, let us use the translation $h(z) = z - \langle v_{d+1}, y\rangle$ as an integral transformation. Then $h'(z) = 1$, hence we get

$$\hat{\chi}_{tS}(y) = \frac{(-1)^d d!}{(2\pi i)^{d+1}} \lambda(S) \int_{\gamma} \frac{e^{-2\pi i z t}}{(z - \langle y, v_1\rangle) \cdots (z - \langle y, v_{d+1}\rangle)} \, \mathrm{d}z,$$

where the path γ is the circle $|z| = R$ translated by $\langle v_{d+1}, y\rangle$. It is easy to see that every singularity $\langle v_k, y\rangle$ is inside γ. The residue theorem implies that we can replace γ by a circle centered at the origin, with a radius large enough so that every singularity is inside it.

□

3.2 The polyhedral sphere problem

3.2.1 The main term

Consider the polytope

$$P = \left\{ x \in \mathbb{R}^d : \frac{|x_1|}{a_1} + \cdots + \frac{|x_d|}{a_d} \leq 1 \right\},$$

where $a_1, \ldots, a_d > 0$ are algebraic numbers. Notice that the map

$$x \mapsto \frac{|x_1|}{a_1} + \cdots + \frac{|x_d|}{a_d}$$

is a norm in $\mathbb{R}^d$. Thus P is the unit ball, while its boundary, ∂P, is the unit sphere with respect to this norm. Since P is also a polyhedron, we call the problem of estimating the number of lattice points in tP, as a function of the magnifying factor t, the polyhedral sphere problem. Section 3.2 is devoted to studying this problem. We start by identifying the main term of $\left|tP \cap \mathbb{Z}^d\right|$.

Theorem 2.5 applies directly to our polytope, thus we can use the Cesàro means $C(tP, N)$ defined in Definition 2.3 to estimate $\left|tP \cap \mathbb{Z}^d\right|$. To actually carry out this

approximation, we need to find the Fourier transform $\hat{\chi}_{tP}$, and then the partial sums $S(tP, M)$ defined in Definition 2.2. Since P can easily be cut into simplices using affine hyperplanes, we can use Theorem 3.1 to compute $S(tP, M)$ as follows.

Proposition 3.2 *Let $a_1, \dots, a_d > 0$, and consider the polytope*

$$P = \left\{ x \in \mathbb{R}^d : \frac{|x_1|}{a_1} + \cdots + \frac{|x_d|}{a_d} \leq 1 \right\}.$$

Let $M_1, \dots, M_d \geq 0$ be integers and $M = (M_1, \dots, M_d)$. For any $t > 0$ we have

$$S(tP, M) = \frac{(-1)^d 2^d a_1 \cdots a_d}{(2\pi i)^{d+1}} \sum_{m_1=-M_1}^{M_1} \cdots \sum_{m_d=-M_d}^{M_d} \int_{|z|=R} \frac{e^{-2\pi i z t}}{z(z - m_1 a_1) \cdots (z - m_d a_d)} \, \mathrm{d}z,$$

where $R > \max_k M_k a_k$ and $S(tP, M)$ is as in Definition 2.2.

Proof: Let us cut P using the hyperplanes $x_1 = 0, \dots, x_d = 0$. We obtain the 2^d simplices

$$S_\sigma = \left\{ x \in \mathbb{R}^d : \sigma_1 x_1 \geq 0, \dots, \sigma_d x_d \geq 0, \frac{\sigma_1 x_1}{a_1} + \cdots + \frac{\sigma_d x_d}{a_d} \leq 1 \right\},$$

where $\sigma \in \{1, -1\}^d$. Then tP is also decomposed into tS_σ for $\sigma \in \{1, -1\}^d$, therefore

$$\hat{\chi}_{tP} = \sum_{\sigma \in \{1,-1\}^d} \hat{\chi}_{tS_\sigma}. \tag{3.6}$$

Let us consider the particular simplex $S = S_{(1,1,\dots,1)}$. Using an integral transformation in the definition of the Fourier transform, it is easy to see that $\hat{\chi}_{tS_\sigma}(y) = \hat{\chi}_{tS}(y_\sigma)$, where $y_\sigma = (\sigma_1 y_1, \dots, \sigma_d y_d)$. By evaluating (3.6) at lattice points $(m_1, \dots, m_d)$ and summing it over the integral points of the rectangle $[-M_1, M_1] \times \cdots \times [-M_d, M_d]$ we get

$$S(tP, M) = \sum_{m_1=-M_1}^{M_1} \cdots \sum_{m_d=-M_d}^{M_d} \sum_{\sigma \in \{1,-1\}^d} \hat{\chi}_{tS_\sigma}(m_1, \dots, m_d) =$$

$$\sum_{\sigma \in \{1,-1\}^d} \sum_{m_1=-M_1}^{M_1} \cdots \sum_{m_d=-M_d}^{M_d} \hat{\chi}_{tS}(\sigma_1 m_1, \dots, \sigma_d m_d).$$

Here every $\sigma \in \{1, -1\}^d$ yields the same inner sum, as the effect of σ is simply a reordering of the terms of the inner sum. Therefore we get $S(tP, M) = 2^d S(tS, M)$.

Now we can use Theorem 3.1 to express $2^d S(tS, M)$ as a complex line integral. The Lebesgue measure of S is $\lambda(S) = \frac{a_1 \cdots a_d}{d!}$, while the vertices of S are the zero vector, and a_k times the kth standard basis vector in $\mathbb{R}^d$ for $k = 1, \ldots, d$. Substituting these values into the formula given in Theorem 3.1 finishes the proof.

□

The formula we found for $S(tP, M)$ in Proposition 3.2 involves a complex line integral, which can be evaluated using the residue theorem. We will consider the contribution of the residue at zero, and the contribution of all the other residues separately. Note that in case of the terms indexed by lattice points $(m_1, \ldots, m_d)$ which have one or more zero coordinates, the pole of the integrand at zero has an order higher than 1. Nevertheless it is possible to compute the sum of the residues at zero over all lattice points $(m_1, \ldots, m_d)$ in the rectangle $[-M_1, M_1] \times \ldots \times [-M_d, M_d]$ up to an effective error term. This sum will serve as the main term of $S(tP, M)$. The contribution of all the other residues will be considered a randomly fluctuating term. The rest of subsection 3.2.1 is devoted to computing and analyzing the main term.

Proposition 3.3 *Let $a_1, \ldots, a_d > 0$ be reals, and $M_1, \ldots, M_d \geq 0$ be integers. For any $t > 1$ we have*

$$\frac{(-1)^d 2^d a_1 \cdots a_d}{(2\pi i)^d} \sum_{m_1=-M_1}^{M_1} \cdots \sum_{m_d=-M_d}^{M_d} \mathrm{Res}_0 \frac{e^{-2\pi i z t}}{z(z - m_1 a_1) \cdots (z - m_d a_d)} =$$

$$\frac{2^d a_1 \cdots a_d}{d!} t^d + \sum_{k=0}^{d-2} \frac{2^d a_1 \cdots a_d}{(2\pi i)^{d-k} k!} t^k \sum_{\ell=1}^{d} \sum_{1 \leq j_1 < \ldots < j_\ell \leq d} \sum_{\substack{i_1 + \cdots + i_\ell = d-k \\ i_1, \ldots, i_\ell \geq 2 \\ 2 | i_1, \ldots, i_\ell}} \frac{-2\zeta(i_1)}{a_{j_1}^{i_1}} \cdots \frac{-2\zeta(i_\ell)}{a_{j_\ell}^{i_\ell}} +$$

$$O\left(\frac{t^{d-2}}{M_1 + 1} + \cdots + \frac{t^{d-2}}{M_d + 1} \right),$$

where ζ is the Riemann zeta function. The implied constant depends only on $a_1, \ldots, a_d$, and is effective.

Proof: Let L denote the left hand side of the formula we are trying to prove. Let us first find the term $m_1 = \ldots = m_d = 0$. Then the function the residue of which we are

interested in is simply $\frac{e^{-2\pi izt}}{z^{d+1}}$. Using the Taylor expansion of the exponential function around zero, we get the Laurent series

$$\frac{e^{-2\pi izt}}{z^{d+1}} = \frac{1 + (-2\pi it)z + \cdots + \frac{(-2\pi it)^d}{d!}z^d + \cdots}{z^{d+1}} =$$

$$\frac{1}{z^{d+1}} + \frac{-2\pi it}{z^d} + \cdots + \frac{\frac{(-2\pi it)^d}{d!}}{z} + \cdots.$$

Therefore the $m_1 = \ldots = m_d = 0$ term of L is

$$\frac{(-1)^d 2^d a_1 \cdots a_d}{(2\pi i)^d} \cdot \mathrm{Res}_0 \frac{e^{-2\pi izt}}{z^{d+1}} = \frac{2^d a_1 \cdots a_d}{d!} t^d.$$

Let us now consider a term indexed by a lattice point $(m_1, \ldots, m_d)$ such that not all of its coordinates are zero. Let the number of nonzero coordinates be $1 \leq \ell \leq d$. Suppose for the sake of simplicity, that $m_1, \ldots, m_\ell \neq 0$ and $m_{\ell+1} = \ldots = m_d = 0$. Then the function the residue of which we are interested in is

$$\frac{e^{-2\pi izt}}{z^{d-\ell+1}(z - m_1 a_1) \cdots (z - m_\ell a_\ell)} = \frac{1}{z^{d+1}} \cdot e^{-2\pi izt} \cdot \frac{z}{z - m_1 a_1} \cdots \frac{z}{z - m_\ell a_\ell}.$$

Consider the Taylor expansions

$$e^{-2\pi izt} = \sum_{k=0}^{\infty} \frac{(-2\pi it)^k}{k!} z^k,$$

$$\frac{z}{z - m_1 a_1} = \sum_{i_1=1}^{\infty} \frac{-1}{(m_1 a_1)^{i_1}} z^{i_1}, \ldots, \frac{z}{z - m_\ell a_\ell} = \sum_{i_\ell=1}^{\infty} \frac{-1}{(m_\ell a_\ell)^{i_\ell}} z^{i_\ell},$$

which hold in an open neighborhood of $z = 0$. These expansions imply that

$$\mathrm{Res}_0 \frac{e^{-2\pi izt}}{z^{d-\ell+1}(z - m_1 a_1) \cdots (z - m_\ell a_\ell)}$$

equals the coefficient of z^d in the power series

$$\left(\sum_{k=0}^{\infty} \frac{(-2\pi it)^k}{k!} z^k \right) \left(\sum_{i_1=1}^{\infty} \frac{-1}{(m_1 a_1)^{i_1}} z^{i_1} \right) \cdots \left(\sum_{i_\ell=1}^{\infty} \frac{-1}{(m_\ell a_\ell)^{i_\ell}} z^{i_\ell} \right).$$

The largest k value which contributes to the coefficient of z^d is $k = d - \ell \leq d - 1$. Therefore to find this coefficient, we can first fix a value $0 \leq k \leq d - 1$, then consider all positive integers $i_1, \ldots, i_\ell$ for which $i_1 + \cdots + i_\ell = d - k$. This yields

$$\mathrm{Res}_0 \frac{e^{-2\pi izt}}{z^{d-\ell+1}(z-m_1a_1)\cdots(z-m_\ell a_\ell)} =$$

$$\sum_{k=0}^{d-1} \frac{(-2\pi i)^k}{k!} t^k \sum_{\substack{i_1+\cdots+i_\ell=d-k \\ i_1,\ldots,i_\ell\geq 1}} \frac{-1}{(m_1a_1)^{i_1}} \cdots \frac{-1}{(m_\ell a_\ell)^{i_\ell}}.$$

Let us add up this equation for $m_1 \in [-M_1, M_1]\backslash\{0\}, \ldots, m_\ell \in [-M_\ell, M_\ell]\backslash\{0\}$. Notice, that any term for which at least one out of $i_1, \ldots, i_\ell$ is odd will cancel. This also implies that the term $k = d-1$ cancels. Thus we obtain

$$\sum_{\substack{m_1=-M_1 \\ m_1\neq 0}}^{M_1} \cdots \sum_{\substack{m_\ell=-M_\ell \\ m_\ell\neq 0}}^{M_\ell} \mathrm{Res}_0 \frac{e^{-2\pi izt}}{z^{d-\ell+1}(z-m_1a_1)\cdots(z-m_\ell a_\ell)} =$$

$$\sum_{k=0}^{d-2} \frac{(-2\pi i)^k}{k!} t^k \sum_{\substack{i_1+\cdots+i_\ell=d-k \\ i_1,\ldots,i_\ell\geq 2 \\ 2|i_1,\ldots,i_\ell}} \sum_{\substack{m_1=-M_1 \\ m_1\neq 0}}^{M_1} \cdots \sum_{\substack{m_\ell=-M_\ell \\ m_\ell\neq 0}}^{M_\ell} \frac{-1}{(m_1a_1)^{i_1}} \cdots \frac{-1}{(m_\ell a_\ell)^{i_\ell}}.$$

Let us use the general formula

$$\sum_{\substack{m=-M \\ m\neq 0}}^{M} \frac{1}{m^i} = 2\zeta(i) + O\left(\frac{1}{(M+1)^{i-1}}\right),$$

which holds for any positive even integer i, to compute the inner sums to get

$$\sum_{\substack{m_1=-M_1 \\ m_1\neq 0}}^{M_1} \cdots \sum_{\substack{m_\ell=-M_\ell \\ m_\ell\neq 0}}^{M_\ell} \mathrm{Res}_0 \frac{e^{-2\pi izt}}{z^{d-\ell+1}(z-m_1a_1)\cdots(z-m_\ell a_\ell)} =$$

$$\sum_{k=0}^{d-2} \frac{(-2\pi i)^k}{k!} t^k \sum_{\substack{i_1+\cdots+i_\ell=d-k \\ i_1,\ldots,i_\ell\geq 2 \\ 2|i_1,\ldots,i_\ell}} \frac{-2\zeta(i_1)}{a_1^{i_1}} \cdots \frac{-2\zeta(i_\ell)}{a_\ell^{i_\ell}} + O\left(\frac{t^{d-2}}{M_1+1} + \cdots + \frac{t^{d-2}}{M_\ell+1}\right).$$

Up to the factor $\frac{(-1)^d 2^d a_1\cdots a_d}{(2\pi i)^d}$, this is the contribution of the terms $m_1, \ldots, m_\ell \neq 0$, $m_{\ell+1} = \ldots = m_d = 0$ in the sum defining L. To find the contribution of all lattice points $(m_1, \ldots, m_d)$ with exactly ℓ nonzero coordinates, simply replace $a_1, \ldots, a_\ell$ by $a_{j_1}, \ldots, a_{j_\ell}$, and sum over $1 \leq j_1 < \ldots < j_\ell \leq d$. Finally, to find the contribution of all lattice points with at least one nonzero coordinate, sum over $1 \leq \ell \leq d$.

□

The formula found in Proposition 3.3 will serve as the main term of $\left|tP \cap \mathbb{Z}^d\right|$ in the polyhedral sphere problem. Since it is quite long and complicated, we introduce a notation for it as follows.

Definition 3.1 *Let $a_1, \ldots, a_d > 0$ be real numbers, and let ζ denote the Riemann zeta function. The function $p = p_{(a_1,\ldots,a_d)}$ of the real variable t is defined as*

$$p(t) = p_{(a_1,\ldots,a_d)}(t) =$$

$$\frac{2^d a_1 \cdots a_d}{d!} t^d + \sum_{k=0}^{d-2} \frac{2^d a_1 \cdots a_d}{(2\pi i)^{d-k} k!} t^k \sum_{\ell=1}^{d} \sum_{1 \le j_1 < \ldots < j_\ell \le d} \sum_{\substack{i_1+\cdots+i_\ell = d-k \\ i_1,\ldots,i_\ell \ge 2 \\ 2 \mid i_1,\ldots,i_\ell}} \frac{-2\zeta(i_1)}{a_{j_1}^{i_1}} \cdots \frac{-2\zeta(i_\ell)}{a_{j_\ell}^{i_\ell}}.$$

The main properties of $p(t)$ are the following.

Proposition 3.4 *Let $a_1, \ldots, a_d > 0$ be real numbers, and let the function $p(t)$ be as in Definition 3.1.*

(i) *$p(t)$ is a polynomial of degree d.*

(ii) *In every term of $p(t)$ the exponent of t is congruent to d modulo 2.*

(iii) *The coefficients of $p(t)$ are symmetric rational functions of $a_1, \ldots, a_d$ with rational coefficients.*

(iv) *The coefficient of t^{d-2} in $p(t)$ is*

$$\frac{2^{d-2} a_1 \cdots a_d}{3(d-2)!} \sum_{1 \le i \le d} \frac{1}{a_i^2}.$$

(v) *If $d \ge 4$, the coefficient of t^{d-4} in $p(t)$ is*

$$\frac{2^{d-4} a_1 \cdots a_d}{9(d-4)!} \left(\sum_{1 \le i < j \le d} \frac{1}{a_i^2 a_j^2} - \frac{1}{5} \sum_{1 \le i \le d} \frac{1}{a_i^4} \right).$$

Proof:

(i): Trivial from the definition of $p(t)$.

(ii): If k is incongruent to d modulo 2, then it is impossible to write $d - k$ as a sum of positive even integers. This results in an empty sum in the coefficient of t^k in the definition of $p(t)$.

(iii): The coefficients of $p(t)$ are clearly symmetric rational functions of $a_1, \ldots, a_d$. In any term indexed by $0 \leq k \leq d-2$ which is congruent to d modulo 2, we raise i to an even power. Also note that for any positive even integers $i_1, \cdots, i_\ell$ such that $i_1 + \cdots + i_\ell = d - k$, we have

$$\zeta(i_1) \cdots \zeta(i_\ell) \in \pi^{i_1} \cdots \pi^{i_\ell} \mathbb{Q} = \pi^{d-k} \mathbb{Q}$$

resulting in rational coefficients.

(iv): To find the coefficient of t^{d-2} consider the $k = d-2$ term in the sum defining $p(t)$. Since the only way to write $d - k = 2$ as the sum of positive even integers is $2 = 2$, in the inner sum we have $\ell = 1$, $i_1 = 2$, and $1 \leq j_1 \leq d$. Thus the coefficient is

$$\frac{2^d a_1 \cdots a_d}{(2\pi i)^2 (d-2)!} \sum_{1 \leq j_1 \leq d} \frac{-2\zeta(2)}{a_{j_1}^2}.$$

Substituting $\zeta(2) = \frac{\pi^2}{6}$ finishes the proof.

(v): To find the coefficient of t^{d-4} consider the $k = d-4$ term in the sum defining $p(t)$. The only two ways of writing $d - k = 4$ as the sum of positive even integers are $4 = 4$ and $4 = 2 + 2$. Therefore in the $\ell = 1$ term of the inner sum we have $i_1 = 4$, while in the $\ell = 2$ term we have $i_1 = i_2 = 2$. The coefficient is altogether

$$\frac{2^d a_1 \cdots a_d}{(2\pi i)^4 (d-4)!} \left(\sum_{1 \leq j_1 \leq d} \frac{-2\zeta(4)}{a_{j_1}^4} + \sum_{1 \leq j_1 < j_2 \leq d} \frac{-2\zeta(2)}{a_{j_1}^2} \cdot \frac{-2\zeta(2)}{a_{j_2}^2} \right).$$

Substituting $\zeta(2) = \frac{\pi^2}{6}$ and $\zeta(4) = \frac{\pi^4}{90}$ finishes the proof.

□

We conclude the analysis of the main term $p(t)$ with two remarks. First note that the highest degree term of $p(t)$ is of course equal to the Lebesgue measure of tP, where P is the polytope in the polyhedral sphere problem. The other terms of $p(t)$, however, do not seem to have a natural geometric interpretation.

The relationship between the terms of the formal series

$$\sum_{m \in \mathbb{Z}^d} \hat{\chi}_{tP}(m)$$

and the terms of the polynomial $p(t)$ might also be worth mentioning. Clearly $\hat{\chi}_{tP}(0)$ is precisely the highest degree term of $p(t)$. A careful analysis of the proof of Proposition 3.3 shows that $\hat{\chi}_{tP}(m)$, where $m \in \mathbb{Z}^d$ has ℓ nonzero coordinates, contributes to the coefficient of t^k only for $k \leq d - 2\ell$. This is simply because in the definition of $p(t)$ the smallest possible values $i_1 = \ldots = i_\ell = 2$ give $i_1 + \cdots + i_\ell = 2\ell = d - k$. Note that $\hat{\chi}_{tP}(m)$, where $m \neq 0$, also contributes to the randomly fluctuating term as well, since in this case the function

$$\frac{e^{-2\pi izt}}{z(z - m_1a_1)\cdots(z - m_da_d)}$$

has a singularity other than $z = 0$.

3.2.2 The expected value of the fluctuating term

Let $a_1, \ldots, a_d > 0$ and consider the polytope P as in the polyhedral sphere problem. In subsection 3.2.1 we identified the main term $p(t)$, as in Definition 3.1. The error $\left|tP \cap \mathbb{Z}^d\right| - p(t)$ will be considered a random fluctuation. This subsection is devoted to studying the expected value of this fluctuation. In other words, we will be interested in

$$\frac{1}{T_2 - T_1} \int_{T_1}^{T_2} \left(\left|tP \cap \mathbb{Z}^d\right| - p(t)\right) \, \mathrm{d}t,$$

where $1 \leq T_1 < T_2$ are fixed constants. We will only consider the case, when $a_1, \ldots, a_d$ are algebraic numbers. The most general result is stated in the following theorem.

