

Chapter 2

Expectation, and Its Connection with Quadratic Fields

2.1 Computing the Expectation in General (I)

The diophantine sum

$$S_\alpha(n) = \sum_{k=1}^n \left(\{k\alpha\} - \frac{1}{2} \right) \quad (2.1)$$

introduced in Sect. 1.2 [see (1.43)] is highly irregular as $n \rightarrow \infty$, but its mean value

$$M_\alpha(N) = \frac{1}{N} \sum_{n=1}^N S_\alpha(n) \quad (2.2)$$

exhibits a particularly simple and elegant asymptotic behavior for quadratic irrationals.

Let

$$\alpha = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \dots}} = [a_0; a_1, a_2, a_3, \dots] \quad (2.3)$$

denote the continued fraction for α ; a_i denote the partial quotients and $[a_0; a_1, \dots, a_{j-1}] = p_j/q_j$ is the j th convergent. By using (2.3) we can formulate

Proposition 2.1. *For any irrational $\alpha > 0$ given with (2.3) and any integer $N \geq 1$,*

$$M_\alpha(N) = \frac{-a_1 + a_2 - a_3 \pm \dots + (-1)^k a_k}{12} + O\left(\max_{1 \leq j \leq k} a_j\right), \quad (2.4)$$

where $k = k(\alpha, N)$ is the last index j for which the j th convergent denominator $q_j \leq N$, i.e., $q_k \leq N < q_{k+1}$, and the implicit constant on the right-hand side of (2.4) is absolute (less than 10).

Proposition 2.1 is particularly useful for quadratic irrationals. Indeed, for a periodic sequence a_i it is easy to evaluate the alternating sum in (2.4). As an illustration, consider first

$$\alpha = \sqrt{3} = [1; 1, 2, 1, 2, 1, 2, \dots] = [1; \overline{1, 2}]. \quad (2.5)$$

The least solution of Pell's equation $x^2 - 3y^2 = 1$ is $x = 2$, $y = 1$, and so

$$p_{2j} \pm q_{2j}\sqrt{3} = (2 \pm \sqrt{3})^j, \quad j = 1, 2, 3, \dots \quad (2.6)$$

where p_{2j}/q_{2j} is the $2j$ th convergent of $\sqrt{3}$ (we get every second convergent in (2.6), because the length of the period of $\sqrt{3}$ is 2 [see (2.5)]. By (2.6)

$$q_{2j} = \frac{1}{2\sqrt{2}} \left((2 + \sqrt{3})^j - (2 - \sqrt{3})^j \right),$$

and so we have

$$N = q_{2j} \implies j = \frac{\log N}{\log(2 + \sqrt{3})} + O(1). \quad (2.7)$$

Combining (2.4) with (2.7), for $\alpha = \sqrt{3}$ we have with $k = 2j$

$$\begin{aligned} M_{\sqrt{3}}(N) &= \frac{-a_1 + a_2 - a_3 \pm \dots + (-1)^k a_k}{12} + O(1) = \\ &= \frac{-1 + 2 - 1 + 2 \mp \dots - 1 + 2}{12} + O(1) = \frac{-1 + 2}{12} \cdot \frac{\log N}{\log(2 + \sqrt{3})} + O(1) = \\ &= \frac{\log N}{12 \log(2 + \sqrt{3})} + O(1), \end{aligned} \quad (2.8)$$

proving our claim in (1.53).

Here are two more examples like (2.8): for $\sqrt{7} = [2; \overline{1, 1, 1, 4}]$ the least solution $x = 8$, $y = 3$ of Pell's equation $x^2 - 7y^2 = 1$ comes from the fourth convergent $[2; 1, 1, 1] = 8/3$ of $\sqrt{7}$, and so

$$\begin{aligned} M_{\sqrt{7}}(N) &= \frac{-1 + 1 - 1 + 4}{12} \cdot \frac{\log N}{\log(8 + 3\sqrt{7})} + O(1) = \\ &= \frac{\log N}{4 \log(8 + 3\sqrt{7})} + O(1), \end{aligned}$$

and for $\sqrt{67} = [8; \overline{5, 2, 1, 1, 7, 1, 1, 2, 5, 16}]$ the least solution $x = 48,842$, $y = 5,967$ of Pell's equation $x^2 - 67y^2 = 1$ comes from the tenth convergent $[8; 5, 2, 1, 1, 7, 1, 1, 2, 5] = 48842/5967$ of $\sqrt{67}$, and so

$$M_{\sqrt{67}}(N) = \frac{-5 + 2 - 1 + 1 - 7 + 1 - 1 + 2 - 5 + 16}{12} \cdot \frac{\log N}{\log(48842 + 5967\sqrt{67})} \\ + O(1) = \frac{\log N}{4 \log(48842 + 5967\sqrt{67})} + O(1).$$

In sharp contrast, for $\alpha = \sqrt{2} = [1; \overline{2}]$ the alternating sum in (2.4) *cancels out*, and $M_{\sqrt{2}}(N) = O(1)$; this proves (1.52).

Similarly, any quadratic irrational α , for which the length of the period (of the continued fraction) is *odd*, has the property that the mean value is basically zero: $M_\alpha(N) = O(1) = O_\alpha(1)$ (because the alternating sum in (2.4) cancels out). Note that in Sect. 1.5 we proved the fact $M_\alpha(N) = O(1)$ in the special case of the golden ratio

$$\alpha = (\sqrt{5} - 1)/2 = [1, 1, 1, 1, \dots] = [\overline{1}]$$

by a long, direct computation; see (1.177). This direct computation becomes hopelessly messy even for an arbitrary quadratic irrational, not to mention the general case of an arbitrary irrational number.

Unfortunately, we cannot characterize the quadratic irrationals for which the period is odd/even (what we mean here is that the *length* of period in the continued fraction is odd or even). However, if $\alpha = \sqrt{p}$ where p is an odd prime, we have a perfect characterization: the period is odd if $p \equiv 1 \pmod{4}$, and the period is even if $p \equiv 3 \pmod{4}$.

The proof of this elegant characterization is based on the well-known number-theoretic fact that the “negative” Pell equation $x^2 - dy^2 = -1$ (where $d > 0$ is an integer, but not a complete square) has an integral solution if and only if the period of \sqrt{d} is odd. If p is a prime with $p \equiv 1 \pmod{4}$, then we will *find* an integral solution of $x^2 - py^2 = -1$, and this will imply that the period of \sqrt{p} is odd. To find a solution of $x^2 - py^2 = -1$, we start with the fundamental solution (x_1, y_1) of the ordinary Pell's equation $x^2 - py^2 = 1$, which always has a solution (the fundamental solution is the least positive solution). The equation $x^2 - 1 = py^2$ leads to the factorization

$$(x_1 - 1)(x_1 + 1) = py_1^2. \quad (2.9)$$

If $p \equiv 1 \pmod{4}$ then (2.9) implies that x_1 is odd, and also by using that p is a prime, we have either (1) $x_1 - 1 = 2pu^2$ and $x_1 + 1 = 2v^2$ or (2) $x_1 + 1 = 2pu^2$ and $x_1 - 1 = 2v^2$ holds for some positive integers u and v satisfying $y_1 = 2uv$. Hence $v^2 - pu^2 = \pm 1$. The case $v^2 - pu^2 = 1$ is impossible, since (v, u) is a smaller solution

than (x_1, y_1) , the fundamental solution—a contradiction. Thus $v^2 - pu^2 = -1$, i.e., the negative Pell's equation *does* have a solution, and we obtain the following.

Corollary 2.2. *If p is a prime with $p \equiv 1 \pmod{4}$ then*

$$M_{\sqrt{p}}(N) = O(1).$$

The proof above is prime specific: if $d \equiv 1 \pmod{4}$ is not a prime, then the length of the period of \sqrt{d} can be both even and odd. For example, $\sqrt{21} = [4; \overline{1, 1, 2, 1, 1, 8}]$ gives length 6 (even) and $\sqrt{65} = [8; \overline{16}]$ gives length 1 (odd).

On the other hand, if $d \equiv 3 \pmod{4}$, then by a simple $\pmod{4}$ analysis we have $x^2 - dy^2 \not\equiv -1 \pmod{4}$ (it is irrelevant that d is a prime or not), implying that the length of the period of \sqrt{d} has to be even.