Theorem 3.5 *Let $2 \leq k \leq d$ be integers, and let $a_1, \ldots, a_d > 0$ be algebraic. Suppose that any k numbers out of $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$ are linearly independent over $\mathbb{Q}$. Consider the polytope*

$$P = \left\{ x \in \mathbb{R}^d : \frac{|x_1|}{a_1} + \cdots + \frac{|x_d|}{a_d} \leq 1 \right\},$$

and the polynomial $p(t)$ as in Definition 3.1. Let $1 \leq T_1 < T_2$ and $\varepsilon > 0$.

(i) *If $k = d$, then*

$$\frac{1}{T_2 - T_1} \int_{T_1}^{T_2} \left(\left|tP \cap \mathbb{Z}^d\right| - p(t)\right) \, \mathrm{d}t = O\left(1 + \left(\frac{1}{T_2 - T_1}\right)^{\frac{d-2}{d-1}+\varepsilon}\right).$$

(ii) *If* $2 \le k \le d-1$, *then*

$$\frac{1}{T_2 - T_1} \int_{T_1}^{T_2} \left(\left| tP \cap \mathbb{Z}^d \right| - p(t) \right) \, \mathrm{d}t =$$

$$O\left(T_2^{d-k} + T_2^{\frac{2(d-1)(d-k-1)}{2d-k-3}+\varepsilon} \left(\frac{1}{T_2 - T_1} \right)^{\frac{k-1}{2d-k-3}} + \left(\frac{1}{T_2 - T_1} \right)^{\frac{d-2}{k-1}+\varepsilon} \right).$$

The implied constants in (i) and (ii) depend only on $a_1, \ldots, a_d$ *and* ε, *and are ineffective.*

In the conditions of the theorem the numbers $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$ have a clear geometric meaning: they are the coordinates of the normal vector of a hyperface of P. The significance of their linear independence over $\mathbb{Q}$ has already been seen in Theorem 2.5. Now we need a stronger assumption, however. The value of k still intuitively measures how irrational the polytope P is. The case $k = 2$ simply means that the ratio $\frac{a_i}{a_j}$ is irrational for any $i \neq j$. The strongest case $k = d$ means that $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$ are linearly independent over $\mathbb{Q}$.

Note that in the special case $k = d$ and $T_2 - T_1 \ge 1$, the error term becomes simply $O(1)$. The most important message of this result is that we correctly identified the polynomial $p(t)$ as the main term. In other words every coefficient of $p(t)$, except for the constant term, has an actual meaning related to the polyhedral sphere problem. This fact was not at all obvious from the way we defined $p(t)$.

In the case $2 \le k \le d-1$ we have three different error terms. The dominating term depends on the length and the location of the interval $[T_1, T_2]$ over which we take the average. Note that the theorem may be applied to very short intervals. The reason why the result is stated and proved in such generality is that, surprisingly, averaging over very short intervals will play a crucial role in finding a uniform bound for the random fluctuation $\left| tP \cap \mathbb{Z}^d \right| - p(t)$ in the following subsection.

In the proof of Theorem 3.5 we are going to encounter a simultaneous Diophantine approximation problem. To solve it, we will need to use Theorem 2.3 of Schmidt. Note that in the proof of Theorem 2.5 we have already used the dual version of Schimdt's theorem, Theorem 2.4. Even though the simultaneous Diophantine approximation problem

we need to solve is somewhat technical, it might be of interest on its own. Therefore we state it as a separate proposition as follows.

Proposition 3.6 *Let $1 \le k \le d$ be integers, and let $\alpha_1, \dots, \alpha_d$ be real algebraic numbers. Suppose that for any $1 \le i_1 < \dots < i_k \le d$ the numbers $1, \alpha_{i_1}, \dots, \alpha_{i_k}$ are linearly independent over $\mathbb{Q}$. Then for any $M > 0$ and $\varepsilon > 0$ we have*

$$\sum_{m=1}^{M} \frac{1}{\|m\alpha_1\| \cdots \|m\alpha_d\|} = O\left(M^{\frac{d+k-1}{k}+\varepsilon}\right).$$

The implied constant depends only on $\alpha_1, \dots, \alpha_d$ and ε, and is ineffective.

Let us analyze this proposition in the special case $k = d$. In this case Schmidt's theorem implies that the terms satisfy

$$\frac{1}{\|m\alpha_1\| \cdots \|m\alpha_d\|} = O\left(m^{1+\varepsilon}\right). \tag{3.7}$$

If we applied this estimate term by term in the sum, we would get

$$\sum_{m=1}^{M} \frac{1}{\|m\alpha_1\| \cdots \|m\alpha_d\|} = O\left(M^{2+\varepsilon}\right).$$

Compared to this, our proposition gives an exponent of $\frac{d+k-1}{k} + \varepsilon = 2 - \frac{1}{d} + \varepsilon$. How is this possible, since we know that the exponent in Schmidt's theorem is best possible? The reason is that even though (3.7) is best possible, there cannot be many values of m in the interval $[1, M]$ for which it is tight. In the proof of Proposition 3.6 we shall use the pigeonhole principle to exploit this fact, the result of which will be the appearance of $-\frac{1}{d}$ in the exponent. In the case $d = 1$ this method is well known, moreover it is known that the exponent $2 - \frac{1}{d} + \varepsilon = 1 + \varepsilon$ obtained is best possible. The case $d \ge 2$ seems to be a new result. Unfortunately the question whether the exponent $2 - \frac{1}{d} + \varepsilon$ is best possible is left open even for $d = 2$.

Generalizing the result to smaller values of k is simple, it does not require any new ideas. It might be worth mentioning that in the proof of Theorem 3.5 we are going to use Proposition 3.6 on the numbers $\alpha_1 = \frac{a_1}{a_2}, \dots, \alpha_{d-1} = \frac{a_1}{a_d}$, and on other similar $(d-1)$-tuples of pairwise ratios. Therefore if d and k are as in Theorem 3.5, we shall apply Proposition 3.6 with parameters $d-1$ and $k-1$.

Proof of Proposition 3.6: Let $[d]$ denote the set $\{1, 2, \ldots, d\}$, and let $\binom{[d]}{k}$ denote the family of subsets of $[d]$ of size k. Fix an index set $I \in \binom{[d]}{k}$. The numbers $\{\alpha_i : i \in I\}$ satisfy the conditions of Theorem 2.3 of Schmidt, therefore there exists a constant $0 < K_I < 1$ such that

$$\prod_{i \in I} \|m\alpha_i\| \geq \frac{K_I}{m^{1+\varepsilon}}$$

for every integer $m > 0$. Note that the ineffectiveness of Schmidt's theorem means that we cannot find an explicit value for K_I. In particular we have

$$\prod_{i \in I} \|m\alpha_i\| \geq f(M) \tag{3.8}$$

for every $1 \leq m \leq M$, where

$$f(M) = \frac{K_I}{M^{1+\varepsilon}}. \tag{3.9}$$

Our goal is to show that there cannot be many values of $1 \leq m \leq M$, for which 3.8 is close to being tight. In other words, we want to find an upper bound to the cardinality of the set

$$A_{I,t} = \left\{ 1 \leq m \leq M : \prod_{i \in I} \|m\alpha_i\| < tf(M) \right\},$$

where $t > 1$ is a constant.

Consider the map $g : A_{I,t} \to \left[-\frac{1}{2}, \frac{1}{2}\right)^k$ defined as

$$g(m) = (m\alpha_i : i \in I) \pmod 1.$$

For every $0 < c < \frac{1}{2^k}$ also consider the set

$$S_c = \left\{ x \in \left[-\frac{1}{2}, \frac{1}{2}\right)^k : |x_1 \cdots x_k| < c \right\}.$$

We start by showing that $\lambda(S_c) = O\left(c \log^{k-1} \frac{1}{c}\right)$, where the implied constant depends only on k. We can prove this claim by induction on k. The case $k = 1$ is trivial, since in this case S_c is simply an interval of length $2c$. Suppose the claim is true for $k - 1$

and let us prove it for k. By fixing the value of the last variable $x_k \in \left[-\frac{1}{2}, \frac{1}{2}\right)$, we can consider the section of the set S_c:

$$(S_c)_{x_k} = \left\{ (x_1, \dots, x_{k-1}) \in \left[-\frac{1}{2}, \frac{1}{2}\right)^{k-1} : |x_1 \cdots x_{k-1}| < \frac{c}{|x_k|} \right\}.$$

If the last variable lies in the interval $-2^{k-1}c < x_k < 2^{k-1}c$, then the section is $(S_c)_{x_k} = \left[-\frac{1}{2}, \frac{1}{2}\right)^{k-1}$, thus $\lambda\left((S_c)_{x_k}\right) = 1$. If $2^{k-1}c < |x_k| \le \frac{1}{2}$, then by the inductive hypothesis

$$\lambda\left((S_c)_{x_k}\right) = O\left(\frac{c}{|x_k|} \log^{k-2} \frac{|x_k|}{c}\right) = O\left(\frac{c}{|x_k|} \log^{k-2} \frac{1}{c}\right).$$

Therefore by Fubini's theorem

$$\lambda(S_c) = \int_{\left(-\frac{1}{2}, \frac{1}{2}\right)} \lambda\left((S_c)_{x_k}\right) \mathrm{d}x_k =$$

$$\int_{\left(-2^{k-1}c, 2^{k-1}c\right)} 1 \, \mathrm{d}x_k + \int_{\left(-\frac{1}{2}, -2^{k-1}c\right) \cup \left(2^{k-1}c, \frac{1}{2}\right)} \lambda\left((S_c)_{x_k}\right) \mathrm{d}x_k =$$

$$2^k c + O\left(c \log^{k-2} \frac{1}{c} \int_{\left(-\frac{1}{2}, -2^{k-1}c\right) \cup \left(2^{k-1}c, \frac{1}{2}\right)} \frac{1}{|x_k|} \, \mathrm{d}x_k\right) = O\left(c \log^{k-1} \frac{1}{c}\right).$$

Consider a partition of $\left[-\frac{1}{2}, \frac{1}{2}\right)^k$ into axis parallel cubes with side lengths between $\frac{1}{2} f(M)^{\frac{1}{k}}$ and $f(M)^{\frac{1}{k}}$. It is easy to see that such a partition exists. Indeed, from $0 < K_I < 1$ and from (3.9) we see that $0 < f(M) < 1$, thus $0 < f(M)^{\frac{1}{k}} < 1$. Then we can find the reciprocal of a positive integer between $\frac{1}{2} f(M)^{\frac{1}{k}}$ and $f(M)^{\frac{1}{k}}$, which can be chosen to be the side lengths of the axis parallel cubes in our partition. Let $\mathcal{C}$ denote the family of cubes in the partition.

Note that for every $m \in A_{I,t}$ we have $g(m) \in S_{tf(M)}$. Also note, that every cube in $\mathcal{C}$ contains at most one $g(m)$ with $1 \le m \le M$. Indeed, if $1 \le m < m' \le M$ and $g(m)$ and $g(m')$ belong to the same cube in $\mathcal{C}$, then

$$\left\| (m' - m)\alpha_i \right\| < f(M)^{\frac{1}{k}}$$

for every $i \in I$, therefore

$$\prod_{i\in I} \|(m'-m)\alpha_i\| < f(M).$$

This is a contradiction, since $1 \leq m' - m \leq M$. By the pigeonhole principle, we have that the cardinality of $A_{I,t}$ is at most as big, as the number of cubes in $\mathcal{C}$ which intersect the set $S_{tf(M)}$. We will show that the cubes in $\mathcal{C}$ which intersect $S_{tf(M)}$ are all contained in a similar set S_c, with c not too big. Indeed, let $x \in S_{tf(M)}$, or in other words $|x_1 \cdots x_k| < tf(M)$. Consider

$$\left(|x_1| + f(M)^{\frac{1}{k}}\right) \cdots \left(|x_k| + f(M)^{\frac{1}{k}}\right).$$

When expanding the product, every term we get will be the product of certain $|x_i|$'s and $f(M)^{\frac{1}{k}}$ raised to a certain power. Let us use the bound $|x_1 \cdots x_k| < tf(M)$ on the first term, and let us simply use the bound $|x_i| \leq \frac{1}{2}$ on all the other terms. This way we get

$$\left(|x_1| + f(M)^{\frac{1}{k}}\right) \cdots \left(|x_k| + f(M)^{\frac{1}{k}}\right) = O\left(tf(M) + f(M)^{\frac{1}{k}} + f(M)^{\frac{2}{k}} + \cdots + f(M)\right).$$

Since $0 < f(M) < 1$, we have

$$\left(|x_1| + f(M)^{\frac{1}{k}}\right) \cdots \left(|x_k| + f(M)^{\frac{1}{k}}\right) = O\left(tf(M) + f(M)^{\frac{1}{k}}\right).$$

This estimate shows that if $x \in S_{tf(M)}$, then the cube in $\mathcal{C}$ containing x lies completely within S_c with $c = O\left(tf(M) + f(M)^{\frac{1}{k}}\right)$. In other words, the cubes in $\mathcal{C}$ which intersect $S_{tf(M)}$, are all contained in S_c with $c = O\left(tf(M) + f(M)^{\frac{1}{k}}\right)$. The Lebesgue measure of the cubes in $\mathcal{C}$ is at least $\frac{1}{2^k} f(M)$. The Lebesgue measure of S_c is

$$O\left(c \log^{k-1} \frac{1}{c}\right) = O\left(\left(tf(M) + f(M)^{\frac{1}{k}}\right) \log^{k-1} \frac{1}{tf(M) + f(M)^{\frac{1}{k}}}\right).$$

Using (3.9) we can simplify the logarithmic factor as

$$\log^{k-1} \frac{1}{tf(M) + f(M)^{\frac{1}{k}}} \leq \log^{k-1} \frac{1}{f(M)^{\frac{1}{k}}} = O\left(\log^{k-1} M\right),$$

thus

$$\lambda(S_c) = O\left(\left(tf(M) + f(M)^{\frac{1}{k}}\right)\log^{k-1} M\right).$$

Comparing the Lebesgue measure of the cubes and that of S_c, we obtain that the number of cubes in $\mathcal{C}$ which intersect $S_{tf(M)}$ is

$$O\left(\frac{\lambda(S_c)}{f(M)}\right) = O\left(\left(t + f(M)^{\frac{1}{k}-1}\right)\log^{k-1} M\right) = O\left(\left(t + M^{\left(1-\frac{1}{k}\right)(1+\varepsilon)}\right)\log^{k-1} M\right).$$

This means that

$$|A_{I,t}| = O\left(\left(t + M^{\left(1-\frac{1}{k}\right)(1+\varepsilon)}\right)\log^{k-1} M\right) \tag{3.10}$$

holds for any $I \in \binom{[d]}{k}$ and $t > 1$.

We are now ready to prove the proposition. We will decompose the original sum into terms of the same order of magnitude. Let

$$K = \min\left\{K_I : I \in \binom{[d]}{k}\right\}.$$

Then from (3.8) and (3.9) we have

$$\prod_{i\in I} \|m\alpha_i\| \geq \frac{K}{M^{1+\varepsilon}}$$

for every $I \in \binom{[d]}{k}$ and $1 \leq m \leq M$. Let us multiply this inequality together for every index set I to obtain

$$\prod_{I\in\binom{[d]}{k}} \prod_{i\in I} \|m\alpha_i\| \geq \left(\frac{K}{M^{1+\varepsilon}}\right)^{\binom{d}{k}}.$$

When switching the order of the two products, notice that every factor $\|m\alpha_i\|$ appears $\binom{d-1}{k-1}$ times, thus

$$\left(\|m\alpha_1\| \cdots \|m\alpha_d\|\right)^{\binom{d-1}{k-1}} \geq \left(\frac{K}{M^{1+\varepsilon}}\right)^{\binom{d}{k}}.$$

Finally, using $\frac{\binom{d}{k}}{\binom{d-1}{k-1}} = \frac{d}{k}$ we get that

$$\|m\alpha_1\| \cdots \|m\alpha_d\| \geq \left(\frac{K}{M^{1+\varepsilon}}\right)^{\frac{d}{k}}$$

holds for every $1 \leq m \leq M$. For every integer $\ell \geq 0$ let

$$B_\ell = \left\{1 \leq m \leq M : 2^\ell \left(\frac{K}{M^{1+\varepsilon}}\right)^{\frac{d}{k}} \leq \|m\alpha_1\| \cdots \|m\alpha_d\| < 2^{\ell+1} \left(\frac{K}{M^{1+\varepsilon}}\right)^{\frac{d}{k}}\right\}.$$

Then $B_0, B_1, \ldots$ is a partition of $[1, M]$. First of all note, that if $2^\ell \left(\frac{K}{M^{1+\varepsilon}}\right)^{\frac{d}{k}} > 1$, then $B_\ell = \emptyset$. Therefore it will be enough to consider $0 \leq \ell \leq L$, where $L = O(\log M)$. Also note, that for every ℓ we have

$$B_\ell \subseteq \bigcup_{I \in \binom{[d]}{k}} A_{I,t}$$

with $t = 2^{\frac{k}{d}(\ell+1)}$. Indeed, for any $m \in B_\ell$ we have

$$\|m\alpha_1\| \cdots \|m\alpha_d\| < \left(2^{\frac{k}{d}(\ell+1)} \frac{K}{M^{1+\varepsilon}}\right)^{\frac{d}{k}},$$

$$(\|m\alpha_1\| \cdots \|m\alpha_d\|)^{\binom{d-1}{k-1}} < \left(2^{\frac{k}{d}(\ell+1)} \frac{K}{M^{1+\varepsilon}}\right)^{\binom{d}{k}},$$

$$\left(\prod_{I \in \binom{[d]}{k}} \prod_{i \in I} \|m\alpha_i\|\right)^{\frac{1}{\binom{d}{k}}} < 2^{\frac{k}{d}(\ell+1)} \frac{K}{M^{1+\varepsilon}}.$$

Since the left hand side is a geometric mean, we get that there exists an index set $I \in \binom{[d]}{k}$ such that

$$\prod_{i \in I} \|m\alpha_i\| < 2^{\frac{k}{d}(\ell+1)} \frac{K}{M^{1+\varepsilon}} \leq 2^{\frac{k}{d}(\ell+1)} \frac{K_I}{M^{1+\varepsilon}}.$$

This means that $m \in A_{I,t}$ for $t = 2^{\frac{k}{d}(\ell+1)}$, as claimed. Using (3.10) this implies, that

$$|B_\ell| \leq \sum_{I \in \binom{[d]}{k}} \left|A_{I, 2^{\frac{k}{d}(\ell+1)}}\right| = O\left(\left(2^{\frac{k}{d}(\ell+1)} + M^{\left(1-\frac{1}{k}\right)(1+\varepsilon)}\right) \log^{k-1} M\right).$$

Finally, decomposing our original sum using the sets $B_0, B_1, \ldots, B_L$ we get

$$\sum_{m=1}^{M} \frac{1}{\|m\alpha_1\| \cdots \|m\alpha_d\|} = \sum_{\ell=0}^{L} \sum_{m \in B_\ell} \frac{1}{\|m\alpha_1\| \cdots \|m\alpha_d\|} \leq \sum_{\ell=0}^{L} |B_\ell| \frac{1}{2^\ell} \left(\frac{M^{1+\varepsilon}}{K} \right)^{\frac{d}{k}} =$$

$$O\left(\sum_{\ell=0}^{L} \left(\left(2^{\left(\frac{k}{d}-1\right)\ell} M^{\frac{d}{k}(1+\varepsilon)} + \frac{1}{2^\ell} M^{\left(1-\frac{1}{k}+\frac{d}{k}\right)(1+\varepsilon)} \right) \log^{k-1} M \right) \right) =$$

$$O\left(\left(LM^{\frac{d}{k}(1+\varepsilon)} + M^{\frac{d+k-1}{k}(1+\varepsilon)} \right) \log^{k-1} M \right) = O\left(M^{\frac{d+k-1}{k}(1+\varepsilon)} \log^k M \right).$$

This is true for any $\varepsilon > 0$. By switching to a different $\varepsilon > 0$, we get that the same sum is $O\left(M^{\frac{d+k-1}{k}+\varepsilon}\right)$ for any $\varepsilon > 0$.