Actually, we have a stronger result: if d has a prime factor $q \equiv 3 \pmod{4}$, the period of \sqrt{d} is always even. Indeed, then $x^2 - dy^2 = -1$ implies $x^2 \equiv -1 \pmod{q}$, which contradicts Fermat's little theorem:

$$1 \equiv x^{q-1} = (x^2)^{(q-1)/2} \equiv (-1)^{(q-1)/2} = -1 \pmod{q}.$$

What happens in Proposition 2.1 if we go beyond quadratic irrationals? How about the special number e :

$$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, \dots, 1, 2i, 1, \dots]?$$

Well, the alternating sum $(-1 + 2 - 1) + (1 - 4 + 1) + (-1 + 6 - 1) + \dots + (-1)^i (1 - 2i + 1)$ equals $i - 1$ if i is odd and $-i$ if i is even. Thus by Proposition 2.1 we have

$$M_e(N) = O(\log N / \log \log N), \quad (2.10)$$

which is the true order of magnitude.

Note in advance that Proposition 2.1 also gives the constant factor $C_1(\alpha, x)$ in Theorem 1.1 in the special case $x = 1/2$. It is a consequence of the identity

$$\chi_{1/2}(y) - \frac{1}{2} = \left(\{2y\} - \frac{1}{2} \right) - 2 \left(\{y\} - \frac{1}{2} \right),$$

where of course $\{y\}$ denotes the fractional part of y , and $\chi_{1/2}(y)$ is 1 if $\{y\} < 1/2$ and 0 otherwise. We will return to this later in Sect. 2.2; see (2.87) and (2.88).

2.1.1 An Important Detour: How to Guess Proposition 2.1?

The proof of Proposition 2.1 is not easy, but it was equally difficult to *find* the right conjecture. What was our motivation to guess formula (2.4)? Well, this is an

interesting long story, which involves algebraic number theory. To explain it, we briefly outline an alternative approach to find the average $M_\alpha(N)$. We start with the well-known Fourier series expansion of the fractional part function (warning: it is not absolutely convergent)

$$\{x\} = \frac{1}{2} - \sum_{n=1}^{\infty} \frac{\sin(2\pi nx)}{\pi n}. \quad (2.11)$$

Substituting it back to (2.1) and (2.2), after some long but standard manipulations we end up with

$$M_\alpha(N) = -\frac{1}{2\pi} \sum_{n=1}^N \frac{1}{n \tan(\pi n\alpha)} + O(1), \quad (2.12)$$

if $a_i = O(1)$, i.e., the partial quotients of α are bounded (this is certainly true for the quadratic irrationals). (Note that Eq. (2.12) is exactly our Proposition 2.16 coming later.)

Let $\alpha = \sqrt{d}$, where $d \equiv 3 \pmod{4}$ is a positive square-free integer. We clearly have (m denotes the nearest integer to $n\sqrt{d}$)

$$\frac{1}{\pi} \tan(\pi n\sqrt{d}) \approx \pm \|n\sqrt{d}\| = n\sqrt{d} - m \approx \frac{-(m^2 - dn^2)}{2n\sqrt{d}}. \quad (2.13)$$

In view of (2.12) and (2.13), the following formula is not too surprising:

$$M_{\sqrt{d}}(N) = \frac{\sqrt{d}}{\pi^2} \left(\sum_{\substack{(x,y) \neq (0,0): \\ \text{primary representations}}} \frac{1}{x^2 - dy^2} \right) \frac{\log N}{\log \eta_d} + O((\log \log N)^3), \quad (2.14)$$

where η_d is the fundamental unit of $\mathbf{Q}(\sqrt{d})$. Note that Eq. (2.14) is exactly Proposition 2.20; the meaning of “primary representations” will be explained later at the beginning of Sect. 2.6—actually the reader can jump ahead and read it right now.

If $d \equiv 3 \pmod{4}$ then $x^2 - dy^2$ is the norm of the algebraic integer $x + y\sqrt{d}$ in the real quadratic field $\mathbf{Q}(\sqrt{d})$.

2.1.2 Quadratic Fields in a Nutshell

Let D be a square-free positive or negative integer, and consider the quadratic field $\mathbf{Q}(\sqrt{D})$. The *discriminant* Δ of $\mathbf{Q}(\sqrt{D})$ is $4D$ if $D \equiv 2$ or $3 \pmod{4}$, and D if $D \equiv 1 \pmod{4}$. The quadratic irrational $(a + b\sqrt{D})/2$ is an *algebraic integer* in

$\mathbf{Q}(\sqrt{D})$ iff a and $b \in \mathbb{Z}$ are integers satisfying $a \equiv b \equiv 0 \pmod{2}$ when $D \equiv 2$ or $3 \pmod{4}$, and $a \equiv b \pmod{2}$ when $D \equiv 1 \pmod{4}$. So the *norm*

$$\frac{a + b\sqrt{D}}{2} \cdot \frac{a - b\sqrt{D}}{2} = \frac{a^2 - b^2D}{4}$$

of $(a + b\sqrt{D})/2$ is always an integer. An algebraic integer in $\mathbf{Q}(\sqrt{D})$ is called a *unit* if its norm is ± 1 . If $D > 0$, then there exists a unit $\eta = \eta_D$ in $\mathbf{Q}(\sqrt{D})$ such that any unit in $\mathbf{Q}(\sqrt{D})$ is representable as $\pm \eta^n$, $n = 0, \pm 1, \pm 2, \dots$. This number $\eta = \eta_D$ is called the *fundamental unit* in $\mathbf{Q}(\sqrt{D})$.

Let $F(x, y) = ax^2 + bxy + cy^2$ be an integral binary quadratic form of discriminant $\Delta = b^2 - 4ac$ ($a, b, c \in \mathbb{Z}$ are integers). If an integral binary quadratic form $F(x, y)$ is transformed into the form $F_1(x_1, y_1)$ by an integral unimodular transformation $x = Ux_1 + Vy_1$, $y = Wx_1 + Zy_1$ where $UZ - VW = 1$, then F and F_1 are called *equivalent*. The *class number* $h(D)$ (where $\Delta = 4D$ or D) is basically the number of nonequivalent integral binary quadratic forms of discriminant Δ . More precisely, by computing the class number we do not distinguish a quadratic form from its negative, though they may be nonequivalent (which is exactly the case if $D > 0$, and $x^2 - Dy^2 = -1$ does not have an integer solution). For example, let $D = 79$, then the discriminant is $4 \cdot 79 = 316$, and there are six nonequivalent integral binary forms of discriminant 316: $F_1 = x^2 - 79y^2$, $-F_1 = -x^2 + 79y^2$, $F_2 = 3x^2 + 4xy - 25y^2$, $-F_2 = -3x^2 - 4xy + 25y^2$, $F_3 = 3x^2 + 2xy - 26y^2$, $-F_3 = -3x^2 - 2xy + 26y^2$. So the class number $h(79)$ of the quadratic field $\mathbf{Q}(\sqrt{79})$ is 3 (and not 6). If $h(D) = 1$ then the algebraic integers in $\mathbf{Q}(\sqrt{D})$ have *unique factorization into algebraic primes*. The “first” quadratic field with class number > 1 is $\mathbf{Q}(\sqrt{-5})$. The discriminant is $4 \cdot (-5) = -20$, and there are two nonequivalent integral binary quadratic forms of discriminant -20 : $x^2 + 5y^2$ and $2x^2 + 2xy + 3y^2$. So the class number $h(-5)$ is 2. A counterexample to the unique prime factorization is

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

where all the 4 factors $(1 + \sqrt{-5})$, $(1 - \sqrt{-5})$, 2, and 3 are primes in the ring of integers of $\mathbf{Q}(\sqrt{-5})$.

Now let us return to (2.14). If we make the extra hypothesis that $d = p \equiv 3 \pmod{4}$ is a prime and the class number $h(p)$ of the real quadratic field $\mathbf{Q}(\sqrt{p})$ is one, then the middle sum on the right-hand side of (2.14) becomes a special L-function at $s = 1$:

$$\sum_{\substack{(x,y) \neq (0,0): \\ \text{primary representations}}} \frac{1}{x^2 - py^2} = L(1, \chi^*). \quad (2.15)$$

Here χ^* is the so-called norm-sign character: a unique character with values ± 1 defined for all ideals in the ring of the algebraic integers of $\mathbf{Q}(\sqrt{d})$ (in fact, χ^*

depends only on the narrow ideal class), and satisfies $\chi^*((a)) = \text{sign Norm}(a)$ for the principal ideals (a) . Note that, in our special case $d = p$ with $h(p) = 1$, every ideal is principal.