□

Proof of Theorem 3.5: The polytope P has 2^d hyperfaces, the normal vectors of which are of the form $\left(\frac{\pm 1}{a_1}, \ldots, \frac{\pm 1}{a_d}\right)$. Thus P satisfies the conditions of Theorem 2.5 with the same k value as in the theorem. Hence for any integer $N > 1$ we have

$$\left| tP \cap \mathbb{Z}^d \right| - p(t) = C(tP, N) - p(t) + O\left(t^{d-k} + t^{d-1+\varepsilon} \sqrt{\frac{\log N}{N}} \right),$$

where $C(tP, N)$ is as in Definition 2.3. By taking the average over the interval $[T_1, T_2]$ we get

$$\frac{1}{T_2 - T_1} \int_{T_1}^{T_2} \left(\left| tP \cap \mathbb{Z}^d \right| - p(t) \right) \mathrm{d}t =$$

$$\frac{1}{T_2 - T_1} \int_{T_1}^{T_2} \left(C(tP, N) - p(t) \right) \mathrm{d}t + O\left(T_2^{d-k} + T_2^{d-1+\varepsilon} \sqrt{\frac{\log N}{N}} \right). \tag{3.11}$$

Let us fix integers $M_1, \ldots, M_d \geq 0$. Let $M = (M_1, \ldots, M_d)$, and consider the partial sums $S(tP, M)$ as in Definition 2.2. According to Proposition 3.2 we have

$$S(tP, M) = \frac{(-1)^d 2^d a_1 \cdots a_d}{(2\pi i)^{d+1}} \sum_{m_1=-M_1}^{M_1} \cdots \sum_{m_d=-M_d}^{M_d} \int_{|z|=R} \frac{e^{-2\pi i z t}}{z(z - m_1 a_1) \cdots (z - m_d a_d)} \, \mathrm{d}z,$$

where $R > \max_k M_k a_k$. For any given term indexed by $(m_1, \ldots, m_d)$ the function

$$\frac{e^{-2\pi izt}}{z(z-m_1a_1)\cdots(z-m_da_d)}$$

has a singularity at zero, and singularities at m_ja_j for every j such that $m_j \neq 0$. Notice that the conditions of the theorem imply that the ratio $\frac{a_j}{a_{j'}}$ is irrational for any $j \neq j'$, therefore the singularity at m_ja_j is a simple pole. The order of the pole at zero depends on how many zero coordinates $(m_1, \ldots, m_d)$ has. Let us use the residue theorem to evaluate the complex line integral, and consider the residues at zero, and the residues at m_ja_j with $m_j \neq 0$ separately. In Proposition 3.3 we computed the contribution of the residues at zero. If $m_j \neq 0$ then

$$\mathrm{Res}_{m_ja_j} \frac{e^{-2\pi izt}}{z(z-m_1a_1)\cdots(z-m_da_d)} = \frac{e^{-2\pi im_ja_jt}}{m_ja_j\prod_{j'\neq j}(m_ja_j-m_{j'}a_{j'})}.$$

Therefore we get

$$S(tP,M) = \frac{(-1)^d 2^d a_1\cdots a_d}{(2\pi i)^d} \sum_{m_1=-M_1}^{M_1} \cdots \sum_{m_d=-M_d}^{M_d} \sum_{\substack{1\leq j\leq d\\ m_j\neq 0}} \frac{e^{-2\pi im_ja_jt}}{m_ja_j\prod_{j'\neq j}(m_ja_j-m_{j'}a_{j'})} +$$

$$p(t) + O\left(\frac{t^{d-2}}{M_1+1} + \cdots + \frac{t^{d-2}}{M_d+1}\right).$$

By subtracting the main term $p(t)$, and taking the average over $[T_1, T_2]$, we get

$$\frac{1}{T_2-T_1}\int_{T_1}^{T_2} (S(tP,M)-p(t))\,\mathrm{d}t =$$

$$O\left(\sum_{m_1=-M_1}^{M_1} \cdots \sum_{m_d=-M_d}^{M_d} \sum_{\substack{1\leq j\leq d\\ m_j\neq 0}} \frac{1}{T_2-T_1}\int_{T_1}^{T_2} \frac{e^{-2\pi im_ja_jt}}{m_ja_j\prod_{j'\neq j}(m_ja_j-m_{j'}a_{j'})}\,\mathrm{d}t\right) +$$

$$O\left(\frac{T_2^{d-2}}{M_1+1} + \cdots + \frac{T_2^{d-2}}{M_d+1}\right). \tag{3.12}$$

We want to show that the first error term is small. It is more convenient to switch the order of summation by replacing

$$\sum_{m_1=-M_1}^{M_1} \cdots \sum_{m_d=-M_d}^{M_d} \sum_{\substack{1\leq j\leq d\\ m_j\neq 0}}$$

by

$$\sum_{j=1}^{d} \sum_{m_1=-M_1}^{M_1} \cdots \sum_{\substack{m_j=-M_j \\ m_j\neq 0}}^{M_j} \cdots \sum_{m_d=-M_d}^{M_d}.$$

For the sake of simplicity, we will work with the $j = 1$ term, the others being similar. Note that the integral satisfies

$$\left|\frac{1}{T_2-T_1}\int_{T_1}^{T_2} e^{-2\pi i m_1 a_1 t}\,\mathrm{d}t\right| \le \min\left(1, \frac{1}{(T_2-T_1)\pi|m_1|a_1}\right).$$

Thus it is enough to find an upper bound to

$$\sum_{\substack{m_1=-M_1 \\ m_1\neq 0}}^{M_1} \frac{1}{|m_1|}\left|\sum_{m_2=-M_2}^{M_2} \frac{1}{m_1a_1-m_2a_2}\right| \cdots \left|\sum_{m_d=-M_d}^{M_d} \frac{1}{m_1a_1-m_da_d}\right|.$$

$$\min\left(1, \frac{1}{(T_2-T_1)|m_1|}\right). \tag{3.13}$$

For the sake of simplicity, we will work with the sum over m_2, the others being similar. First separate the $m_2 = 0$ term, then combine the m_2 and $-m_2$ terms to get

$$\left|\sum_{m_2=-M_2}^{M_2} \frac{1}{m_1a_1-m_2a_2}\right| \le \frac{1}{|m_1a_1|} + \left|\sum_{m_2=1}^{M_2} \frac{2m_1a_1}{(m_1a_1)^2-(m_2a_2)^2}\right|.$$

Let $\alpha = \frac{a_1}{a_2}$. Then

$$\left|\sum_{m_2=-M_2}^{M_2} \frac{1}{m_1a_1-m_2a_2}\right| \le \frac{1}{|m_1a_1|} + \frac{2|m_1a_1|}{a_2^2}\sum_{m_2=1}^{M_2} \frac{1}{\left|(m_1\alpha)^2-m_2^2\right|}.$$

Let $b = b(m_1\alpha)$ be the integer for which

$$b < |m_1\alpha| < b+1.$$

The $m_2 = b$ term satisfies

$$\frac{1}{|(m_1\alpha)^2-b^2|} = \frac{1}{(|m_1\alpha|-b)\cdot(|m_1\alpha|+b)} \le \frac{1}{\|m_1\alpha\|\cdot|m_1\alpha|}.$$

The same bound holds for the $m_2 = b+1$ term. The sum of all the other terms is small. To see this, first consider

$$\sum_{m_2=1}^{b-1} \frac{1}{\left|(m_1\alpha)^2 - m_2^2\right|} = \sum_{m_2=1}^{b-1} \frac{1}{(|m_1\alpha| - m_2)\,(|m_1\alpha| + m_2)} \leq$$

$$\frac{1}{|m_1\alpha|} \sum_{m_2=1}^{b-1} \frac{1}{|m_1\alpha| - m_2} = O\left(\frac{\log|m_1|}{|m_1|}\right).$$

We also have

$$\sum_{m_2=b+2}^{M_2} \frac{1}{\left|(m_1\alpha)^2 - m_2^2\right|} \leq \sum_{m_2=b+2}^{\infty} \frac{1}{m_2^2 - (m_1\alpha)^2} \leq$$

$$\frac{1}{(b+2)^2 - (m_1\alpha)^2} + \int_{b+2}^{\infty} \frac{1}{x^2 - (m_1\alpha)^2}\,\mathrm{d}x \leq$$

$$\frac{1}{2|m_1\alpha|} + \int_{b+2}^{\infty} \left(\frac{\frac{1}{2|m_1\alpha|}}{x - |m_1\alpha|} - \frac{\frac{1}{2|m_1\alpha|}}{x + |m_1\alpha|}\right) \mathrm{d}x =$$

$$\frac{1}{2|m_1\alpha|} + \frac{1}{2|m_1\alpha|} \log \frac{b+2+|m_1\alpha|}{b+2-|m_1\alpha|} = O\left(\frac{\log|m_1|}{|m_1|}\right).$$

Altogether we got, that

$$\left|\sum_{m_2=-M_2}^{M_2} \frac{1}{m_1a_1 - m_2a_2}\right| = O\left(\frac{1}{|m_1|} + \frac{1}{\left\|m_1\frac{a_1}{a_2}\right\|} + \log|m_1|\right) = O\left(\frac{\log|m_1|}{\left\|m_1\frac{a_1}{a_2}\right\|}\right).$$

Similar bounds hold for the sums over $m_3, \ldots, m_d$. Applying these bounds, we find that (3.13) is at most constant times

$$\sum_{m_1=1}^{M_1} \frac{\log^{d-1} m_1}{m_1 \left\|m_1\frac{a_1}{a_2}\right\| \cdots \left\|m_1\frac{a_1}{a_d}\right\|} \cdot \min\left(1, \frac{1}{(T_2 - T_1)m_1}\right).$$

This is exactly the setup of Proposition 3.6 with $\alpha_1 = \frac{a_1}{a_2}, \ldots, \alpha_{d-1} = \frac{a_1}{a_d}$. The conditions of the theorem imply, that for any $1 \leq i_1 < \ldots < i_{k-1} \leq d-1$ the numbers $\frac{1}{a_1}, \frac{1}{a_{i_1+1}}, \ldots, \frac{1}{a_{i_{k-1}+1}}$ are linearly independent over $\mathbb{Q}$. Multiplying these numbers by a_1 preserves the linear independence, therefore $1, \alpha_{i_1}, \ldots, \alpha_{i_{k-1}}$ are also linearly independent.

We will consider the terms $1 \le m_1 \le \frac{1}{T_2-T_1}$ and $\frac{1}{T_2-T_1} < m_1 \le M_1$ separately. The factor $\min\left(1, \frac{1}{(T_2-T_1)m_1}\right)$ equals 1 in the first case, while $\frac{1}{(T_2-T_2)m_1}$ in the second case. To estimate the sum of the terms over $1 \le m_1 \le \frac{1}{T_2-T_1}$, let us fix an integer $\ell \ge 0$ first, and consider

$$\sum_{2^\ell \le m_1 < 2^{\ell+1}} \frac{\log^{d-1} m_1}{m_1 \left\|m_1 \frac{a_1}{a_2}\right\| \cdots \left\|m_1 \frac{a_1}{a_d}\right\|} \le \frac{\log^{d-1} 2^{\ell+1}}{2^\ell} \sum_{2^\ell \le m_1 < 2^{\ell+1}} \frac{1}{\left\|m_1 \frac{a_1}{a_2}\right\| \cdots \left\|m_1 \frac{a_1}{a_d}\right\|}.$$

Proposition 3.6 with $M = 2^{\ell+1}$ and parameters $d-1$, $k-1$ implies, that this sum is

$$O\left(\frac{\ell^{d-1}}{2^\ell}\left(2^\ell\right)^{\frac{d+k-3}{k-1}+\varepsilon}\right) = O\left(\left(2^\ell\right)^{\frac{d-2}{k-1}+2\varepsilon}\right).$$

Summing this error term over all integers $\ell \ge 0$ such that $2^\ell \le \frac{1}{T_2-T_1}$ yields

$$\sum_{1 \le m_1 \le \frac{1}{T_2-T_1}} \frac{\log^{d-1} m_1}{m_1 \left\|m_1 \frac{a_1}{a_2}\right\| \cdots \left\|m_1 \frac{a_1}{a_d}\right\|} = O\left(\sum_{2^\ell \le \frac{1}{T_2-T_1}} \left(2^\ell\right)^{\frac{d-2}{k-1}+2\varepsilon}\right) =$$

$$O\left(\left(\frac{1}{T_2-T_1}\right)^{\frac{d-2}{k-1}+2\varepsilon}\right). \tag{3.14}$$

Consider now the sum over $\frac{1}{T_2-T_1} < m_1 \le M_1$. For any fixed integer $\ell \ge 0$ we have

$$\sum_{2^\ell \le m_1 < 2^{\ell+1}} \frac{\log^{d-1} m_1}{m_1 \left\|m_1 \frac{a_1}{a_2}\right\| \cdots \left\|m_1 \frac{a_1}{a_d}\right\|} \cdot \frac{1}{(T_2-T_1)m_1} \le$$

$$\frac{\log^{d-1} 2^{\ell+1}}{(T_2-T_1)2^{2\ell}} \sum_{2^\ell \le m_1 < 2^{\ell+1}} \frac{1}{\left\|m_1 \frac{a_1}{a_2}\right\| \cdots \left\|m_1 \frac{a_1}{a_d}\right\|}.$$

Proposition 3.6 with $M = 2^{\ell+1}$ and parameters $d-1$, $k-1$ implies that this sum is

$$O\left(\frac{\ell^{d-1}}{(T_2-T_1)2^{2\ell}}\left(2^\ell\right)^{\frac{d+k-3}{k-1}+\varepsilon}\right) = O\left(\frac{1}{T_2-T_1}\left(2^\ell\right)^{\frac{d-k-1}{k-1}+2\varepsilon}\right).$$

Summing this error term over all integers $\ell \ge 0$ such that $\frac{1}{2(T_2-T_1)} \le 2^\ell \le M_1$ yields

$$\sum_{\frac{1}{T_2-T_1} < m_1 \le M_1} \frac{\log^{d-1} m_1}{m_1 \left\|m_1 \frac{a_1}{a_2}\right\| \cdots \left\|m_1 \frac{a_1}{a_d}\right\|} \cdot \frac{1}{(T_2-T_1)m_1} =$$

$$O\left(\sum_{\frac{1}{2(T_2-T_1)} \le 2^\ell \le M_1} \frac{1}{T_2-T_1}\left(2^\ell\right)^{\frac{d-k-1}{k-1}+2\varepsilon}\right).$$

This is the point, where the case $k = d$ and the case $2 \le k \le d-1$ are qualitatively different. If $k = d$, the exponent $\frac{d-k-1}{k-1} + 2\varepsilon = \frac{-1}{d-1} + 2\varepsilon$ is negative, if we choose ε to be small enough. Thus if $k = d$, we have

$$\sum_{\frac{1}{T_2-T_1}<m_1\le M_1} \frac{\log^{d-1} m_1}{m_1 \left\|m_1\frac{a_1}{a_2}\right\| \cdots \left\|m_1\frac{a_1}{a_d}\right\|} \cdot \frac{1}{(T_2-T_1)m_1} = O\left(\left(\frac{1}{T_2-T_1}\right)^{\frac{d-2}{d-1}+2\varepsilon}\right). \quad (3.15)$$

On the other hand if $2 \le k \le d-1$, then the exponent $\frac{d-k-1}{k-1} + 2\varepsilon > 0$, therefore

$$\sum_{\frac{1}{T_2-T_1}<m_1\le M_1} \frac{\log^{d-1} m_1}{m_1 \left\|m_1\frac{a_1}{a_2}\right\| \cdots \left\|m_1\frac{a_1}{a_d}\right\|} \cdot \frac{1}{(T_2-T_1)m_1} = O\left(\frac{1}{T_2-T_1} \cdot M_1^{\frac{d-k-1}{k-1}+2\varepsilon}\right). \quad (3.16)$$

Adding (3.14), and (3.15) or (3.16) we get that (3.13) is

$$O\left(\left(\frac{1}{T_2-T_1}\right)^{\frac{d-2}{d-1}+2\varepsilon}\right), \quad (3.17)$$

if $k = d$, and

$$O\left(\left(\frac{1}{T_2-T_1}\right)^{\frac{d-2}{k-1}+2\varepsilon} + \frac{1}{T_2-T_1} \cdot M_1^{\frac{d-k-1}{k-1}+2\varepsilon}\right), \quad (3.18)$$

if $2 \le k \le d-1$.

To prove (i), let $k = d$ and let us use (3.17) in (3.12) to get

$$\frac{1}{T_2-T_1}\int_{T_1}^{T_2} (S(tP, M) - p(t))\, \mathrm{d}t =$$

$$O\left(\left(\frac{1}{T_2-T_1}\right)^{\frac{d-2}{d-1}+2\varepsilon} + \frac{T_2^{d-2}}{M_1+1} + \cdots + \frac{T_2^{d-2}}{M_d+1}\right).$$

Taking the average of this over $(M_1, \ldots, M_d) \in [0, N-1]^d$ yields

$$\frac{1}{T_2-T_1}\int_{T_1}^{T_2} (C(tP, N) - p(t))\, \mathrm{d}t = O\left(\left(\frac{1}{T_2-T_1}\right)^{\frac{d-2}{d-1}+2\varepsilon} + T_2^{d-2}\frac{\log N}{N}\right).$$

Applying this bound in (3.11) in the special case $k = d$ gives us

$$\frac{1}{T_2-T_1}\int_{T_1}^{T_2}\left(\left|tP\cap\mathbb{Z}^d\right|-p(t)\right)\,\mathrm{d}t=$$

$$O\left(1+T_2^{d-1+\varepsilon}\sqrt{\frac{\log N}{N}}+\left(\frac{1}{T_2-T_1}\right)^{\frac{d-2}{d-1}+2\varepsilon}+T_2^{d-2}\frac{\log N}{N}\right).$$

Taking the limit, as $N\to\infty$ finishes the proof of (i).

To prove (ii) consider the special case $2\le k\le d-1$. Let us use (3.18) in (3.12) to get

$$\frac{1}{T_2-T_1}\int_{T_1}^{T_2}\left(S(tP,M)-p(t)\right)\,\mathrm{d}t=$$

$$O\left(\left(\frac{1}{T_2-T_1}\right)^{\frac{d-2}{k-1}+2\varepsilon}+\frac{1}{T_2-T_1}\sum_{j=1}^{d}M_j^{\frac{d-k-1}{k-1}+2\varepsilon}+\sum_{j=1}^{d}\frac{T_2^{d-2}}{M_j+1}\right).$$

Taking the average of this over $(M_1,\dots,M_d)\in[0,N-1]^d$ yields

$$\frac{1}{T_2-T_1}\int_{T_1}^{T_2}\left(C(tP,N)-p(t)\right)\,\mathrm{d}t=$$

$$O\left(\left(\frac{1}{T_2-T_1}\right)^{\frac{d-2}{k-1}+2\varepsilon}+\frac{1}{T_2-T_1}\cdot N^{\frac{d-k-1}{k-1}+2\varepsilon}+T_2^{d-2}\frac{\log N}{N}\right).$$

Applying this bound in (3.11), and noticing $T_2^{d-2}\frac{\log N}{N}\le T_2^{d-1+\varepsilon}\sqrt{\frac{\log N}{N}}$ gives us

$$\frac{1}{T_2-T_1}\int_{T_1}^{T_2}\left(\left|tP\cap\mathbb{Z}^d\right|-p(t)\right)\,\mathrm{d}t=$$

$$O\left(T_2^{d-k}+T_2^{d-1+\varepsilon}\sqrt{\frac{\log N}{N}}+\left(\frac{1}{T_2-T_1}\right)^{\frac{d-2}{k-1}+2\varepsilon}+\frac{1}{T_2-T_1}\cdot N^{\frac{d-k-1}{k-1}+2\varepsilon}\right).$$

We want to choose N to be the integer closest to the solution of

$$T_2^{d-1}\cdot\frac{1}{\sqrt{N}}=\frac{1}{T_2-T_1}\cdot N^{\frac{d-k-1}{k-1}}. \tag{3.19}$$

Since it is easy to see, that for this choice of N we have both $\sqrt{\log N}=O\left(T_2^{\varepsilon}\right)$ and $N^{2\varepsilon}=O\left(T_2^{\varepsilon'}\right)$ for some $\varepsilon'=O(\varepsilon)$, in this case the error term we get will be exactly

the same as in the theorem. The only case when this choice of N is not admissible, is when $\frac{1}{T_2-T_1} \geq T_2^{d-1}$, in which case the solution of (3.19) possibly has limit zero as $T_2 \to \infty$. But if $\frac{1}{T_2-T_1} \geq T_2^{d-1}$, then the third error term in the theorem satisfies

$$\left(\frac{1}{T_2 - T_1}\right)^{\frac{d-2}{k-1}+\varepsilon} = \Omega\left(T_2^{d-1}\right),$$

which means that the bound we are trying to prove is weaker than the trivial discrepancy bound in Corollary 1.2.

□

3.2.3 Uniform bound on the fluctuating term

Let $a_1, \ldots, a_d > 0$ be algebraic, and let P be as in the polyhedral sphere problem. In subsection 3.2.1 we found that the main term of $\left|tP \cap \mathbb{Z}^d\right|$ is the polynomial $p(t)$ defined in Definition 3.1. We now want to study how large the random fluctuation

$$\left|tP \cap \mathbb{Z}^d\right| - p(t)$$

can be as a function of t. The most general result is stated as follows.

Theorem 3.7 *Let $2 \leq k \leq d$ be integers, and let $a_1, \ldots, a_d > 0$ be algebraic. Suppose that any k numbers out of $\frac{1}{a_1}, \ldots, \frac{1}{a_d}$ are linearly independent over $\mathbb{Q}$. Consider the polytope*

$$P = \left\{x \in \mathbb{R}^d : \frac{|x_1|}{a_1} + \cdots + \frac{|x_d|}{a_d} \leq 1\right\},$$

and the polynomial $p(t)$ as in Definition 3.1. Let $\varepsilon > 0$ and $t > 1$.