The L-function

$$L(s, \chi^*) = \sum_{A: \text{ideals}} \frac{\chi^*(A)}{\text{Norm}(A)^s}$$

(here we don't have to write $|\text{Norm}(A)|$, because the norm of an ideal is by definition an integer ≥ 1 ; in sharp contrast the norm of an algebraic integer in a real field can be both positive and negative) has the product decomposition

$$L(s, \chi^*) = L(s, \chi_{-4})L(s, \chi_{-p}) \quad (2.16)$$

where

$$L(s, \chi_{-4}) = \sum_{n=1}^{\infty} \frac{\chi_{-4}(n)}{n^s} \quad \text{and} \quad L(s, \chi_{-p}) = \sum_{n=1}^{\infty} \frac{\chi_{-p}(n)}{n^s}$$

are the (ordinary) L-functions of the complex quadratic fields $\mathbf{Q}(\sqrt{-4}) = \mathbf{Q}(\sqrt{-1})$ ("Gauss integers") and $\mathbf{Q}(\sqrt{-p})$; the characters χ_{-4} and χ_{-p} are defined as follows: $\chi_{-4}(n) = \pm 1$ if $n \equiv \pm 1 \pmod{4}$ and $\chi_{-4}(n) = 0$ if n is even, and

$$\chi_{-p}(n) = \left(\frac{n}{p} \right)$$

is the usual Legendre symbol (quadratic residue symbol). Note that (2.16) is basically an Euler product, and it is "explained" by the elementary factorization $4p = (-4)(-p)$ of the discriminant of $x^2 - py^2$; see, e.g., Zagier's book [Za4].

In the special case $s = 1$ Eq. (2.16) gives

$$L(1, \chi^*) = L(1, \chi_{-4})L(1, \chi_{-p}), \quad (2.17')$$

and by Dirichlet's (analytic) class number formula,

$$L(1, \chi_{-4}) = \frac{\pi}{4} \quad \text{and} \quad L(1, \chi_{-p}) = \frac{\pi h(-p)}{\sqrt{p}}, \quad (2.17'')$$

if $p > 3$. Now this is where the remarkable Hirzebruch–Meyer–Zagier formula (HMZ-formula, in short) enters the story: $h(-p)$ can be expressed in terms of an alternating sum of the partial quotients (i.e., the "digits" of the continued fraction) in the period of \sqrt{p} ; see, e.g., in Zagier [Za1].

But before formulating the HMZ-formula, we note that quadratic irrationals all have periodic continued fraction, and the least solution of Pell's equation

$x^2 - dy^2 = 1$ can be determined from the period of \sqrt{d} ; the least solution is basically the fundamental unit. Moreover, the *parity* of the length of the period describes the sign of the norm of the fundamental unit: odd length means $+1$, even length means -1 . Combining Dirichlet's class number formulas with the *ineffective* Siegel theorem, we obtain the deep asymptotic formulas

$$h(d) \log \eta_d = d^{1/2 \pm \varepsilon}, \quad (2.18')$$

$$h(-d) = d^{1/2 \pm \varepsilon}, \quad (2.18'')$$

where $h(d)$ and $h(-d)$ are the class numbers of the real and complex quadratic fields $\mathbf{Q}(\sqrt{d})$ and $\mathbf{Q}(\sqrt{-d})$, respectively, η_d is the fundamental unit of $\mathbf{Q}(\sqrt{d})$, and $\varepsilon > 0$ is arbitrarily small but fixed. Note that the order of magnitude of $\log \eta_d$ is roughly around the *length* of the period of the continued fraction for \sqrt{d} .

The elegant Hirzebruch-Meyer-Zagier formula (HMZ-formula) was discovered in the 1970s. It states that

$$h(-p) = \frac{-a_1 + a_2 - a_3 \pm \cdots + a_{2s}}{3}, \quad (2.19)$$

where $p \equiv 3 \pmod{4}$ is a prime > 3 , $h(p) = 1$, and a_1, a_2, \dots, a_{2s} forms the period of \sqrt{p} (since $p \equiv 3 \pmod{4}$, the length of the period has to be even). (Note that both (2.17) and (2.19) fail for $p = 3$, because $\mathbf{Q}(\sqrt{-3})$ has too many automorphisms: 6 instead of the usual 2—a technical nuisance in algebraic number theory.)

Combining the HMZ-formula with (2.14)–(2.17), we conclude

$$\begin{aligned} M_{\sqrt{p}}(N) &= \frac{h(-p)}{4} \cdot \frac{\log N}{\log \eta} + O((\log \log N)^3) = \\ &= \frac{-a_1 + a_2 \mp \cdots + a_{2s}}{12} \cdot \frac{\log N}{\log \eta} + O((\log \log N)^3) = \\ &= \frac{-a_1 + a_2 - a_3 \pm \cdots + (-1)^\ell a_\ell}{12} + O((\log \log N)^3), \end{aligned} \quad (2.20)$$

where ℓ is the last index for which $q_\ell \leq N$ and η is the fundamental unit of $\mathbf{Q}(\sqrt{p})$ (in the last equation we heavily used the periodicity of the continued fraction for \sqrt{p}).

Summarizing, by using the HMZ-formula, we just managed to prove (2.20), at least under some strong technical conditions (for example, we assumed that $p \equiv 3 \pmod{4}$ is a prime > 3 with $h(p) = 1$, and also in (2.20) we have the ugly but negligible error term $O((\log \log N)^3)$). Nevertheless, from (2.20) it was quite easy to guess that Proposition 2.1 must hold for *arbitrary* α (not just for quadratic irrationals), and this is exactly how we came up with the right conjecture (2.4).

Because we know a completely elementary proof of Proposition 2.1, reversing the argument, we can produce an elementary proof for the HMZ-formula. Later we will give a precise proof of (2.12) and (2.14); (2.12) is Proposition 2.16 and (2.14) is Proposition 2.20.

(The interested reader can find all the details, and much more, about quadratic fields in the well-written book of Zagier [Za4] (it is in German), or in the classic Borevich–Safarevich: Number Theory.)

2.1.3 Another Detour: Formulating a “Positivity Conjecture”

The first line in (2.20) raises a very interesting question. If a prime p satisfies the condition of the HMZ-formula, the expectation equals

$$M_{\sqrt{p}}(N) = \frac{h(-p)}{4} \cdot \frac{\log N}{\log \eta} + \text{negligible error.}$$

Here the class number is trivially ≥ 1 , and also $\eta \geq \sqrt{p} > 1$, implying $\log \eta > 0$; therefore,

$$M_{\sqrt{p}}(N) = c \cdot \log N + \text{negligible error,}$$

where $c = c(p) > 0$ is a positive constant. By Proposition 2.1, the error term here is in fact $O(1)$, and in general, for any quadratic irrational α ,

$$M_{\alpha}(N) = c \cdot \log N + O(1),$$

where $c = c(\alpha)$ is a constant (expressed in terms of the period of α). Is it true that if $\alpha = \sqrt{d}$, the corresponding constant factor is always nonnegative, that is, $M_{\sqrt{d}}(N) = c \cdot \log N + O(1)$ with $c \geq 0$? We guess the answer is “yes,” and I refer to this as the “positivity conjecture.”

If the length of the period of \sqrt{d} is odd, the “positivity conjecture” is trivial. Indeed, by formula (2.4) the corresponding alternating sum “cancels out,” implying that the constant factor is zero, i.e., $M_{\sqrt{d}}(N) = O(1)$ (the same holds for any quadratic irrational with odd period). Thus, the nontrivial case is when the length of the period of \sqrt{d} is even. It is well known that then the period has the symmetric form with a central term

$$\sqrt{d} = [a_0; \overline{a_1, a_2, \dots, a_t, a_{t+1}, a_t, \dots, a_2, a_1, 2a_0}]$$

where $a_0 = \lfloor \sqrt{d} \rfloor$ and a_{t+1} denotes the central term. Applying the alternating sum in formula (2.4), we have

$$\begin{aligned}
M_{\sqrt{d}} &= \frac{-a_1 + a_2 - a_3 \pm \cdots}{12} + O(1) = \\
&= \left(2 \left(\sum_{j=1}^t (-1)^j a_j \right) + (-1)^{t+1} a_{t+1} + 2a_0 \right) \cdot \frac{\log N}{\log \eta} + O(1).
\end{aligned}$$