(i) *If $k = d$, then*

$$\left|tP \cap \mathbb{Z}^d\right| - p(t) = O\left(t^{\frac{(d-1)(d-2)}{2d-3}+\varepsilon}\right).$$

(ii) *If $2 \leq k \leq d-1$, then*

$$\left|tP \cap \mathbb{Z}^d\right| - p(t) = O\left(t^{\frac{(d-1)(2d-k-3)}{2d-4}+\varepsilon}\right).$$

The implied constants in (i) and (ii) depend only on $a_1, \dots, a_d$ and ε, and are ineffective.

Note that the conditions of the theorem are the same as those of Theorem 3.5. Under the strongest condition $k = d$, in other words when $\frac{1}{a_1}, \dots, \frac{1}{a_d}$ are linearly independent over $\mathbb{Q}$, the exponent $\frac{(d-1)(d-2)}{2d-3} + \varepsilon$ of the error term is roughly $\frac{d}{2}$ for large values of d. Even under the weakest assumption $k = 2$, in other words when all we assume is that the pairwise ratios $\frac{a_i}{a_j}$ are irrational for every $i \neq j$, the error term $t^{\frac{(d-1)(2d-5)}{2d-4}+\varepsilon}$ is smaller than the trivial discrepancy bound t^{d-1} of Corollary 1.2. To get a better intuition of the behavior of the exponent for intermediate values of k, note that for all $2 \leq k \leq d-1$ we have

$$\frac{(d-1)(2d-k-3)}{2d-4} < d - \frac{k+1}{2}.$$

The proof of Theorem 3.7 relies on the simple observation, that $\left|tP \cap \mathbb{Z}^d\right|$ is a monotone increasing function of t. This will allow us to use a "mean to maximum" type estimate. The basic idea is that if the fluctuation $\left|tP \cap \mathbb{Z}^d\right| - p(t)$ has a large positive value at some t, then it has a large positive value on a short interval $[t, t+\delta]$, where $\delta = o(t)$. But applying Theorem 3.5 on the short interval $[t, t+\delta]$ shows that the average fluctuation cannot be large, which is a contradiction.

The idea of using a mean to maximum type estimate is from Lemma 6.3.2 in [13]. The method is used there to study the L^2 discrepancy of a planar region. Our proof of Theorem 3.7 will be a modified version of the proof of Lemma 6.3.2 in [13] to study the maximal fluctuation instead.

Proof of Theorem 3.7: First note, that since the origin belongs to P, we have that $tP \subseteq t'P$ for $0 < t \leq t'$. Therefore $\left|tP \cap \mathbb{Z}^d\right|$ is a monotone increasing function of t. Fix $t \geq 1$. Since p is a polynomial of degree d, there exists a constant $K > 1$ such that for any $u \in [t-1, t+1]$ we have

$$|p(t) - p(u)| \leq Kt^{d-1}|t-u|. \tag{3.20}$$

Fix a real number $0 < a < t^{d-1}$, and let $\delta = \frac{a}{2Kt^{d-1}}$. Clearly $0 < \delta < 1$.

Suppose first, that

$$\left|tP \cap \mathbb{Z}^d\right| - p(t) \geq a.$$

Then using the monotonicity of the number of lattice points as a function and (3.20), we get that for any $u \in [t, t+\delta]$ we have

$$\left|uP \cap \mathbb{Z}^d\right| - p(u) \geq \left|tP \cap \mathbb{Z}^d\right| - p(t) + p(t) - p(u) \geq a - Kt^{d-1}\delta = \frac{a}{2}.$$

Taking the average of this inequality over $u \in [t, t+\delta]$ we get

$$\frac{1}{\delta}\int_t^{t+\delta} \left(\left|uP \cap \mathbb{Z}^d\right| - p(u)\right) \,\mathrm{d}u \geq \frac{a}{2}.$$

Now we want to apply Theorem 3.5 on the interval $[T_1, T_2] = [t, t+\delta]$. Using $\delta = \frac{a}{2Kt^{d-1}}$ we get the error bound $E(a,t)$ in terms of a and t

$$E(a,t) = \left(\frac{t^{d-1}}{a}\right)^{\frac{d-2}{d-1}+\varepsilon},$$

if $k = d$, and

$$E(a,t) = t^{d-k} + t^{\frac{2(d-1)(d-k-1)}{2d-k-3}+\varepsilon}\left(\frac{t^{d-1}}{a}\right)^{\frac{k-1}{2d-k-3}} + \left(\frac{t^{d-1}}{a}\right)^{\frac{d-2}{k-1}+\varepsilon},$$

if $2 \leq k \leq d-1$.

With this notation Theorem 3.5 says, that

$$\frac{1}{\delta}\int_t^{t+\delta} \left(\left|uP \cap \mathbb{Z}^d\right| - p(u)\right) \,\mathrm{d}u = O\left(E(a,t)\right).$$

Thus we showed that if $0 < a < t^{d-1}$ is such that $\left|tP \cap \mathbb{Z}^d\right| - p(t) \geq a$, then $a = O\left(E(a,t)\right)$.

A similar argument shows the same is true, when $\left|tP \cap \mathbb{Z}^d\right| - p(t) \leq -a$. Indeed, in this case for every $u \in [t-\delta, t]$ we have

$$\left|uP \cap \mathbb{Z}^d\right| - p(u) \leq \left|tP \cap \mathbb{Z}^d\right| - p(t) + p(t) - p(u) \leq -a + Kt^{d-1}\delta = -\frac{a}{2}.$$

Taking the average of this inequality over $u \in [t-\delta, t]$ we get

$$\frac{1}{\delta}\int_{t-\delta}^{t} \left(\left|uP \cap \mathbb{Z}^d\right| - p(u)\right) \,\mathrm{d}u \leq -\frac{a}{2}.$$

Applying Theorem 3.5 on the interval $[T_1, T_2] = [t-\delta, t]$ we get the same error bound as before, therefore $a = O\left(E(a,t)\right)$.

Altogether we showed, that if $0 < a < t^{d-1}$ is such that $\left|\left|tP \cap \mathbb{Z}^d\right| - p(t)\right| \geq a$, then $a = O\left(E(a,t)\right)$. The trivial discrepancy bound in Corollary 1.2 implies, that we can choose $0 < a < t^{d-1}$ with $a = \Theta\left(\left|\left|tP \cap \mathbb{Z}^d\right| - p(t)\right|\right)$. Therefore

$$\left|\left|tP \cap \mathbb{Z}^d\right| - p(t)\right| = O\left(E\left(\left|\left|tP \cap \mathbb{Z}^d\right| - p(t)\right|, t\right)\right).$$

Now we claim, for the sake of simplicity, that

$$\left|\left|tP \cap \mathbb{Z}^d\right| - p(t)\right| = O\left(a + E(a,t)\right) \tag{3.21}$$

for any real number $0 < a < t^{d-1}$. This is easy to see by observing, that $E(\cdot, t)$ is monotone decreasing. Therefore for any choice of $0 < a < t^{d-1}$, either a or $E(a,t)$ will have a larger order of magnitude, than $\left|\left|tP \cap \mathbb{Z}^d\right| - p(t)\right|$.

If $k = d$, (3.21) becomes

$$\left|\left|tP \cap \mathbb{Z}^d\right| - p(t)\right| = O\left(a + \frac{t^{d-2+(d-1)\varepsilon}}{a^{\frac{d-2}{d-1}+\varepsilon}}\right).$$

The optimal choice for a is $a = t^{\frac{(d-1)(d-2)}{2d-3}}$, in which case the two error terms have a similar order of magnitude.

If $2 \leq k \leq d-1$, then (3.21) becomes

$$\left|\left|tP \cap \mathbb{Z}^d\right| - p(t)\right| = O\left(a + t^{d-k} + t^{\frac{2(d-1)(d-k-1)}{2d-k-3}+\varepsilon}\left(\frac{t^{d-1}}{a}\right)^{\frac{k-1}{2d-k-3}} + \left(\frac{t^{d-1}}{a}\right)^{\frac{d-2}{k-1}+\varepsilon}\right).$$

Note that the third error term simplifies, thus we have

$$\left|\left|tP \cap \mathbb{Z}^d\right| - p(t)\right| = O\left(a + t^{d-k} + \frac{t^{d-1+\varepsilon}}{a^{\frac{k-1}{2d-k-3}}} + \left(\frac{t^{d-1}}{a}\right)^{\frac{d-2}{k-1}+\varepsilon}\right).$$

The optimal choice of a is when the first and third error terms are equal. Indeed, if we choose a to be

$$a = t^{\frac{(d-1)(2d-k-3)}{2d-4}},$$

then we have

$$a = \frac{t^{d-1}}{a^{\frac{k-1}{2d-k-3}}} = t^{\frac{(d-1)(2d-k-3)}{2d-4}}.$$

Elementary calculation shows that both

$$t^{d-k} \leq t^{\frac{(d-1)(2d-k-3)}{2d-4}}$$

and

$$\left(\frac{t^{d-1}}{a}\right)^{\frac{d-2}{k-1}} \leq t^{\frac{(d-1)(2d-k-3)}{2d-4}}$$

hold for any $2 \leq k \leq d-1$.

□

3.2.4 The orthogonal simplex problem

Given algebraic numbers $a_1, \ldots, a_d > 0$, consider the orthogonal simplex

$$S = \left\{ x \in \mathbb{R}^d : x_1, \ldots, x_d \geq 0, \frac{x_1}{a_1} + \cdots + \frac{x_d}{a_d} \leq 1 \right\}.$$

The simplex tS, where $t > 1$, is a direct generalization of the right triangle studied by Hardy and Littlewood in [11] and [12]. As a tribute, we will reduce the problem of estimating the number of lattice points in tS to the polyhedral sphere problem.

Recall that in the polyhedral sphere problem we considered the polytope

$$P = \left\{ x \in \mathbb{R}^d : \frac{|x_1|}{a_1} + \cdots + \frac{|x_d|}{a_d} \leq 1 \right\}.$$

For any $\sigma \in \{1, -1\}^d$ let

$$S_\sigma = \left\{ x \in \mathbb{R}^d : \sigma_1 x_1, \ldots, \sigma_d x_d \geq 0, \frac{\sigma_1 x_1}{a_1} + \cdots + \frac{\sigma_d x_d}{a_d} \leq 1 \right\}.$$

Then the polytope tP decomposes into the simplices tS_σ, for $\sigma \in \{1, -1\}^d$. For any index set $I \subseteq [d]$, where $[d] = \{1, 2, \ldots, d\}$, also consider the polytope

$$P_I = \left\{ x \in \mathbb{R}^d : \sum_{i \in I} \frac{|x_i|}{a_i} \leq 1, \forall j \in [d] \backslash I : x_j = 0 \right\}.$$

Then we have

$$\sum_{\sigma\in\{1,-1\}^d}\left|tS_\sigma\cap\mathbb{Z}^d\right|=\sum_{I\subseteq[d]}\left|tP_I\cap\mathbb{Z}^d\right|.$$

Indeed, a lattice point $n\in tP\cap\mathbb{Z}^d$ with k zero coordinates is counted 2^k times on both sides. Since the terms of the sum on the left hand side are all equal, we get

$$\left|tS\cap\mathbb{Z}^d\right|=\frac{1}{2^d}\sum_{I\subseteq[d]}\left|tP_I\cap\mathbb{Z}^d\right|. \tag{3.22}$$

This reduces the problem of estimating $\left|tS\cap\mathbb{Z}^d\right|$ to the polyhedral sphere problem. Let us introduce the main term of the estimate as follows.

Definition 3.2: *Let $a_1,\dots,a_d>0$ be real numbers. The function $q=q_{(a_1,\dots,a_d)}$ of the real variable t is defined as*

$$q(t)=q_{(a_1,\dots,a_d)}(t)=\frac{1}{2^d}\sum_{I\subseteq[d]}p_{(a_i:i\in I)}(t),$$

where $p_{(a_i:i\in I)}(t)$ is as in Definition 3.1 for $I\neq\emptyset$, and $p_\emptyset(t)=1$.

The main term $q(t)$ is a polynomial of degree d, and the three highest degree terms are

$$q(t)=\frac{a_1\cdots a_d}{d!}t^d+\frac{a_1\cdots a_d}{2(d-1)!}\sum_{i=1}^{d}\frac{1}{a_i}t^{d-1}+$$

$$\frac{a_1\cdots a_d}{(d-2)!}\left(\frac{1}{12}\sum_{1\le i\le d}\frac{1}{a_i^2}+\frac{1}{4}\sum_{1\le i<j\le d}\frac{1}{a_ia_j}\right)t^{d-2}+\cdots.$$

Indeed, since $p_{(a_i:i\in I)}(t)$ is a polynomial of degree $|I|$ for $I\neq\emptyset$, the only contribution to the three highest degree terms of $q(t)$ come from index sets I of size $d-2,d-1$ and d. Note that the degree d term of $q(t)$ is the Lebesgue measure of tS, while the degree $d-1$ term is one half of the surface area of the orthogonal hyperfaces of tS. The lower degree terms of $q(t)$ do not seem to have a natural geometric interpretation.

Using the reduction (3.22), the following result is a direct corollary of Theorem 3.7.

Corollary 3.8 *Let $2\le k\le d$ be integers, and let $a_1,\dots,a_d>0$ be algebraic. Suppose that any k numbers out of $\frac{1}{a_1},\dots,\frac{1}{a_d}$ are linearly independent over $\mathbb{Q}$. Consider the polytope*

$$S = \left\{ x \in \mathbb{R}^d : x_1, \dots, x_d \geq 0, \frac{x_1}{a_1} + \cdots + \frac{x_d}{a_d} \leq 1 \right\},$$

and the polynomial $q(t)$ as in Definition 3.2. Let $\varepsilon > 0$ and $t > 1$.

(i) *If $k = d$, then*

$$\left| tS \cap \mathbb{Z}^d \right| - q(t) = O\left(t^{\frac{(d-1)(d-2)}{2d-3} + \varepsilon} \right).$$

(ii) *If $2 \leq k \leq d-1$, then*

$$\left| tS \cap \mathbb{Z}^d \right| - q(t) = O\left(t^{\frac{(d-1)(2d-k-3)}{2d-4} + \varepsilon} \right).$$

The implied constants in (i) and (ii) depend only on $a_1, \dots, a_d$ and ε, and are ineffective.

Proof: Using the reduction (3.22) and Definition 3.2 we have, that

$$\left| \left| tS \cap \mathbb{Z}^d \right| - q(t) \right| \leq \frac{1}{2^d} \sum_{I \subseteq [d]} \left| \left| tP_I \cap \mathbb{Z}^d \right| - p_{(a_i : i \in I)} \right|.$$

We can apply Theorem 3.7 to get an upper bound on the terms of the right hand side.

If $k = d$, then for any $I \neq \emptyset$ the numbers $\left\{ \frac{1}{a_i} : i \in I \right\}$ are linearly independent over $\mathbb{Q}$, therefore we can apply Theorem 3.7 (i) in dimension $|I|$. We get an error bound of

$$\left| tP_I \cap \mathbb{Z}^d \right| - p_{(a_i : i \in I)} = O\left(t^{\frac{(|I|-1)(|I|-2)}{2|I|-3} + \varepsilon} \right).$$

Since the function $\frac{(x-1)(x-2)}{2x-3}$ is increasing on $[2, \infty)$, we have that every exponent satisfies

$$\frac{(|I|-1)(|I|-2)}{2|I|-3} \leq \frac{(d-1)(d-2)}{2d-3}.$$

If $2 \leq k \leq d-1$, then we can still apply Theorem 3.7 on $\left| tP_I \cap \mathbb{Z}^d \right| - p_{(a_i : i \in I)}$. If $|I| \geq k+1$ we get

$$\left| tP_I \cap \mathbb{Z}^d \right| - p_{(a_i : i \in I)} = O\left(t^{\frac{(|I|-1)(2|I|-k-3)}{2|I|-4} + \varepsilon} \right).$$

Since the function $\frac{(x-1)(2x-k-3)}{2x-4}$ is increasing on $[k+1, \infty)$, we get

$$\frac{(|I|-1)(2|I|-k-3)}{2|I|-4} \leq \frac{(d-1)(2d-k-3)}{2d-4},$$

therefore

$$\left|tP_I \cap \mathbb{Z}^d\right| - p_{(a_i:i\in I)} = O\left(t^{\frac{(d-1)(2d-k-3)}{2d-4}+\varepsilon}\right).$$

If $0 < |I| \leq k$, then Theorem 3.7 (i) says, that

$$\left|tP_I \cap \mathbb{Z}^d\right| - p_{(a_i:i\in I)} = O\left(t^{\frac{(|I|-1)(|I|-2)}{2|I|-3}+\varepsilon}\right).$$

It is easy to see that the exponent satisfies

$$\frac{(|I|-1)(|I|-2)}{2|I|-3} \leq \frac{(d-1)(d-2)}{2d-3} \leq \frac{(d-1)(2d-k-3)}{2d-4}.$$

Thus for any $I \subseteq [d]$ we have

$$\left|tP_I \cap \mathbb{Z}^d\right| - p_{(a_i:i\in I)} = O\left(t^{\frac{(d-1)(2d-k-3)}{2d-4}+\varepsilon}\right).$$

□

3.3 Effective bound on the discrepancy

In this section we try to find an effective error bound on the discrepancy

$$\left|tP \cap \mathbb{Z}^d\right| - \lambda(P)t^d$$

for a special class of polytopes. As observed in section 2.3, we have an effective error term for the Poisson summation formula only in case every hyperface of the polytope has a normal vector with coordinates from a real quadratic field, as in Theorem 2.7.

In the special case

$$P = \left\{x \in \mathbb{R}^d : \frac{|x_1|}{a_1} + \cdots + \frac{|x_d|}{a_d} \leq 1\right\},$$

which was studied in section 3.2, this means we have to assume $a_1, \ldots, a_d \in \mathbb{Q}\left(\sqrt{D}\right)$ for some square-free integer $D > 1$ to obtain an effective bound on the discrepancy.

Theorem 3.7 in the case $k = 2$ gives an error bound of $O\left(t^{\frac{(d-1)(2d-5)}{2d-4}+\varepsilon}\right)$. It would not be difficult to rewrite the proof of Theorem 3.7 in this special case to prove the same error bound with an effective implied constant; the factor t^{ε} might even be replaced by a polylogharithmic factor of t. Note, however, that the exponent $\frac{(d-1)(2d-5)}{2d-4}$ for large values of d is roughly $d - \frac{3}{2}$, which is only a modest improvement to the trivial error bound $O\left(t^{d-1}\right)$ of Corollary 1.2. To avoid repetitions, we will prove a slightly weaker result under much more general conditions.

Theorem 3.8 *Let P be a polytope in $\mathbb{R}^d$ and let $D > 1$ be a square-free integer. Suppose that every coordinate of every vertex of P is in $\mathbb{Q}\left(\sqrt{D}\right)$. Suppose also, that every hyperface of P has a normal vector of the form $p + \sqrt{D}q$, where $p, q \in \mathbb{Z}^d$ are linearly independent. For any $t > 1$ we have*

$$\left|tP \cap \mathbb{Z}^d\right| - \lambda(P)t^d = O\left(t^{d-\frac{8}{7}} \log^{\frac{4}{7}}(t+1)\right).$$

The implied constant depends only on P, and is effective.

By the normal vector of a hyperface we simply mean a nonzero vector of arbitrary length orthogonal to the hyperface. The exponent $d - \frac{8}{7}$ of t in the error bound makes the result only a slight improvement to the trivial error bound $O\left(t^{d-1}\right)$ of Corollary 1.2. The reason why Theorem 3.8 is still relevant is that the implied constant is effective. In particular this means that we cannot use Schmidt's theorem on simultaneous Diophantine approximation, or any other ineffective method in the proof.

In a sense the class of polytopes the vertices of which have coordinates in a given real quadratic field is the simplest class of non lattice polytopes. The significance of Theorem 3.8 is that it shows that the order of magnitude of the discrepancy of polytopes in this class exhibits a "gap". Indeed, consider a polytope P in $\mathbb{R}^d$ the vertices of which lie in $\mathbb{Q}\left(\sqrt{D}\right)^d$, and suppose that the origin is in the interior of P. Since the normal vectors of the hyperfaces of P can be expressed in terms of the vertices using linear algebra, the normal vectors also lie in $\mathbb{Q}\left(\sqrt{D}\right)^d$. After scaling, the normal vectors can be written in the form $p+\sqrt{D}q$ for some $p, q \in \mathbb{Z}^d$. If p, q are linearly dependent for some hyperface, then the magnified polytope tP contains a $d-1$ dimensional sublattice of $\mathbb{Z}^d$ for special values of t. This means that the discrepancy $\left|tP \cap \mathbb{Z}^d\right| - \lambda(P)t^d$ has jumps of

size constant times t^{d-1}, therefore the trivial discrepancy bound $O\left(t^{d-1}\right)$ of Corollary 1.2 is best possible. If, on the other hand, p, q are linearly independent for every hyperface of P, then Theorem 3.8 applies and the discrepancy is $O\left(t^{d-\frac{8}{7}} \log^{\frac{4}{7}}(t+1)\right)$. Thus the order of magnitude of the discrepancy exhibits a gap between $O\left(t^{d-1}\right)$ and $O\left(t^{d-\frac{8}{7}} \log^{\frac{4}{7}}(t+1)\right)$.