The positivity of the constant factor $c = c(d)$ in $M_{\sqrt{d}} = c \log N + O(1)$ is, therefore, equivalent to the positivity of the alternating sum formed from the period

$$2 \sum_{j=0}^t (-1)^j a_j + (-1)^{t+1} a_{t+1} > 0.$$

We checked the tables for $d < 100$, and this alternating sum is indeed positive when the period of \sqrt{d} is even. Since the “positivity conjecture” is certainly not true for arbitrary quadratic irrational α , its hypothetical truth in the special case $\alpha = \sqrt{d}$ is probably closely related to the arithmetic of the real quadratic field $\mathbf{Q}(\sqrt{d})$ (or perhaps the complex field $\mathbf{Q}(\sqrt{-d})$).

Let’s return now to Proposition 2.1. We include an elementary (but far from easy) proof.

Proof of Proposition 2.1. We use Dedekind sums. To explain where the Dedekind sum comes from, we rewrite (2.1) and (2.2) in the following form:

$$\begin{aligned}
M_\alpha(N) &= \frac{1}{N} \sum_{k=1}^N (N+1-k) \left(\{k\alpha\} - \frac{1}{2} \right) = \\
&= \left(\frac{N+1}{N} - \frac{1}{2} \right) \sum_{k=1}^N \left(\{k\alpha\} - \frac{1}{2} \right) - \sum_{k=1}^N \left(\frac{k}{N} - \frac{1}{2} \right) \left(\{k\alpha\} - \frac{1}{2} \right), \quad (2.21)
\end{aligned}$$

where the last sum

$$\sum_{k=1}^N \left(\frac{k}{N} - \frac{1}{2} \right) \left(\{k\alpha\} - \frac{1}{2} \right)$$

in (2.21) strongly resembles a Dedekind sum

$$D(H, K) = \sum_{j=1}^{K-1} \left(\frac{j}{K} - \frac{1}{2} \right) \left(\{jH/K\} - \frac{1}{2} \right), \quad (2.22)$$

where we always assume that H and $K \geq 1$ are relatively prime integers.

Dedekind sums [i.e., (2.22)] originally appeared in Dedekind's study of elliptic functions and theta-functions. Luckily we don't need to know anything about these (rather technical) subjects; we can just work with definition (2.22). The key fact about Dedekind sums is the following reciprocity formula, a highly surprising and nontrivial result.

Lemma 2.3 (Dedekind's reciprocity formula). *We have*

$$D(H, K) + D(K, H) = \frac{1}{12} \left(\frac{H}{K} + \frac{K}{H} + \frac{1}{HK} \right) - \frac{1}{4}. \quad (2.23)$$

Note that the definition of $D(H, K)$ and $D(K, H)$ automatically includes the condition that " $H \geq 1$ and $K \geq 1$ are relatively prime integers."

For a proof of this classical result, see, e.g., the book [Ra-Gr].

From Lemma 2.3 we will derive

Lemma 2.4. *If $1 \leq H < K$ are relatively prime then*

$$D(H, K) = \frac{a_1 - a_2 + a_3 \mp \cdots + (-1)^{\ell-1} a_\ell}{12} + O(1), \quad (2.24)$$

where

$$\frac{H}{K} = \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \dots}}} = [a_1, a_2, a_3, \dots, a_\ell]. \quad (2.25)$$

Note that the error term $O(1)$ in (2.24) has absolute value $\leq 1/4$.