Proof of Theorem 3.8: The proof consists of two steps. In the first step we show that it is enough to prove the theorem for simplices. The second step is to prove the theorem in the special case when P is a simplex.

We first prove, that any polytope P satisfying the conditions of the theorem can be decomposed into simplices satisfying the same conditions. To see this fact, let us first consider a triangulation of the hyperfaces of P, in which every coordinate of every vertex is in $\mathbb{Q}\left(\sqrt{D}\right)$. Next, choose a point $c \in \mathbb{Q}\left(\sqrt{D}\right)^d$ in the interior of P, and let $r > 0$ such that the open ball centered at c of radius r is a subset of P. Let $x, y \in \mathbb{Q}^d$ such that $\left|x + \sqrt{D}y\right| < r$, and add the vertex $v_0 = c + x + \sqrt{D}y$. The convex hulls of v_0 and the $d-1$ dimensional simplices triangulating the hyperfaces triangulate P. Our goal is to show that x and y can be chosen in such a way, that the simplices in this triangulation satisfy the conditions of the theorem.

Consider thus a $d-1$ dimensional simplex with vertices $v_1, \ldots, v_d$ in the triangulation of one of the hyperfaces of P, and the convex hull S of $v_0, v_1, \ldots, v_d$. The hyperface of S with vertices $v_1, \ldots, v_d$ automatically satisfies the conditions of the theorem. Consider now a hyperface H of S which contains v_0. For the sake of simplicity we will work with the hyperface containing $v_0, v_1, \ldots, v_{d-1}$, the others being similar. It will be enough to show that x and y can be chosen in such a way, that there does not exist a nonzero rational vector orthogonal to H. Indeed, since H clearly has a normal vector with coordinates in $\mathbb{Q}\left(\sqrt{D}\right)$, after rescaling, it can be written in the form $p + \sqrt{D}q$ with $p, q \in \mathbb{Z}^d$. If p, q were linearly dependent, then $p + \sqrt{D}q$ could be rescaled again to a nonzero rational vector orthogonal to H. Let (C) denote the condition that there does not exist a nonzero rational vector orthogonal to H.

Let the rational linear subspace $V \subseteq \mathbb{Q}^d$ be defined as

$$V = \left\{ u \in \mathbb{Q}^d : \langle v_1, u \rangle = \cdots = \langle v_{d-1}, u \rangle \right\}.$$

Note that any $u \in V$ such that $\langle v_1 - v_d, u \rangle = 0$, is orthogonal to a hyperface of P. Since the normal vectors of the hyperfaces of P are not in $\mathbb{Q}^d$, this implies that every such u has to be the zero vector. But $v_1 - v_d$ is of the form $a + \sqrt{D}b$ for some $a, b \in \mathbb{Q}^d$, hence the equation $\langle v_1 - v_d, u \rangle = 0$ can be written as a system of two linear equations with rational coefficients. Since a system of two linear equations in V has only the trivial solution $u = 0$, the dimension of V over $\mathbb{Q}$ is at most 2.

Note that (C) is equivalent to the fact that the only $u \in V$ for which

$$\langle v_0 - v_1, u \rangle = \langle c + x + \sqrt{D}y - v_1, u \rangle = 0,$$

is $u = 0$. We claim that there exists a polynomial $p(x, y) = p_V(x, y)$ of the $2d$ variables $x, y \in \mathbb{Q}^d$ depending on the subspace V, which is not the constant zero polynomial, such that (C) is satisfied if and only if $p(x, y) \neq 0$. If $V = \{0\}$, then we can choose $p(x, y) = 1$, as (C) is automatically satisfied.

If V has dimension 1 over $\mathbb{Q}$, then (C) holds if and only if the vector $c+x+\sqrt{D}y-v_1$ is not orthogonal to the basis vector of V. Therefore we can choose $p(x, y)$ to be the scalar product of $c + x + \sqrt{D}y - v_1$ and the basis vector of V.

If V has dimension 2 over $\mathbb{Q}$, then let $n_1, n_2 \in \mathbb{Q}^d$ be a basis of V. Let us write the vector $c - v_1$ in the form $c - v_1 = a + \sqrt{D}b$, where $a, b \in \mathbb{Q}^d$. Then the equation $\langle c + x + \sqrt{D}y - v_1, u \rangle = 0$ is equivalent to the system

$$\langle a + x, u \rangle = 0,$$

$$\langle b + y, u \rangle = 0.$$

Expressing everything in the basis n_1, n_2 of V, we get that (C) is equivalent to the fact, that

$$\det \begin{pmatrix} \langle a + x, n_1 \rangle & \langle a + x, n_2 \rangle \\ \langle b + y, n_1 \rangle & \langle b + y, n_2 \rangle \end{pmatrix} \neq 0.$$

This means that we can choose $p(x, y)$ to be this determinant. We have to check, however, that this determinant, as a function of x and y, is not the constant zero polynomial. To see this, it is enough to find a particular value for x and y which result in a nonzero value for the determinant. Substituting $x = n_1 - a$ and $y = n_2 - b$, we get that the determinant is

$$|n_1|^2 |n_2|^2 - |\langle n_1, n_2 \rangle|^2 .$$

If this determinant were zero, we would get equality in the Cauchy–Schwarz inequality applied on n_1, n_2. This is impossible, since equality in the Cauchy–Schwarz inequality holds only for parallel vectors, but n_1, n_2 is a basis of V. Therefore the polynomial

$$p(x, y) = \det \begin{pmatrix} \langle a + x, n_1 \rangle & \langle a + x, n_2 \rangle \\ \langle b + y, n_1 \rangle & \langle b + y, n_2 \rangle \end{pmatrix}$$

is not constant zero.

This way we obtain finitely many polynomials of degree 1 or 2 of the variables x, y, such that by adding the vertex $v_0 = c + x + \sqrt{D}y$, every simplex in the triangulation of P satisfies the conditions of the theorem if and only if none of the polynomials evaluated at x, y are zero. Finding rational values for x, y in a small neighborhood of the origin to ensure $\left| x + \sqrt{D}y \right| < r$, where a finite family of polynomials all take nonzero values is easy. For example, the product of all the polynomials has to take a nonzero value on a large rectangular grid within a small neighborhood of the origin.

We now claim that it is enough to prove the theorem in the special case when P is a simplex. Indeed, suppose the theorem is true for simplices, and consider a polytope P satisfying the conditions of the theorem. Consider a triangulation of P into simplices $S_1, \ldots, S_k$ satisfying the conditions of the theorem. Since the normal vectors of the hyperfaces of P and $S_1, \ldots, S_k$ are not in $\mathbb{Q}^d$, the hyperfaces of tP and $tS_1, \ldots, tS_k$ cannot contain a $d-1$ dimensional sublattice of $\mathbb{Z}^d$ for any $t > 1$. Thus the hyperfaces of tP and $tS_1, \ldots, tS_k$ contain $O\left(t^{d-2}\right)$ lattice points, hence

$$\left| tP \cap \mathbb{Z}^d \right| - \lambda(P) t^d = \sum_{i=1}^{k} \left(\left| tS_i \cap \mathbb{Z}^d \right| - \lambda(S_i) t^d \right) + O\left(t^{d-2}\right) .$$

Applying the theorem on the simplices $S_1, \dots, S_k$ shows that the theorem holds for P. Finally note that the triangulation of P, including finding the extra vertex v_0, was done using an effective algorithm, hence the implied constant in the theorem is effective.

Let us now prove the theorem in the special case, when $P = S$ is a simplex with vertices $v_1, \dots, v_{d+1}$. Fix an arbitrary $t > 1$ and an arbitrary integer $N > 1$. Applying Theorem 2.7 we get

$$\left|tS \cap \mathbb{Z}^d\right| = C(tS, N) + O\left(t^{d-2} + t^{d-1}\sqrt{\frac{\log N}{N}}\right), \tag{3.23}$$

where $C(tS, N)$ is as in Definition 2.3. We will use the representation in Theorem 3.1 to estimate the Fourier coefficients, and thus $C(tS, N)$. According to Theorem 3.1, for any lattice point $m \in [0, N-1]^d$ we have

$$\hat{\chi}_{tS}(m) = \frac{(-1)^d d!}{(2\pi i)^{d+1}}\lambda(S)\int_{|z|=R} \frac{e^{-2\pi i z t}}{(z - \langle v_1, m\rangle)\cdots(z - \langle v_{d+1}, m\rangle)}\,\mathrm{d}z,$$

where $R > 0$ is a fixed constant such that $R > |\langle v_j, m\rangle|$ for any $1 \le j \le d+1$ and any $m \in [0, N-1]^d$. The only main term we will isolate is $\hat{\chi}_{tS}(0) = \lambda(S)t^d$. We will use the residue theorem to evaluate the complex line integral. The problem is that for a given lattice point m several singularities $\langle v_j, m\rangle$ might coincide resulting in a high order pole. To handle such cases, for any lattice point $m \neq 0$ let us define the equivalence relation $\sim$ on the index set $[d+1]$ as

$$i \sim j \Longleftrightarrow \langle v_i, m\rangle = \langle v_j, m\rangle.$$

Let $\mathcal{P}(m)$ denote the partition of $[d+1]$ defined by this equivalence relation. For the sake of simplicity let us introduce the notation

$$R(J, m) = \mathrm{Res}_{\langle v_j, m\rangle} \frac{e^{-2\pi i z t}}{(z - \langle v_1, m\rangle)\cdots(z - \langle v_{d+1}, m\rangle)}$$

for any lattice point $m \neq 0$ and $j \in J \in \mathcal{P}(m)$. Then the residue theorem can be written in the form

$$\hat{\chi}_{tS}(m) = \frac{(-1)^d d!}{(2\pi i)^d}\lambda(S)\sum_{J \in \mathcal{P}(m)} R(J, m). \tag{3.24}$$

Note that for any $m \neq 0$ and $J \in \mathcal{P}(m)$ we have $|J| \leq d-1$. Indeed, the nonzero rational vector m is orthogonal to the face of S containing $\{v_j : j \in J\}$, which therefore cannot be a hyperface of S, or S itself. We now find an upper bound to $|R(J,m)|$ for any $m \neq 0$ and $J \in \mathcal{P}(m)$.

To find an upper bound to $|R(J,m)|$, note that we may assume that J is of the form $J = [k] = \{1, 2, \ldots, k\}$ for some $1 \leq k \leq d-1$, the other index sets resulting in similar residues. In other words, we will assume for the sake of simplicity that the lattice point $m \neq 0$ satisfies

$$\langle v_1, m\rangle = \cdots = \langle v_k, m\rangle,$$

but $\langle v_j, m\rangle \neq \langle v_1, m\rangle$ for every $k+1 \leq j \leq d+1$. We want to estimate the residue

$$R([k], m) = \mathrm{Res}_{\langle v_1, m\rangle} \frac{e^{-2\pi i z t}}{(z - \langle v_1, m\rangle)^k \, (z - \langle v_{k+1}, m\rangle) \cdots (z - \langle v_{d+1}, m\rangle)} =$$

$$\mathrm{Res}_0 \frac{e^{-2\pi i z t} e^{-2\pi i \langle v_1, m\rangle t}}{z^k \, (z - \langle v_{k+1} - v_1, m\rangle) \cdots (z - \langle v_{d+1} - v_1, m\rangle)} =$$

$$\mathrm{Res}_0 \frac{1}{z^{d+1}} e^{-2\pi i z t} e^{-2\pi i \langle v_1, m\rangle t} \frac{z}{z - \langle v_{k+1} - v_1, m\rangle} \cdots \frac{z}{z - \langle v_{d+1} - v_1, m\rangle}.$$

Let us use the Taylor expansions

$$e^{-2\pi i z t} = \sum_{n=0}^{\infty} \frac{(-2\pi i t)^n}{n!} z^n,$$

$$\frac{z}{z - \langle v_j - v_1, m\rangle} = \sum_{i_j=1}^{\infty} \frac{-1}{\langle v_j - v_1, m\rangle^{i_j}} z^{i_j}$$

for $k+1 \leq j \leq d+1$, which hold in an open neighborhood of $z = 0$. We get, that $R([k], m)$ equals the coefficient of z^d in the power series

$$e^{-2\pi i \langle v_1, m\rangle t} \left(\sum_{n=0}^{\infty} \frac{(-2\pi i t)^n}{n!} z^n \right) \left(\sum_{i_{k+1}=1}^{\infty} \frac{-1}{\langle v_{k+1} - v_1, m\rangle^{i_{k+1}}} z^{i_{k+1}} \right) \cdots$$

$$\left(\sum_{i_{d+1}=1}^{\infty} \frac{-1}{\langle v_{d+1} - v_1, m\rangle^{i_{d+1}}} z^{i_{d+1}} \right).$$

The only terms of the first power series which contribute to the coefficient of z^d are indexed by $0 \leq n \leq k-1$. Therefore to find the coefficient of z^d, we can first fix an integer $0 \leq n \leq k-1$, then consider all integers $i_{k+1}, \ldots, i_{d+1} \geq 1$, such that $i_{k+1} + \cdots + i_{d+1} = d - n$. Altogether we get, that

$$|R([k], m)| = O\left(\sum_{n=0}^{k-1} t^n \sum_{\substack{i_{k+1},\ldots,i_{d+1} \geq 1 \\ i_{k+1}+\cdots+i_{d+1}=d-n}} \frac{1}{|\langle v_{k+1} - v_1, m\rangle|^{i_{k+1}} \cdots |\langle v_{d+1} - v_1, m\rangle|^{i_{d+1}}}\right). \tag{3.25}$$

For any $k+1 \leq j \leq d+1$ we can write $v_j - v_1$ in the form $v_j - v_1 = \frac{a_j}{Q} + \sqrt{D}\frac{b_j}{Q}$ for some $a_j, b_j \in \mathbb{Z}^d$ and $Q \in \mathbb{N}$. Therefore we have

$$|\langle v_j - v_1, m\rangle| = \frac{1}{Q}\left|\langle a_j, m\rangle + \sqrt{D}\langle b_j, m\rangle\right|.$$

Here both $\langle a_j, m\rangle$ and $\langle b_j, m\rangle$ are integers. Estimating this factor is thus equivalent to the classical problem of approximating $\sqrt{D}$ with rational numbers. We can use the standard trick of multiplying by the conjugate to obtain

$$|\langle v_j - v_1, m\rangle| = \frac{\left|\langle a_j, m\rangle^2 - D\langle b_j, m\rangle^2\right|}{Q\left|\langle a_j, m\rangle - \sqrt{D}\langle b_j, m\rangle\right|}.$$

The numerator is the absolute value of a nonzero integer. Applying the triangle inequality and the Cauchy–Schwarz inequality on the denominator, we get that

$$|\langle v_j - v_1, m\rangle| = \Omega\left(\frac{1}{|m|}\right). \tag{3.26}$$

On the other hand, we also have

$$|\langle v_j - v_1, m\rangle| = \Omega\left(\left\|\langle b_j, m\rangle\sqrt{D}\right\|\right). \tag{3.27}$$

Unfortunately (3.27) is useless if $\langle b_j, m\rangle = 0$. But if $\langle b_j, m\rangle = 0$ then

$$|\langle v_j - v_1, m\rangle| \geq \frac{1}{Q} = \Omega(1). \tag{3.28}$$

Since we are trying to prove an effective bound, we cannot use simultaneous Diophantine approximation. Therefore let us apply (3.26) on all but one factor of (3.25), and use (3.27) or (3.28) on one factor. Since we have $i_{k+1} + \cdots + i_{d+1} = d - n$ we get

$$|R([k], m)| = O\left(\sum_{n=0}^{k-1} t^n \frac{|m|^{d-n-1}}{\left\|\langle b_{k+1}, m\rangle\sqrt{D}\right\|}\right),$$

if $\langle b_{k+1}, m\rangle \neq 0$, and

$$|R([k], m)| = O\left(\sum_{n=0}^{k-1} t^n |m|^{d-n-1}\right),$$

if $\langle b_{k+1}, m\rangle = 0$. Clearly the same bounds hold for any other index set $J \subseteq [d+1]$ of size k. Altogether we found, that for any index set $J \subseteq [d+1]$ of size k there exists an integral vector $b_J \in \mathbb{Z}^d$ depending only on the simplex S such that

$$|R(J, m)| = O\left(\sum_{n=0}^{k-1} t^n \frac{|m|^{d-n-1}}{\left\|\langle b_J, m\rangle\sqrt{D}\right\|}\right), \tag{3.29}$$

if $\langle b_J, m\rangle \neq 0$, and

$$|R(J, m)| = O\left(\sum_{n=0}^{k-1} t^n |m|^{d-n-1}\right), \tag{3.30}$$

if $\langle b_J, m\rangle = 0$.

Fix integers $0 \leq M_1, \ldots, M_d \leq N-1$ and let $M = (M_1, \ldots, M_d)$. Summing (3.24) over lattice points m in the rectangle $[-M_1, M_1] \times \cdots \times [-M_d, M_d]$, and isolating the main term $\hat{\chi}_{tS}(0) = \lambda(S)t^d$ we get

$$S(tS, M) = \lambda(S)t^d + \frac{(-1)^d d!}{(2\pi i)^d}\lambda(S) \sum_{m \in [-M_1, M_1]\times\cdots\times[-M_d, M_d]\setminus\{0\}} \sum_{J \in \mathcal{P}(m)} R(J, m),$$

where $S(tS, M)$ is as in Definition 2.2. To bound the double sum, it will be more convenient to switch the order of summation to obtain

$$S(tS, M) = \lambda(S)t^d + \frac{(-1)^d d!}{(2\pi i)^d}\lambda(S) \sum_{J \subseteq [d+1]} \sum_{\substack{m \in [-M_1, M_1]\times\cdots\times[-M_d, M_d]\setminus\{0\} \\ J \in \mathcal{P}(m)}} R(J, m). \tag{3.31}$$

Let us thus fix an index set $J \subseteq [d+1]$ of size k. Recall that $1 \leq k \leq d-1$. Consider the inner sum indexed by J. From the definition of $\mathcal{P}(m)$ we see that every lattice point m which shows up in the inner sum is contained in the linear subspace

$$V_J = \left\{ x \in \mathbb{R}^d : \forall i,j \in J \quad \langle v_i, x \rangle = \langle v_j, x \rangle \right\}.$$

Note that V_J can be described by $k-1$ linearly independent linear equations, therefore it has dimension $d-k+1$. Therefore the inner sum in (3.31) satisfies

$$\left| \sum_{\substack{m \in [-M_1, M_1] \times \cdots \times [-M_d, M_d] \setminus \{0\} \\ J \in \mathcal{P}(m)}} R(J,m) \right| \leq \sum_{\substack{m \in V_J \cap \mathbb{Z}^d \\ 0 < |m| < \sqrt{d}N}} |R(J,m)|. \tag{3.32}$$

We thus have to sum the bounds (3.29) or (3.30) over the integral points in a ball of radius $\sqrt{d}N$ in a linear subspace of dimension $d-k+1$. We have to distinguish between two cases: either b_J is orthogonal to the linear subspace V_J, or it is not. If b_J is orthogonal to V_J, then $\langle b_J, m \rangle = 0$ for every $m \in V_J \cap \mathbb{Z}^d$, therefore we can use (3.30) on every term to get

$$\sum_{\substack{m \in V_J \cap \mathbb{Z}^d \\ 0 < |m| < \sqrt{d}N}} |R(J,m)| = O\left(\sum_{n=0}^{k-1} t^n N^{d-n-1} N^{d-k+1} \right) = O\left(\sum_{n=0}^{k-1} t^n N^{2d-k-n} \right).$$

Suppose now, that b_J is not orthogonal to V_J. The scalar product $\langle b_J, m \rangle$ takes integral values in an interval of length cN centered at zero, where $c = |b_J|\sqrt{d} = O(1)$. Moreover every value is attained at most $O\left(N^{d-k}\right)$ times. Indeed, the set of $m \in V_J \cap \mathbb{Z}^d$ for which $\langle b_J, m \rangle = a$ for a fixed integer a is contained in an affine hyperplane of V_J. Therefore (3.29) and (3.30) imply that

$$\sum_{\substack{m \in V_J \cap \mathbb{Z}^d \\ 0 < |m| < \sqrt{d}N}} |R(J,m)| = O\left(\sum_{n=0}^{k-1} t^n N^{d-n-1} N^{d-k} \left(1 + \sum_{a=1}^{cN} \frac{1}{\left\| a\sqrt{D} \right\|} \right) \right).$$

The sum over a is a well-known Diophantine sum. Using an argument based on the pigeonhole principle it is easy to see that

$$\sum_{a=1}^{cN} \frac{1}{\left\| a\sqrt{D} \right\|} = O\left(N \log N \right).$$

For a proof of this fact see Proposition 4.2 (i) in chapter 4. Thus if b_J is not orthogonal to V_J, we have

$$\sum_{\substack{m\in V_J\cap\mathbb{Z}^d\\ 0<|m|<\sqrt{d}N}} |R(J,m)| = O\left(\sum_{n=0}^{k-1} t^n N^{2d-k-n}\log N\right).$$

Applying this bound in (3.32) for every index set $J\subseteq[d+1]$ of size $1\le k\le d-1$, (3.31) yields

$$S(tS,M) = \lambda(S)t^d + O\left(\sum_{k=1}^{d-1}\sum_{n=0}^{k-1} t^n N^{2d-k-n}\log N\right).$$

By taking the average over integral points $M\in[0,N-1]^d$ we obtain

$$C(tS,N) = \lambda(S)t^d + O\left(\sum_{k=1}^{d-1}\sum_{n=0}^{k-1} t^n N^{2d-k-n}\log N\right), \tag{3.33}$$

where $C(tS,N)$ is as in Definition 2.3.

Finally (3.23) and (3.33) yield

$$\left|tS\cap\mathbb{Z}^d\right| = \lambda(S)t^d + O\left(t^{d-2} + t^{d-1}\sqrt{\frac{\log N}{N}} + \sum_{k=1}^{d-1}\sum_{n=0}^{k-1} t^n N^{2d-k-n}\log N\right).$$

The optimal choice for the integer $N>1$ is the integer closest to $t^{\frac{2}{7}}\log^{-\frac{1}{7}}(t+1)$. Note that for this choice we have $\frac{t}{N}=\Omega(1)$, therefore the largest error term in

$$\sum_{n=0}^{k-1} t^n N^{2d-k-n}\log N$$

is when $n=k-1$. Thus the last error term simplifies to

$$\sum_{k=1}^{d-1} t^{k-1}N^{2d-2k+1}\log N.$$

We also have $\frac{t}{N^2}=\Omega(1)$, hence the largest error term is simply $t^{d-2}N^3\log N$. For our choice of N the second and third error terms $t^{d-1}\sqrt{\frac{\log N}{N}}$ and $t^{d-2}N^3\log N$ have the same order of magnitude $t^{d-\frac{8}{7}}\log^{\frac{4}{7}}(t+1)$.

□

Chapter 4

Lattice point counting problems on the plane

4.1 Continued fractions

This section is devoted to recalling some basic facts from the theory of continued fractions. We follow the notation of the book [4] by Cassels. There are no new results proved.

Let us fix the terminology and the notation first. The continued fraction representation of an irrational real number α is denoted by

$$\alpha = [a_0; a_1, a_2, \ldots],$$

where a_0 is an integer, and $a_1, a_2, \ldots$ are positive integers, called the partial quotients of α. For integers $k \geq 1$ the fractions

$$\frac{p_k}{q_k} = [a_0; a_1, a_2, \ldots, a_{k-1}] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\cdots + \cfrac{1}{a_{k-1}}}}}$$

are called the convergents to α. The distance of a real number x from the nearest integer is denoted by $\|x\|$.

The properties of continued fractions we shall use later are listed in the following proposition. Since these properties can be found in any introductory textbook on continued fractions, e.g. in Chapter I of [4], the proof will be omitted.

Proposition 4.1 *With the notation above, for any irrational real α:*

(i) *For any $k \geq 1$ and $0 < m < q_{k+1}$ we have $\|q_k\alpha\| \leq \|m\alpha\|$.*

(ii) *For any $k \geq 2$ we have $\frac{1}{q_{k+1}+q_k} < \|q_k\alpha\| < \frac{1}{q_{k+1}}$. If $a_1 > 1$, the same is true for $k = 1$.*

(iii) *The convergent denominators satisfy the recurrence* $q_{k+1} = a_k q_k + q_{k-1}$ *with initial conditions* $q_1 = 1$ *and* $q_2 = a_1$.

(iv) *For any* $k \geq 1$ *the numbers* p_k *and* q_k *are relatively prime.*

(v) *For any* $k \geq 2$ *we have* $p_k q_{k-1} - q_k p_{k-1} = (-1)^k$.

(vi) *For any* $k \geq 1$ *we have* $\operatorname{sign}(q_k \alpha - p_k) = (-1)^{k+1}$.

Properties (i) and (ii) describe the connection between continued fractions and Diophantine approximation. Consider the sequence $\|m\alpha\|$. Property (i) says that the terms of this sequence which are as small as possible in terms of m are precisely the terms for which the index m equals a convergent denominator of α. In addition, property (ii) quantifies how small those terms are. In the lattice point counting problem to be studied in the following section, we shall need to know more about the distribution of the values of $\|m\alpha\|$. In particular, we will be interested in sums of the form

$$\sum_{m=1}^{M} \frac{1}{m^a \|m\alpha\|^b}$$

for some exponents a, b. As noted before, the main contribution of these sums come from the terms where m equals a convergent denominator q_k of α. A simple application of the pigeonhole principle allows us to bound the contribution of the terms as m runs between two consecutive convergent denominators. Even though this method is well-known, for the sake of completeness we shall formulate it as a proposition and include a proof as follows.

Proposition 4.2 *Let* $\alpha = [a_0; a_1, a_2, \ldots]$ *be the continued fraction representation of an irrational real number* α, *and let* $\frac{p_k}{q_k} = [a_0; a_1, a_2, \ldots, a_{k-1}]$ *denote its convergents.*

(i) *For any* $k \geq 2$ *we have*

$$\sum_{0<m<q_k} \frac{1}{\|m\alpha\|} \leq 8 q_k \log_2 (2q_k).$$

(ii) *For any* $k \geq 2$ *and* $b \geq 2$ *we have*

$$\sum_{0<m<q_k} \frac{1}{\|m\alpha\|^b} \leq 8 (2q_k)^b.$$

Proof: From Proposition 4.1 (i) and (ii) we know that for every $0 < m < q_k$ we have

$$\|m\alpha\| \geq \|q_{k-1}\alpha\| > \frac{1}{q_k + q_{k-1}} \geq \frac{1}{2q_k}.$$

For any integer $\ell \geq 0$ consider the set

$$A_{k,\ell} = \left\{ 0 < m < q_k : 2^\ell \frac{1}{2q_k} \leq \|m\alpha\| < 2^{\ell+1} \frac{1}{2q_k} \right\}.$$

First note, that if $2^\ell \frac{1}{2q_k} \geq \frac{1}{2}$, in other words if $\ell \geq \log_2 q_k$, then $A_{k,\ell} = \emptyset$. Therefore

$$\bigcup_{0 \leq \ell < \log_2 q_k} A_{k,\ell}$$

is a partition of the interval of integers $(0, q_k)$.

Now we find an upper bound for the cardinality of $A_{k,\ell}$. For every $m \in A_{k,\ell}$ consider the point in $\left[-\frac{1}{2}, \frac{1}{2}\right)$ equivalent to $m\alpha$ modulo 1. These points all lie in the open interval $\left(-2^{\ell+1}\frac{1}{2q_k}, 2^{\ell+1}\frac{1}{2q_k}\right)$. On the other hand, the distance of any two points is larger, than $\frac{1}{2q_k}$. Indeed, if $0 < m < m' < q_k$ then $0 < m' - m < q_k$ and thus

$$\left\|(m' - m)\alpha\right\| > \frac{1}{2q_k}.$$

Therefore, by the pigeonhole principle we have $|A_{k,\ell}| \leq 2^{\ell+2}$.

To see (i) consider

$$\sum_{0<m<q_k} \frac{1}{\|m\alpha\|} = \sum_{0 \leq \ell < \log_2 q_k} \sum_{m \in A_{k,\ell}} \frac{1}{\|m\alpha\|} \leq \sum_{0 \leq \ell < \log_2 q_k} \frac{2q_k}{2^\ell} \cdot 2^{\ell+2} \leq 8q_k \left(\log_2 q_k + 1\right).$$

To see (ii) consider

$$\sum_{0<m<q_k} \frac{1}{\|m\alpha\|^b} = \sum_{0 \leq \ell < \log_2 q_k} \sum_{m \in A_{k,\ell}} \frac{1}{\|m\alpha\|^b} \leq \sum_{\ell=0}^{\infty} \left(\frac{2q_k}{2^\ell}\right)^b \cdot 2^{\ell+2} \leq$$

$$(2q_k)^b \sum_{\ell=0}^{\infty} \frac{4}{2^\ell} = 8(2q_k)^b.$$

□

4.2 Lattice points in a right triangle

We have seen in chapter 3 that the lattice point counting problem for polytopes in $\mathbb{R}^d$ is related to certain simultaneous Diophantine approximation problems for $d-1$ irrational reals. To study the lattice point counting problem for polygons in the Euclidean plane, we therefore only need to consider the classical Diophantine approximation problem for a single irrational number. It is not difficult to come up with the intuition, that the irrational numbers we have to consider are the slopes of the sides of our polygon. Since every polygon in the plane can be decomposed into right triangles with axis parallel legs, the lattice point counting problem for an arbitrary polygon can be reduced to estimating the number of lattice points in such triangles.

For the sake of simplicity we will consider the closed right triangle

$$S = \left\{(x, y) \in \mathbb{R}^2 : x, y \geq 0, \alpha x + y \leq \alpha\right\}$$

with vertices $(0,0), (1,0), (0,\alpha)$, where $\alpha > 0$ is irrational. We want to estimate $\left|tS \cap \mathbb{Z}^2\right|$, where $t > 0$ is real. The same exact problem has been studied by Hardy and Littlewood in [11] and [12], and more recently by Beck in [1].

We identify the main term of $\left|tS \cap \mathbb{Z}^2\right|$ as

$$g(t) = \frac{\alpha}{2}t^2 + \frac{\alpha+1}{2}t + \frac{(\alpha\{t\}+1)(1-\{t\})}{2}.$$

The first term is the area, while the second term is one half of the total length of the legs of tS. Note that the last term, which is a bounded and periodic function of t, makes $g(t)$ different from the main term $q(t)$ we used in subsection 3.2.4 in the orthogonal simplex problem.

The difference $\left|tS \cap \mathbb{Z}^2\right| - g(t)$ is considered a random fluctuation. We studied the expected value and the standard deviation of this random fluctuation on the interval $[0, T]$, where $T \geq 1$ is a real number. In other words, we consider

$$\frac{1}{T}\int_0^T \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right) \mathrm{d}t \tag{4.1}$$

and

$$\sqrt{\frac{1}{T}\int_0^T \left(|tS\cap\mathbb{Z}^2| - g(t)\right)^2 \,\mathrm{d}t}. \tag{4.2}$$

The main difference between this problem and the high dimensional lattice point counting problems studied in chapter 3, is that the continued fraction representation of α provides an effective solution of the related Diophantine approximation problem. Therefore instead of assuming that α is algebraic, we will express our results in terms of the partial quotients of α. As a result we will be able to find the standard deviation (4.2) up to an explicit error term. Note that we do not have a similar formula for higher dimensional lattice point counting problems.

This shows a sharp contrast between lattice point counting problems in polytopes and smooth convex bodies. Let us illustrate the difference by recalling that the sphere problem is completely solved in high dimensions, whereas the centuries-old Gauss circle problem is wide open to this date.

In section 4.5 of [1] Beck states without a proof that for an arbitrary irrational α the expected value (4.1) is negligible compared to the standard deviation (4.2), and that the standard deviation (4.2) is the sum of

$$\sqrt{\sum_{m=1}^{\lfloor T\rfloor} \frac{1}{4\pi^2 m^2 \sin^2(m\alpha\pi)}} \tag{4.3}$$

and a negligible error term. Since the order of magnitude of these negligible error terms is not specified, and no proof is given in [1], for the sake of completeness we formulate and prove these claims as a proposition for a wide range of irrationals as follows.

Proposition 4.3 *Let $\alpha > 0$ be irrational, and let S be the closed right triangle with vertices $(0,0),(1,0),(0,\alpha)$. Let*

$$g(t) = \frac{\alpha}{2}t^2 + \frac{\alpha+1}{2}t + \frac{(\alpha\{t\}+1)(1-\{t\})}{2}.$$

(i) *For any real $T \geq 1$ we have*

$$\left|\frac{1}{T}\int_0^T \left(\left|tS\cap\mathbb{Z}^2\right| - g(t)\right) \,\mathrm{d}t\right| \leq \frac{3}{8\alpha}.$$

(ii) *Suppose that the continued fraction representation* $\alpha = [a_0; a_1, a_2, \ldots]$ *satisfies* $a_k = O(k^d)$ *for some real number* $d \geq 0$. *Then for any real* $T \geq 3$ *we have*

$$\frac{1}{T}\int_0^T \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right)^2 \,\mathrm{d}t = \sum_{m=1}^{\lfloor T \rfloor} \frac{1}{4\pi^2 m^2 \sin^2(m\alpha\pi)} + O\left(\log^{2d} T \log\log T\right).$$

The implied constant depends only on α *and is effective.*

Proof: Suppose that the magnifying factor $0 \leq t \leq T$ has integer part $\lfloor t \rfloor = n$. The number of lattice points in tS on the line $x = k$ is

$$\lfloor (t-k)\alpha \rfloor + 1 = \lfloor (n + \{t\} - k)\alpha \rfloor + 1,$$

where $0 \leq k \leq n$. Therefore

$$\left|tS \cap \mathbb{Z}^2\right| = \sum_{k=0}^{n} \left(\lfloor (n + \{t\} - k)\alpha \rfloor + 1\right) = \sum_{k=0}^{n} \left(\lfloor (k + \{t\})\alpha \rfloor + 1\right).$$

By writing $\lfloor (k + \{t\})\alpha \rfloor = (k + \{t\})\alpha - \{(k + \{t\})\alpha\}$ we obtain

$$\left|tS \cap \mathbb{Z}^2\right| = \sum_{k=0}^{n} \left((k + \{t\})\alpha + \frac{1}{2}\right) + \sum_{k=0}^{n} \left(\frac{1}{2} - \{(k + \{t\})\alpha\}\right).$$

Substituting $n = t - \{t\}$, the first sum evaluates to $g(t)$, as in the statement of the proposition. Thus

$$\left|tS \cap \mathbb{Z}^2\right| - g(t) = \sum_{k=0}^{n} \left(\frac{1}{2} - \{(k + \{t\})\alpha\}\right), \tag{4.4}$$

where $n = \lfloor t \rfloor$.

To find the expected value, it is natural to first integrate (4.4) on the interval $[n, n+1]$ for some integer $n \geq 0$ to get

$$\int_n^{n+1} \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right) \,\mathrm{d}t = \sum_{k=0}^{n} \int_0^1 \left(\frac{1}{2} - \{(k + x)\alpha\}\right) \,\mathrm{d}x.$$

By applying the change of variables $y = (k + x)\alpha$ we obtain

$$\int_n^{n+1} \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right) \,\mathrm{d}t = \sum_{k=0}^{n} \int_{k\alpha}^{(k+1)\alpha} \left(\frac{1}{2} - \{y\}\right) \frac{1}{\alpha} \,\mathrm{d}y =$$

$$\int_0^{(n+1)\alpha} \left(\frac{1}{2} - \{y\}\right) \frac{1}{\alpha}\,\mathrm{d}y = \int_0^{\{(n+1)\alpha\}} \left(\frac{1}{2} - \{y\}\right) \frac{1}{\alpha}\,\mathrm{d}y =$$

$$\frac{\{(n+1)\alpha\}\left(1 - \{(n+1)\alpha\}\right)}{2\alpha} \in \left(0, \frac{1}{8\alpha}\right).$$

Summing this over integers $0 \le n \le \lfloor T \rfloor - 1$ we get

$$0 \le \int_0^{\lfloor T \rfloor} \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right)\,\mathrm{d}t \le \frac{\lfloor T \rfloor}{8\alpha}. \tag{4.5}$$

Finally we can estimate the integral of (4.4) on $[\lfloor T \rfloor, T]$ as

$$\left|\int_{\lfloor T \rfloor}^{T} \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right)\,\mathrm{d}t\right| = \left|\sum_{k=0}^{\lfloor T \rfloor} \int_0^{\{T\}} \left(\frac{1}{2} - \{(k+x)\,\alpha\}\right)\,\mathrm{d}x\right| \le$$

$$\sum_{k=0}^{\lfloor T \rfloor} \left|\int_{k\alpha}^{(k+\{T\})\alpha} \left(\frac{1}{2} - \{y\}\right) \frac{1}{\alpha}\,\mathrm{d}y\right| \le \sum_{k=0}^{\lfloor T \rfloor} \frac{1}{8\alpha} = \frac{\lfloor T \rfloor + 1}{8\alpha}. \tag{4.6}$$

Combining (4.5) and (4.6) concludes the proof of (i):

$$\left|\frac{1}{T}\int_0^T \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right)\,\mathrm{d}t\right| \le \frac{2\lfloor T \rfloor + 1}{8\alpha T} \le \frac{3}{8\alpha}.$$

Now we prove (ii) under the assumption that the partial quotients a_k of α are $O\left(k^d\right)$. Using the notation $\lfloor t \rfloor = n$ and $\{t\} = x$ we get from (4.4) that

$$\int_n^{n+1} \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right)^2\,\mathrm{d}t = \int_0^1 \left(\sum_{k=0}^{n} \left(\frac{1}{2} - \{(k+x)\alpha\}\right)\right)^2\,\mathrm{d}x \tag{4.7}$$

for any integer $n \ge 0$. We will use Fourier analysis to find the right hand side. Consider

$$f(y) = \sum_{k=0}^{n} \left(\frac{1}{2} - \{k\alpha + y\}\right).$$

It is easy to see that its Fourier coefficients are

$$\int_0^1 \left(\sum_{k=0}^{n} \left(\frac{1}{2} - \{k\alpha + y\}\right)\right) e^{-2\pi i m y}\,\mathrm{d}y = \sum_{k=0}^{n} \frac{1}{2\pi i m} e^{2\pi i m k \alpha} = \frac{1}{2\pi i m} \cdot \frac{e^{2\pi i m (n+1)\alpha} - 1}{e^{2\pi i m \alpha} - 1}$$

for $m \ne 0$. Let

$$f_L(y) = \sum_{\substack{m=-L \\ m\neq 0}}^{L} \frac{1}{2\pi i m} \cdot \frac{e^{2\pi i m(n+1)\alpha} - 1}{e^{2\pi i m\alpha} - 1} \cdot e^{2\pi i m y}$$

denote the partial sums of the Fourier series of $f(y)$. Since $f(y)$ belongs to $L^2([0,1])$, we have that $f_L \to f$ in L^2, as $L \to \infty$.

In (4.7) we actually have $f(\alpha x)$, therefore we need to prove that $f_L(\alpha x) \to f(\alpha x)$ in L^2 as well. To see this, consider

$$\int_0^1 (f_L(\alpha x) - f(\alpha x))^2 \, \mathrm{d}x = \int_0^\alpha (f_L(y) - f(y))^2 \, \frac{1}{\alpha} \, \mathrm{d}y \leq$$

$$\frac{\lceil \alpha \rceil}{\alpha} \int_0^1 (f_L(y) - f(y))^2 \, \mathrm{d}y \to 0,$$

as $L \to \infty$, showing that indeed $f_L(\alpha x) \to f(\alpha x)$ in L^2, as $L \to \infty$. This implies that the L^2 norm square of $f_L(\alpha x)$ converges to that of $f(\alpha x)$. This gives us the representation of the right hand side of (4.7)

$$\int_0^1 \left(\sum_{k=0}^{n} \left(\frac{1}{2} - \{(k+x)\alpha\} \right) \right)^2 \mathrm{d}x = \lim_{L\to\infty} \int_0^1 |f_L(\alpha x)|^2 \, \mathrm{d}x =$$

$$\lim_{L\to\infty} \sum_{\substack{m_1,m_2=-L \\ m_1,m_2\neq 0}}^{L} \frac{1}{4\pi^2 m_1 m_2} \cdot \frac{e^{2\pi i m_1(n+1)\alpha} - 1}{e^{2\pi i m_1\alpha} - 1} \cdot \frac{e^{-2\pi i m_2(n+1)\alpha} - 1}{e^{-2\pi i m_2\alpha} - 1} \int_0^1 e^{2\pi i (m_1 - m_2)\alpha x} \, \mathrm{d}x. \tag{4.8}$$

We will separate the terms for which $m_1 = m_2$ and the terms for which $m_1 \neq m_2$. Let us thus consider the sums

$$S_1 = \sum_{\substack{m=-L \\ m\neq 0}}^{L} \frac{1}{4\pi^2 m^2} \cdot \left| \frac{e^{2\pi i m(n+1)\alpha} - 1}{e^{2\pi i m\alpha} - 1} \right|^2 = \sum_{m=1}^{L} \frac{1}{2\pi^2 m^2} \cdot \frac{\sin^2(m(n+1)\alpha\pi)}{\sin^2(m\alpha\pi)},$$

$$S_2 = \sum_{\substack{m_1,m_2\in[-L,L]\setminus\{0\} \\ m_1\neq m_2}} \frac{1}{4\pi^2 m_1 m_2} \cdot \frac{e^{2\pi i m_1(n+1)\alpha} - 1}{e^{2\pi i m_1\alpha} - 1} \cdot \frac{e^{-2\pi i m_2(n+1)\alpha} - 1}{e^{-2\pi i m_2\alpha} - 1} \cdot \frac{e^{2\pi i (m_1 - m_2)\alpha} - 1}{2\pi i (m_1 - m_2)\alpha}.$$

Now we show that $S_2 = O(1)$. To see this let us use the triangle inequality on S_2 together with the estimates

$$\left|\frac{e^{2\pi i m_1(n+1)\alpha}-1}{e^{2\pi i m_1\alpha}-1}\right| = O\left(\frac{1}{\|m_1\alpha\|}\right),$$

$$\left|\frac{e^{-2\pi i m_2(n+1)\alpha}-1}{e^{-2\pi i m_2\alpha}-1}\right| = O\left(\frac{1}{\|m_2\alpha\|}\right),$$

$$\left|e^{2\pi i(m_1-m_2)\alpha}-1\right| = O\left(\|m_1\alpha\|+\|m_2\alpha\|\right)$$

to obtain

$$S_2 = O\left(\sum_{\substack{m_1,m_2\in[-L,L]\setminus\{0\}\\ m_1\neq m_2}} \frac{1}{|m_1m_2|}\cdot\frac{1}{|m_1-m_2|}\cdot\left(\frac{1}{\|m_1\alpha\|}+\frac{1}{\|m_2\alpha\|}\right)\right).$$

By symmetry the terms $\frac{1}{\|m_1\alpha\|}$ and $\frac{1}{\|m_2\alpha\|}$ have the same contribution. It is also easy to see that it is enough to keep the terms for which m_1 and m_2 have the same sign. Thus we get

$$S_2 = O\left(\sum_{m_1=1}^{\infty}\frac{1}{m_1\|m_1\alpha\|}\sum_{\substack{m_2=1\\ m_2\neq m_1}}^{\infty}\frac{1}{m_2|m_1-m_2|}\right).$$

Since

$$\sum_{\substack{m_2=1\\ m_2\neq m_1}}^{\infty}\frac{1}{m_2|m_1-m_2|} = O\left(\frac{\log m_1+1}{m_1}\right),$$

this simplifies to

$$S_2 = O\left(\sum_{m_1=1}^{\infty}\frac{\log m_1+1}{m_1^2\|m_1\alpha\|}\right). \tag{4.9}$$

Consider the convergents

$$\frac{p_k}{q_k} = [a_0; a_1, a_2, \ldots, a_{k-1}]$$

of α. Using Proposition 4.2 (i) we can estimate the terms as m_1 runs between two consecutive convergent denominators as

$$\sum_{q_k \le m_1 < q_{k+1}} \frac{\log m_1 + 1}{m_1^2 \, \|m_1\alpha\|} \le \frac{\log q_{k+1} + 1}{q_k^2} \sum_{0 < m_1 < q_{k+1}} \frac{1}{\|m_1\alpha\|} =$$

$$O\left(\frac{\log q_{k+1}}{q_k^2} \cdot q_{k+1} \log q_{k+1}\right). \tag{4.10}$$

The recurrence $q_{k+1} = a_k q_k + q_{k-1}$ from Proposition 4.1 (iii), together with the assumption $a_k = O\left(k^d\right)$ shows that (4.10) is $O\left(\frac{k^d \log^2 q_k}{q_k}\right)$. Therefore (4.9) and (4.10) yield

$$S_2 = O\left(\sum_{k=1}^{\infty} \frac{k^d \log^2 q_k}{q_k}\right).$$

Since q_k is at least as big as the kth Fibonacci number, we get that indeed $S_2 = O(1)$ with an effective implied constant depending only on α. Altogether, from (4.7) , (4.8) and the definition of the sums S_1 and S_2 we obtain

$$\int_n^{n+1} \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right)^2 \, \mathrm{d}t = \sum_{m=1}^{\infty} \frac{1}{2\pi^2 m^2} \cdot \frac{\sin^2\left(m(n+1)\alpha\pi\right)}{\sin^2\left(m\alpha\pi\right)} + O(1) \tag{4.11}$$

for any integer $n \ge 0$.

We proceed by estimating the tail of this series using summation by parts. First note that

$$\frac{\sin^2\left(m(n+1)\alpha\pi\right)}{\sin^2\left(m\alpha\pi\right)} = \left|\sum_{k=0}^{n} e^{2\pi i k m\alpha}\right|^2 = n + 1 + \sum_{\substack{0 \le k_1, k_2 \le n \\ k_1 \ne k_2}} e^{2\pi i (k_1 - k_2) m\alpha}.$$

For the partial sums b_j of this sequence we have

$$b_j = \sum_{m=1}^{j} \frac{\sin^2\left(m(n+1)\alpha\pi\right)}{\sin^2\left(m\alpha\pi\right)} = j(n+1) + \sum_{\substack{0 \le k_1, k_2 \le n \\ k_1 \ne k_2}} \sum_{m=1}^{j} e^{2\pi i (k_1 - k_2) m\alpha} =$$

$$j(n+1) + \sum_{\substack{0 \le k_1, k_2 \le n \\ k_1 \ne k_2}} \frac{e^{2\pi i (k_1 - k_2) j\alpha} - 1}{1 - e^{-2\pi i (k_1 - k_2)\alpha}}.$$

Here the terms of the sum satisfy

$$\frac{e^{2\pi i(k_1-k_2)j\alpha}-1}{1-e^{-2\pi i(k_1-k_2)\alpha}} = O\left(\frac{1}{\|(k_1-k_2)\alpha\|}\right),$$

thus

$$b_j = O\left(jn + \sum_{\substack{0\le k_1,k_2\le n\\ k_1\neq k_2}} \frac{1}{\|(k_1-k_2)\alpha\|}\right).$$

Consider the two consecutive convergent denominators of α for which $q_k \le n < q_{k+1}$. In the sum above $|k_1-k_2|$ takes integral values between 1 and $n < q_{k+1}$, and each value is attained at most $2n$ times, therefore

$$b_j = O\left(jn + n\sum_{0<\ell<q_{k+1}} \frac{1}{\|\ell\alpha\|}\right).$$

Proposition 4.2 (i) implies that

$$b_j = O\left(jn + nq_{k+1}\log q_{k+1}\right).$$

Using the recurrence $q_{k+1} = a_kq_k + q_{k-1}$ and the assumption $a_k = O\left(k^d\right)$ we get $q_{k+1}\log q_{k+1} = O\left(k^dq_k\log q_k\right)$. Since q_k is at least as big as the kth Fibonacci number, $q_k \le n$ also implies $k = O(\log n)$. Altogether

$$b_j = O\left(jn + n^2\log^{d+1} n\right).$$

Applying summation by parts on the series in (4.11) starting at $m = \lfloor n\sqrt{\log n}\rfloor$ we get

$$\sum_{m=\lfloor n\sqrt{\log n}\rfloor}^{\infty} \frac{1}{2\pi^2m^2}\cdot\frac{\sin^2\left(m(n+1)\alpha\pi\right)}{\sin^2\left(m\alpha\pi\right)} =$$

$$-b_{\lfloor n\sqrt{\log n}\rfloor-1}\frac{1}{2\pi^2\lfloor n\sqrt{\log n}\rfloor^2} + \sum_{m=\lfloor n\sqrt{\log n}\rfloor}^{\infty} b_m\left(\frac{1}{2\pi^2m^2}-\frac{1}{2\pi^2(m+1)^2}\right) =$$

$$O\left(\log^d n + \sum_{m=\lfloor n\sqrt{\log n}\rfloor}^{\infty}\frac{mn+n^2\log^{d+1} n}{m^3}\right) =$$

$$O\left(\log^d n + n\sum_{m=\lfloor n\sqrt{\log n}\rfloor}^{\infty}\frac{1}{m^2} + n^2\log^{d+1} n\sum_{m=\lfloor n\sqrt{\log n}\rfloor}^{\infty}\frac{1}{m^3}\right) = O\left(\log^d n\right).$$

From (4.11) we therefore get that

$$\int_n^{n+1}\left(\left|tS\cap\mathbb{Z}^2\right| - g(t)\right)^2\,\mathrm{d}t = \sum_{1\le m\le M\sqrt{\log M}}\frac{1}{2\pi^2m^2}\cdot\frac{\sin^2\left(m(n+1)\alpha\pi\right)}{\sin^2\left(m\alpha\pi\right)} + O\left(\log^d M\right)$$

for any integers $M \ge 2$ and $0 \le n < M$. By taking the average over the integers $0 \le n < M$, we can express the variance on the interval $[0, M]$ as

$$\frac{1}{M}\int_0^M\left(\left|tS\cap\mathbb{Z}^2\right| - g(t)\right)^2\,\mathrm{d}t =$$

$$\frac{1}{M}\sum_{n=0}^{M-1}\sum_{1\le m\le M\sqrt{\log M}}\frac{1}{2\pi^2m^2}\cdot\frac{\sin^2\left(m(n+1)\alpha\pi\right)}{\sin^2\left(m\alpha\pi\right)} + O\left(\log^d M\right).$$

Let us now switch the order of summation, and use the identity

$$\frac{1}{M}\sum_{n=0}^{M-1}\sin^2\left(m(n+1)\alpha\pi\right) = \frac{1}{2} + \frac{1}{4M}\cdot\frac{\sin(m\alpha\pi) - \sin(m(2M+1)\alpha\pi)}{\sin(m\alpha\pi)}$$

to get

$$\frac{1}{M}\int_0^M\left(\left|tS\cap\mathbb{Z}^2\right| - g(t)\right)^2\,\mathrm{d}t = \sum_{1\le m\le M\sqrt{\log M}}\frac{1}{4\pi^2m^2\sin^2(m\alpha\pi)} +$$

$$\frac{1}{M}\sum_{1\le m\le M\sqrt{\log M}}\frac{\sin(m\alpha\pi) - \sin(m(2M+1)\alpha\pi)}{8\pi^2m^2\sin^3(m\alpha\pi)} + O\left(\log^d M\right). \tag{4.12}$$

We now focus on the terms $1 \le m \le \frac{M}{\log^{3d} M}$ of the second sum in (4.12). Using Proposition 4.2 (ii) we can estimate the terms as m runs between two consecutive convergent denominators of α as

$$\frac{1}{M}\sum_{q_k\le m<q_{k+1}}\left|\frac{\sin(m\alpha\pi) - \sin(m(2M+1)\alpha\pi)}{8\pi^2m^2\sin^3(m\alpha\pi)}\right| = O\left(\frac{1}{M}\sum_{q_k\le m<q_{k+1}}\frac{1}{m^2\left\|m\alpha\right\|^3}\right) =$$

$$O\left(\frac{1}{Mq_k^2}\sum_{0<m<q_{k+1}}\frac{1}{\left\|m\alpha\right\|^3}\right) = O\left(\frac{q_{k+1}^3}{Mq_k^2}\right) = O\left(\frac{k^{3d}q_k}{M}\right).$$

Let us sum this estimate over every positive integer k such that $q_k \leq \frac{M}{\log^{3d} M}$. For every such k we have $k = O(\log M)$, therefore we obtain

$$\frac{1}{M} \sum_{1 \leq m \leq \frac{M}{\log^{3d} M}} \left| \frac{\sin(m\alpha\pi) - \sin(m(2M+1)\alpha\pi)}{8\pi^2 m^2 \sin^3(m\alpha\pi)} \right| = O\left(\frac{\log^{3d} M}{M} \sum_{q_k \leq \frac{M}{\log^{3d} M}} q_k \right). \tag{4.13}$$

We can estimate the sum of the convergent denominators as follows. By summing the recurrence relation in Proposition 4.1 (iii) we get

$$2q_k \geq q_k + q_{k-1} = a_{k-1}q_{k-1} + a_{k-2}q_{k-2} + \cdots + a_2 q_2 + q_2 + q_1 \geq q_{k-1} + q_{k-2} + \cdots + q_2 + q_1,$$

and thus we obtain the general inequality

$$q_1 + q_2 + \cdots + q_k \leq 3q_k.$$

Hence (4.13) simplifies as

$$\frac{1}{M} \sum_{1 \leq m \leq \frac{M}{\log^{3d} M}} \left| \frac{\sin(m\alpha\pi) - \sin(m(2M+1)\alpha\pi)}{8\pi^2 m^2 \sin^3(m\alpha\pi)} \right| = O(1). \tag{4.14}$$

Now we consider the terms $\frac{M}{\log^{3d} M} \leq m \leq M\sqrt{\log M}$. Using the estimate

$$|\sin(m\alpha\pi) - \sin(m(2M+1)\alpha\pi)| \leq (2M+2)\,|\sin(m\alpha\pi)|$$

we get

$$\frac{1}{M} \sum_{\frac{M}{\log^{3d} M} \leq m \leq M\sqrt{\log M}} \left| \frac{\sin(m\alpha\pi) - \sin(m(2M+1)\alpha\pi)}{8\pi^2 m^2 \sin^3(m\alpha\pi)} \right| =$$

$$O\left(\sum_{\frac{M}{\log^{3d} M} \leq m \leq M\sqrt{\log M}} \frac{1}{m^2 \left\| m\alpha \right\|^2} \right). \tag{4.15}$$

Thus using (4.14) and (4.15), (4.12) simplifies to

$$\frac{1}{M} \int_0^M \left(\left| tS \cap \mathbb{Z}^2 \right| - g(t) \right)^2 \, \mathrm{d}t = \sum_{m=1}^{M} \frac{1}{4\pi^2 m^2 \sin^2(m\alpha\pi)} +$$

$$O\left(\sum_{\frac{M}{\log^{3d} M}\le m\le M\sqrt{\log M}} \frac{1}{m^2\left\|m\alpha\right\|^2}+\log^d M\right). \tag{4.16}$$

Using Proposition 4.2 (ii) we can estimate the terms of the error in (4.16) as m runs between two consecutive convergent denominators as

$$\sum_{q_k\le m<q_{k+1}} \frac{1}{m^2\left\|m\alpha\right\|^2}\le \frac{1}{q_k^2}\sum_{0<m<q_{k+1}} \frac{1}{\left\|m\alpha\right\|^2}=O\left(\frac{q_{k+1}^2}{q_k^2}\right)=O\left(k^{2d}\right)=O\left(\log^{2d} M\right).$$

The general inequality

$$\frac{q_{k+2}}{q_k}=\frac{a_{k+1}q_{k+1}+q_k}{q_k}\ge 2$$

shows that the number of convergent denominators which fall into $\left[\frac{M}{\log^{3d} M}, M\sqrt{\log M}\right]$ is $O\left(\log\log M\right)$. Therefore

$$\sum_{\frac{M}{\log^{3d} M}\le m\le M\sqrt{\log M}} \frac{1}{m^2\left\|m\alpha\right\|^2}=O\left(\log^{2d} M\log\log M\right).$$

Thus (4.16) simplifies as

$$\frac{1}{M}\int_0^M \left(\left|tS\cap\mathbb{Z}^2\right|-g(t)\right)^2\,\mathrm{d}t=\sum_{m=1}^{M}\frac{1}{4\pi^2m^2\sin^2(m\alpha\pi)}+O\left(\log^{2d} M\log\log M\right) \tag{4.17}$$

for any integer $M\ge 3$. Finally, for an arbitrary real $T\ge 3$, we can apply (4.17) with $M=\lfloor T\rfloor$ and $M=\lceil T\rceil$ to conclude the proof of (ii).

□

Proposition 4.3 (i) gives a satisfactory result on the expected value (4.1) for an arbitrary irrational real α. Proposition 4.3 (ii) only holds, however, for irrational reals the partial quotients of which grow at most polynomially fast. There are two classes of irrational numbers which satisfy this condition. According to the theorem of Lagrange, the sequence of partial quotients of a quadratic irrational is eventually periodic, which implies that every quadratic irrational real α satisfies the conditions of Proposition 4.3

(ii) with $d = 0$. There is also a class of irrational reals related to Euler's number e which is known to satisfy the same conditions with $d = 1$.

In order to use Proposition 4.3 (ii) to find the standard deviation (4.2), we need to evaluate (4.3). In [1] (4.3) is evaluated for quadratic irrationals in terms of arithmetic quantities of the real quadratic field $\mathbb{Q}(\alpha)$. The question is raised by Beck whether it is possible to evaluate (4.3) for an arbitrary irrational α in terms of its partial quotients. Since

$$\sum_{m=1}^{\lfloor T \rfloor} \frac{1}{4\pi^2 m^2 \sin^2(m\alpha\pi)} = \sum_{m=1}^{\lfloor T \rfloor} \frac{1}{4\pi^4 m^2 \|m\alpha\|^2} + O(1),$$

the following proposition provides such an evaluation up to an explicit error term.

Proposition 4.4 *Let $\alpha = [a_0; a_1, a_2, \ldots]$ be the continued fraction representation of an irrational real number α, and let $\frac{p_k}{q_k} = [a_0; a_1, a_2, \ldots, a_{k-1}]$ denote its convergents. For any integer $k \geq 2$ we have*

$$\sqrt{\sum_{0<m<q_k} \frac{1}{m^2 \|m\alpha\|^2}} = \sqrt{\frac{\pi^4}{90} \sum_{0<\ell<k} a_\ell^2} + O\left(\sqrt{k}\right).$$

The implied constant is absolute and less than 150.

For the class of quadratic irrationals, and more generally for irrationals the partial quotients of which are bounded, the error term has the same order of magnitude as the main term. If, on the other hand the quadratic means of the partial quotients satisfy

$$\sqrt{\frac{1}{k} \sum_{\ell=1}^{k} a_\ell^2} \to \infty,$$

as $k \to \infty$, in particular if α belongs to the class related to Euler's number e, Proposition 4.4 evaluates (4.3). Note that here we do not have to assume that the partial quotients grow at most polynomially fast.

Proposition 4.3 (ii) and Proposition 4.4 make it possible to express the variance in terms of the partial quotients of α as follows.

Corollary 4.5 *Let $\alpha > 0$ be irrational, and let S be the closed right triangle with vertices $(0,0), (1,0), (0,\alpha)$. Let*

$$g(t) = \frac{\alpha}{2} t^2 + \frac{\alpha + 1}{2} t + \frac{(\alpha \{t\} + 1)(1 - \{t\})}{2}.$$

Suppose the continued fraction representation $\alpha = [a_0; a_1, a_2, \ldots]$ satisfies $a_k = O(k^d)$ for some real number $d \geq 0$. Let $\frac{p_k}{q_k} = [a_0; a_1, a_2, \ldots, a_{k-1}]$ denote the convergents to α. Then for any real $T \geq 3$ we have

$$\frac{1}{T}\int_0^T \left(\left|tS \cap \mathbb{Z}^2\right| - g(t)\right)^2 \, \mathrm{d}t = \frac{1}{360}\sum_{q_\ell < T} a_\ell^2 + O\left(\log^{d+1} T + \log^{2d} T \log\log T\right).$$

The sum is over all positive integers ℓ such that $q_\ell < T$. The implied constant depends only on α and is effective.

We conclude chapter 4 with the proofs of Proposition 4.4 and Corollary 4.5.

Proof of Proposition 4.4: Let $\ell \geq 2$ be such that $q_\ell \geq 2$, let $\varepsilon_\ell = q_\ell\alpha - p_\ell$, and consider the sum

$$\sum_{q_\ell \leq m < q_{\ell+1}} \frac{1}{m^2 \left\|m\alpha\right\|^2}.$$

We have

$$\|m\alpha\| = \left\|\frac{mp_\ell}{q_\ell} + \frac{m\varepsilon_\ell}{q_\ell}\right\|. \tag{4.18}$$

Let us decompose the sum using the index sets

$$A = \left\{q_\ell \leq m < q_{\ell+1} : mp_\ell \equiv 0 \pmod{q_\ell}\right\},$$

$$B = \left\{q_\ell \leq m < q_{\ell+1} : mp_\ell \equiv (-1)^\ell \pmod{q_\ell}\right\},$$

$$C = \left\{q_\ell \leq m < q_{\ell+1} : mp_\ell \not\equiv 0, (-1)^\ell \pmod{q_\ell}\right\}.$$

We first show that the contribution of the terms $m \in C$ is negligible. According to Proposition 4.1 (ii) for any $q_\ell \leq m < q_{\ell+1}$ we have

$$\left|\frac{m\varepsilon_\ell}{q_\ell}\right| = \frac{m\left\|q_\ell\alpha\right\|}{q_\ell} < \frac{1}{q_\ell}.$$

Using $\operatorname{sign}(\varepsilon_\ell) = (-1)^{\ell+1}$ from Proposition 4.1 (vi), we therefore get that for any $m \in C$ we have

$$\|m\alpha\| = \left\| \frac{mp_\ell}{q_\ell} + \frac{m\varepsilon_\ell}{q_\ell} \right\| \geq \frac{1}{2} \left\| \frac{mp_\ell}{q_\ell} \right\|.$$

Hence for any integer $1 \leq a \leq a_\ell$ we have the estimate

$$\sum_{\substack{aq_\ell < m < (a+1)q_\ell \\ m \in C}} \frac{1}{m^2 \|m\alpha\|^2} \leq \sum_{aq_\ell < m < (a+1)q_\ell} \frac{4}{a^2 q_\ell^2 \left\| \frac{mp_\ell}{q_\ell} \right\|^2}. \tag{4.19}$$

From Proposition 4.1 (iv) we know that p_ℓ and q_ℓ are relatively prime. Hence as m runs in the interval $aq_\ell < m < (a+1)q_\ell$, the integers mp_ℓ fall into each nonzero residue class modulo q_ℓ exactly once, yielding

$$\sum_{aq_\ell < m < (a+1)q_\ell} \frac{4}{a^2 q_\ell^2 \left\| \frac{mp_\ell}{q_\ell} \right\|^2} = \sum_{j=1}^{q_\ell - 1} \frac{4}{a^2 q_\ell^2 \left\| \frac{j}{q_\ell} \right\|^2} \leq \frac{8}{a^2 q_\ell^2} \sum_{1 \leq j \leq \frac{q_\ell}{2}} \frac{q_\ell^2}{j^2} \leq \frac{8}{a^2} \cdot \sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{4\pi^2}{3a^2}.$$

Thus (4.19) gives

$$\sum_{\substack{aq_\ell < m < (a+1)q_\ell \\ m \in C}} \frac{1}{m^2 \|m\alpha\|^2} \leq \frac{4\pi^2}{3a^2}. \tag{4.20}$$

Since $(a_\ell + 1)q_\ell > a_\ell q_\ell + q_{\ell-1} = q_{\ell+1}$, it is enough to sum (4.20) over integers $1 \leq a \leq a_\ell$ to conclude

$$\sum_{m \in C} \frac{1}{m^2 \|m\alpha\|^2} \leq \sum_{a=1}^{a_\ell} \frac{4\pi^2}{3a^2} \leq \frac{2\pi^4}{9}. \tag{4.21}$$

Now we estimate the contribution of the terms $m \in B$. From Proposition 4.1 (v) we know that $p_\ell q_{\ell-1} - q_\ell p_{\ell-1} = (-1)^\ell$. By taking this equation modulo q_ℓ we learn that the multiplicative inverse of p_ℓ in the ring $\mathbb{Z}_{q_\ell}$ is $(-1)^\ell q_{\ell-1}$. This means that the set B consists of integers $q_\ell \leq m < q_{\ell+1}$ which are congruent to $q_{\ell-1}$ modulo q_ℓ. Therefore

$$B = \{aq_\ell + q_{\ell-1} : 1 \leq a \leq a_\ell - 1\},$$

since the choice $a = a_\ell$ would result in $a_\ell q_\ell + q_{\ell-1} = q_{\ell+1}$ which is outside our interval $q_\ell \leq m < q_{\ell+1}$. For an arbitrary element $m = aq_\ell + q_{\ell-1} \in B$ we get from (4.18) that

$$\|m\alpha\| = \left\| \frac{(-1)^\ell}{q_\ell} + \frac{(aq_\ell + q_{\ell-1})\varepsilon_\ell}{q_\ell} \right\| = \frac{1}{q_\ell} - \frac{(aq_\ell + q_{\ell-1})\,|\varepsilon_\ell|}{q_\ell} = \frac{1 - q_{\ell-1}\,|\varepsilon_\ell|}{q_\ell} - a\,|\varepsilon_\ell|\,.$$

Proposition 4.1 (ii) implies that

$$|\varepsilon_\ell| = \|q_\ell \alpha\| < \frac{1}{q_{\ell+1}} = \frac{1}{a_\ell q_\ell + q_{\ell-1}}.$$

Rearranging this inequality yields

$$a_\ell q_\ell \, |\varepsilon_\ell| \leq 1 - q_{\ell-1} \, |\varepsilon_\ell| \, ,$$

which shows that for every element $m = aq_\ell + q_{\ell-1} \in B$ we have

$$\|m\alpha\| \geq (a_\ell - a) \, |\varepsilon_\ell| \, .$$

Thus we get

$$\sum_{m \in B} \frac{1}{m^2 \, \|m\alpha\|^2} \leq \sum_{a=1}^{a_\ell - 1} \frac{1}{a^2 q_\ell^2 (a_\ell - a)^2 \varepsilon_\ell^2} \leq \frac{2}{q_\ell^2 \varepsilon_\ell^2} \sum_{1 \leq a \leq \frac{a_\ell}{2}} \frac{4}{a^2 a_\ell^2} \leq \frac{4\pi^2}{3 a_\ell^2 q_\ell^2 \varepsilon_\ell^2}.$$

From Proposition 4.1 (ii) we learn that

$$|\varepsilon_\ell| = \|q_\ell \alpha\| > \frac{1}{q_{\ell+1} + q_\ell} \geq \frac{1}{3 a_\ell q_\ell},$$

hence

$$\sum_{m \in B} \frac{1}{m^2 \, \|m\alpha\|^2} \leq 12\pi^2. \tag{4.22}$$

Finally let us consider the contribution of the terms $m \in A$, which is the main term. We have

$$A = \{ aq_\ell : 1 \leq a \leq a_\ell \} \, .$$

Indeed, the choice $a = a_\ell + 1$ would result in $(a_\ell + 1) q_\ell = q_{\ell+1} + q_\ell - q_{\ell-1} > q_{\ell+1}$, which is outside our interval $q_\ell \leq m < q_{\ell+1}$. For every $m = aq_\ell \in A$ we have

$$\|m\alpha\| = a \, \|q_\ell \alpha\| = a \, |\varepsilon_\ell| \, ,$$

and hence

$$\sum_{m \in A} \frac{1}{m^2 \, \|m\alpha\|^2} = \sum_{a=1}^{a_\ell} \frac{1}{a^2 q_\ell^2 a^2 \varepsilon_\ell^2} = \frac{1}{q_\ell^2 \varepsilon_\ell^2} \sum_{a=1}^{\infty} \frac{1}{a^4} - \frac{1}{q_\ell^2 \varepsilon_\ell^2} \sum_{a=a_\ell+1}^{\infty} \frac{1}{a^4}.$$

Here $\sum_{a=1}^{\infty} \frac{1}{a^4} = \frac{\pi^4}{90}$, and

$$\frac{1}{q_\ell^2 \varepsilon_\ell^2} \sum_{a=a_{\ell+1}}^{\infty} \frac{1}{a^4} \le \frac{1}{q_\ell^2 \varepsilon_\ell^2} \int_{a_\ell}^{\infty} \frac{1}{x^4}\, \mathrm{d}x = \frac{1}{q_\ell^2 \varepsilon_\ell^2} \cdot \frac{1}{3a_\ell^3} \le 3.$$

To estimate the main term we can use Proposition 4.1 (ii) and (iii) to obtain

$$\frac{1}{(a_\ell+2)q_\ell} \le \frac{1}{(a_\ell+1)q_\ell + q_{\ell-1}} \le \|q_\ell \alpha\| = |\varepsilon_\ell| \le \frac{1}{a_\ell q_\ell + q_{\ell-1}} \le \frac{1}{a_\ell q_\ell},$$

and hence

$$a_\ell \le \frac{1}{q_\ell\, |\varepsilon_\ell|} \le a_\ell + 2,$$

$$a_\ell^2 \le \frac{1}{q_\ell^2 \varepsilon_\ell^2} \le a_\ell^2 + 4a_\ell + 4.$$

Thus we get

$$\left| \sum_{m \in A} \frac{1}{m^2 \|m\alpha\|^2} - \frac{\pi^4}{90} a_\ell^2 \right| \le \frac{\pi^4}{90} (4a_\ell + 4) + 3. \tag{4.23}$$

Using (4.21), (4.22) and (4.23) we obtain

$$\left| \sum_{q_\ell \le m < q_{\ell+1}} \frac{1}{m^2 \|m\alpha\|^2} - \frac{\pi^4}{90} a_\ell^2 \right| \le \frac{2\pi^4}{9} + 12\pi^2 + \frac{\pi^4}{90} (4a_\ell + 4) + 3.$$

Since

$$\frac{2\pi^4}{9} + 12\pi^2 + \frac{\pi^4}{90} (4+4) + 3 < 152,$$

we conclude that

$$\left| \sum_{q_\ell \le m < q_{\ell+1}} \frac{1}{m^2 \|m\alpha\|^2} - \frac{\pi^4}{90} a_\ell^2 \right| \le 152 a_\ell \tag{4.24}$$

for any $\ell \ge 2$ such that $q_\ell \ge 2$.

We claim that (4.24) holds for every integer $\ell \ge 1$. If $a_1 > 1$ we have

$$1 = q_1 < q_2 < \cdots,$$

thus we only have to prove (4.24) in the special case $\ell = 1$. If on the other hand $a_1 = 1$ we have

$$1 = q_1 = q_2 < q_3 < \cdots,$$

thus we have to prove (4.24) in the cases $\ell = 1$ and $\ell = 2$.

Suppose first that $a_1 > 1$. Since $q_1 = 1$ and $q_2 = a_1$ we have

$$\sum_{q_1 \le m < q_2} \frac{1}{m^2 \left\|m\alpha\right\|^2} = \sum_{m=1}^{a_1-1} \frac{1}{m^2 \left\|m\alpha\right\|^2}.$$

According to the rules of the continued fraction process we have $\lfloor \alpha \rfloor = a_0$ and

$$\frac{1}{a_1+1} < \alpha - a_0 < \frac{1}{a_1}.$$

Hence for any $1 \le m \le \frac{a_1}{2}$ we have

$$\frac{m}{a_1+1} < \left\|m\alpha\right\| = \left|m\alpha - ma_0\right| < \frac{m}{a_1},$$

$$\left| \frac{1}{m^2 \left\|m\alpha\right\|^2} - \frac{a_1^2}{m^4} \right| \le \frac{2a_1+2}{m^4},$$

and thus

$$\left| \sum_{1 \le m \le \frac{a_1}{2}} \frac{1}{m^2 \left\|m\alpha\right\|^2} - \frac{\pi^4}{90} a_1^2 \right| \le \sum_{1 \le m \le \frac{a_1}{2}} \frac{2a_1+2}{m^4} + \sum_{\frac{a_1+1}{2} \le m} \frac{a_1^2}{m^4} \le 4a_1 \frac{\pi^4}{90} + \frac{a_1^2}{3\left(\frac{a_1-1}{2}\right)^3} \le$$

$$\left(\frac{2\pi^4}{45} + \frac{16}{3} \right) a_1. \tag{4.25}$$

If $\frac{a_1+1}{2} \le m \le a_1 - 1$ then

$$\frac{a_1 - m}{a_1} < \left\|m\alpha\right\| = \left|m\alpha - ma_0 - 1\right| < \frac{a_1 + 1 - m}{a_1 + 1},$$

$$\sum_{\frac{a_1+1}{2} \le m \le a_1 - 1} \frac{1}{m^2 \left\|m\alpha\right\|^2} \le \frac{4}{(a_1+1)^2} \sum_{\frac{a_1+1}{2} \le m \le a_1 - 1} \frac{a_1^2}{(a_1 - m)^2} \le \frac{2\pi^2}{3}. \tag{4.26}$$

Since $\frac{2\pi^4}{45} + \frac{16}{3} + \frac{2\pi^2}{3} < 152$, (4.25) and (4.26) imply that (4.24) holds for any $\ell \ge 1$ if $a_1 > 1$.

Finally suppose that $a_1 = 1$. Then $q_1 = q_2 = 1$ and $q_3 = a_2 + 1$, therefore (4.24) trivially holds for $\ell = 1$. Let us thus consider

$$\sum_{q_2 \le m < q_3} \frac{1}{m^2 \|m\alpha\|^2} = \sum_{m=1}^{a_2} \frac{1}{m^2 \|m\alpha\|^2}.$$

According to the rules of the continued fraction process we have $\lfloor \alpha \rfloor = a_0$ and

$$\frac{1}{a_2+1} < \frac{1}{\alpha - a_0} - 1 < \frac{1}{a_2},$$

$$\frac{-1}{a_2+1} < \alpha - a_0 - 1 < \frac{-1}{a_2+2}.$$

Hence for any $1 \le m \le \frac{a_2+1}{2}$ we have

$$\frac{m}{a_2+2} \le \|m\alpha\| = |m\alpha - ma_0 - m| \le \frac{m}{a_2+1},$$

$$\left| \frac{1}{m^2 \|m\alpha\|^2} - \frac{a_2^2}{m^4} \right| \le \frac{4a_2+4}{m^4},$$

and thus

$$\left| \sum_{1 \le m \le \frac{a_2+1}{2}} \frac{1}{m^2 \|m\alpha\|^2} - \frac{\pi^4}{90} a_2^2 \right| \le \sum_{1 \le m \le \frac{a_2+1}{2}} \frac{4a_2+4}{m^4} + \sum_{m \ge \frac{a_2+2}{2}} \frac{a_2^2}{m^4} \le$$

$$\frac{\pi^4}{90}(4a_2+4) + \frac{a_2^2}{3\left(\frac{a_2}{2}\right)^3} \le \left(\frac{4\pi^4}{45} + \frac{8}{3}\right) a_2. \tag{4.27}$$

If $\frac{a_2+2}{2} \le m \le a_2$ then

$$\frac{a_2+1-m}{a_2+1} \le \|m\alpha\| = |m\alpha - ma_0 - m + 1| \le \frac{a_2+2-m}{a_2+2},$$

$$\sum_{\frac{a_2+2}{2} \le m \le a_2} \frac{1}{m^2 \|m\alpha\|^2} \le \frac{4}{(a_2+2)^2} \sum_{\frac{a_2+2}{2} \le m \le a_2} \frac{(a_2+1)^2}{(a_2+1-m)^2} \le \frac{2\pi^2}{3}. \tag{4.28}$$

Since $\frac{4\pi^4}{45} + \frac{8}{3} + \frac{2\pi^2}{3} < 152$, (4.27) and (4.28) shows that (4.24) holds for $\ell = 2$ if $a_1 = 1$. This concludes the proof of the fact that (4.24) holds for every integer $\ell \ge 1$.

Applying (4.24) on the integers $0 < \ell < k$ we get

$$\left| \sum_{0<m<q_k} \frac{1}{m^2 \left\|m\alpha\right\|^2} - \frac{\pi^4}{90} \sum_{0<\ell<k} a_\ell^2 \right| \leq 152 \sum_{0<\ell<k} a_\ell.$$

Let us now use the general inequality

$$\left|\sqrt{A} - \sqrt{B}\right| = \frac{|A-B|}{\sqrt{A}+\sqrt{B}} \leq \frac{|A-B|}{\sqrt{B}}$$

to get

$$\left| \sqrt{\sum_{0<m<q_k} \frac{1}{m^2 \left\|m\alpha\right\|^2}} - \sqrt{\frac{\pi^4}{90} \sum_{0<\ell<k} a_\ell^2} \right| \leq \frac{152 \sum_{0<\ell<k} a_\ell}{\sqrt{\frac{\pi^4}{90} \sum_{0<\ell<k} a_\ell^2}}.$$

Finally, the inequality between the arithmetic and quadratic means implies

$$\frac{152 \sum_{0<\ell<k} a_\ell}{\sqrt{\frac{\pi^4}{90} \sum_{0<\ell<k} a_\ell^2}} \leq \frac{152}{\sqrt{\frac{\pi^4}{90}}} \sqrt{k-1} < 150\sqrt{k}$$

concluding the proof.

□

Proof of Corollary 4.5: According to Proposition 4.3 (ii) we have

$$\frac{1}{T} \int_0^T \left(\left|tS \cap \mathbb{Z}^2\right| - g(t) \right)^2 \, \mathrm{d}t = \sum_{m=1}^{\lfloor T \rfloor} \frac{1}{4\pi^2 m^2 \sin^2(m\alpha\pi)} + O\left(\log^{2d} T \log\log T\right).$$

It is easy to see that for any $-\frac{1}{2} \leq x \leq \frac{1}{2}$, $x \neq 0$ we have

$$\frac{1}{\sin^2(\pi x)} - \frac{1}{\pi^2 x^2} = O(1),$$

therefore for any $x \in \mathbb{R} \backslash \mathbb{Z}$ we also have

$$\frac{1}{\sin^2(\pi x)} - \frac{1}{\pi^2 \left\|x\right\|^2} = O(1).$$

Hence our formula for the variance simplifies as

$$\frac{1}{T} \int_0^T \left(\left|tS \cap \mathbb{Z}^2\right| - g(t) \right)^2 \, \mathrm{d}t = \sum_{m=1}^{\lfloor T \rfloor} \frac{1}{4\pi^4 m^2 \left\|m\alpha\right\|^2} + O\left(\log^{2d} T \log\log T\right).$$

Consider the two consecutive convergent denominators of α for which $q_k \leq \lfloor T \rfloor < q_{k+1}$. Since q_k is at least as big as the kth Fibonacci number, we have $k = O(\log T)$. We can apply Proposition 4.4 to obtain

$$\sum_{0<m<q_k} \frac{1}{4\pi^4 m^2 \|m\alpha\|^2} = \frac{1}{360} \sum_{0<\ell<k} a_\ell^2 + O\left(\sqrt{k}\sqrt{\sum_{0<\ell<k} a_\ell^2}\right) =$$

$$\frac{1}{360} \sum_{0<\ell<k} a_\ell^2 + O\left(k^{d+1}\right) = \frac{1}{360} \sum_{0<\ell<k} a_\ell^2 + O\left(\log^{d+1} T\right).$$

Finally note that the terms of the sum $\sum a_\ell^2$ are $O\left(\log^{2d} T\right)$, therefore

$$\sum_{m=1}^{\lfloor T \rfloor} \frac{1}{4\pi^4 m^2 \|m\alpha\|^2} = \frac{1}{360} \sum_{q_\ell<T} a_\ell^2 + O\left(\log^{d+1} T + \log^{2d} T\right).$$

□

References

[1] József Beck. *Probabilistic Diophantine approximation. Randomness in lattice point counting.* Springer Monographs in Mathematics. Springer, Cham, 2014. xvi+487 pp. ISBN: 978-3-319-10740-0.

[2] Matthias Beck. *Counting lattice points by means of the residue theorem.* The Ramanujan Journal, 4 (2000), 299-310.

[3] Vidmantas Bentkus, Friedrich Götze. *On the lattice point problem for ellipsoids.* Acta Arithmetica, 80 (1997), no. 2, 101-125.

[4] John W. S. Cassels. *An introduction to Diophantine approximation.* Cambridge Tracts in Mathematics and Mathematical Physics, no. 45. Cambridge University Press, New York, 1957. x+166 pp.

[5] Harold S. M. Coxeter. *Regular polytopes.* Pitman Publishing Corporation, New York, 1948. xix+321 pp.

[6] Ricardo Diaz, Sinai Robins. *The Ehrhart polynomial of a lattice n-simplex.* Electronic Research Announcements of the American Mathematical Society, 2 (1996), no. 1, 1-6.

[7] Ricardo Diaz, Sinai Robins. *The Ehrhart polynomial of a lattice polytope.* Annals of Mathematics, Second Series, 145 (1997), no. 3, 503-518.

[8] Eugène Ehrhart. *Sur les polyèdres rationnels homothétiques à n dimensions.* Comptes rendus hebdomadaires des séances de l'Académie des Sciences, 254 (1962), 616-618.

[9] Eugène Ehrhart. *Sur un problème de géométrie diophantienne linéaire. I. Polyèdres et réseaux.* Journal für die Reine und Angewandte Mathematik, 226 (1967), 1-29.

[10] Eugène Ehrhart. *Sur un problème de géométrie diophantienne linéaire. II. Systémes diophantiens linéaires.* Journal für die Reine und Angewandte Mathematik, 227 (1967), 25-49.

[11] Godfrey H. Hardy, John E. Littlewood. *Some problems of Diophantine approximation: The lattice-points of a right-angled triangle.* Proceedings of the London Mathematical Society, S2-20 (1921), no. 1, 15-36.

[12] Godfrey H. Hardy, John E. Littlewood. *Some problems of Diophantine approximation: The lattice-points of a right-angled triangle. (Second memoire.).* Abhandlungen aus dem Mathematischen Seminar der Universitt Hamburg, 1 (1922), no. 1, 211-248.

[13] Martin N. Huxley. *Area, lattice points and exponential sums.* London Mathematical Society Monographs. New Series, 13. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1996. xii+494 pp. ISBN: 0-19-853466-3.

[14] Eugen J. Ionascu, Andrei Markov. *Platonic solids in $\mathbb{Z}^3$.* Journal of Number Theory, 131 (2011), no. 1, 138-145.

[15] Vojtěch Jarník. *Über Gitterpunkte in mehrdimensionalen Ellipsoiden.* Mathematische Annalen, 100 (1928), no. 1, 699-721.

[16] Ian Macdonald. *Polynomials associated with finite cell-complexes.* Journal of the London Mathematical Society, Second Series, 4 (1971), 181-192.

[17] Wolfgang Müller. *Lattice points in large convex bodies.* Monathshefte für Mathematik, 128 (1999), no. 4, 315-330.

[18] Werner G. Nowak. *The lattice point discrepancy of a torus in $\mathbb{R}^3$.* Acta Mathematica Hungarica, 120 (2008), no. 1-2, 179-192.

[19] Mark A. Pinsky. *Introduction to Fourier analysis and wavelets.* Graduate Studies in Mathematics, 102. American Mathematical Society, Providence, RI, 2009. xx+376 pp. ISBN: 978-0-8218-4797-8.

[20] James E. Pommersheim. *Toric varieties, lattice points and Dedekind sums.* Mathematische Annalen, 295 (1993), no. 1, 1-24.

[21] Wolfgang Schmidt. *Simultaneous approximation to algebraic numbers by rationals.* Acta Mathematica, 125 (1970), 189-201.