Proof. The continued fraction $\frac{H}{K} = [a_1, a_2, a_3, \dots, a_\ell]$ is equivalent to the Euclidean algorithm

$$K = a_1 H + H_1, \quad H = a_2 H_1 + H_2, \quad H_1 = a_3 H_2 + H_3, \dots, \quad H_{\ell-2} = a_\ell H_{\ell-1}$$

where $H_{\ell-1} = \gcd(H, K) = 1$ (\gcd denotes the greatest common divisor). We apply Lemma 2.3 with the short notation

$$g(x, y) = \frac{1}{12} \left(\frac{x}{y} + \frac{y}{x} + \frac{1}{xy} \right) - \frac{1}{4}$$

as follows: write $K = H_{-1}$, $H = H_0$, then

$$\begin{aligned} D(H, K) &= D(H_0, H_{-1}) = g(H_{-1}, H_0) - D(H_{-1}, H_0) = \\ &= g(H_{-1}, H_0) - D(H_1, H_0); \end{aligned}$$

here we used the first equation of the Euclidean algorithm. Repeating the same argument, we have

$$\begin{aligned} D(H, K) &= g(H_{-1}, H_0) - D(H_1, H_0) = \\ &= g(H_{-1}, H_0) - (g(H_0, H_1) - D(H_0, H_1)) = \\ &= g(H_{-1}, H_0) - g(H_0, H_1) + D(H_2, H_1); \end{aligned}$$

here we used the second equation of the Euclidean algorithm.

Repeating the same argument several times, we have

$$\begin{aligned} D(H, K) &= g(H_{-1}, H_0) - g(H_0, H_1) + g(H_1, H_2) - g(H_2, H_3) \pm \cdots \\ &\quad \cdots + (-1)^{\ell-1} g(H_{\ell-2}, H_{\ell-1}) + (-1)^\ell D(H_{\ell-2}, H_{\ell-1}). \end{aligned}$$

Note that the last term here is in fact zero; indeed, $H_{\ell-1} = \gcd(H, K) = 1$ implies that $D(H_{\ell-2}, H_{\ell-1}) = 0$.

Moreover, by using the notation

$$f(x, y) = \frac{x}{y} + \frac{y}{x},$$

we have

$$\begin{aligned} \sum_{i=0}^{\ell-1} (-1)^i f(H_{i-1}, H_i) &= \sum_{i=0}^{\ell-1} (-1)^i \left(\frac{H_{i-1}}{H_i} + \frac{H_i}{H_{i-1}} \right) = \\ &= \frac{H_0}{H_{-1}} + \sum_{i=0}^{\ell-1} (-1)^i \frac{H_{i-1} - H_{i+1}}{H_i} = \\ &= \frac{H}{K} + \sum_{i=0}^{\ell-1} (-1)^i \frac{a_{i-1} H_i}{H_i} = \frac{H}{K} + \sum_{i=0}^{\ell-1} (-1)^i a_{i-1}. \end{aligned}$$

Since

$$g(x, y) = \frac{1}{12} f(x, y) + \left(\frac{1}{12xy} - \frac{1}{4} \right),$$

combining the facts above, we conclude

$$\begin{aligned}
 D(H, K) &= g(H_{-1}, H_0) - g(H_0, H_1) + g(H_1, H_2) - g(H_2, H_3) \pm \cdots + \\
 &\quad + (-1)^{\ell-1} g(H_{\ell-2}, H_{\ell-1}) = \frac{a_1 - a_2 + a_3 \mp \cdots + (-1)^{\ell-1} a_\ell}{12} + \\
 &\quad + \frac{H}{12K} - \frac{1 + (-1)^{\ell-1}}{8} + \frac{1}{12} \left(\frac{1}{KH} - \frac{1}{HH_1} + \frac{1}{H_1H_2} \mp \cdots + \frac{(-1)^{\ell-1}}{H_{\ell-2}H_{\ell-1}} \right).
 \end{aligned}$$

The last alternating sum has absolute value $\leq 1/12$, and because $1 \leq H < K$, the total error is at most $\max\{1/4, 1/12 + 1/12\} = 1/4$, completing the deduction of Lemma 2.4 from Lemma 2.3. \square

Next we derive Proposition 2.1 from Lemma 2.4 in the special case $N = q_r$, i.e., when N happens to be a convergent denominator of α ; see Lemma 2.5. But first we introduce a notation that simplifies the treatment of Dedekind sums. Let

$$((x)) = \begin{cases} \{x\} - \frac{1}{2}, & \text{if } x \text{ is not an integer;} \\ 0, & \text{otherwise.} \end{cases}$$

Note that $y = ((x))$ is usually called the “sawtooth function.” By using this new notation, we can rewrite (2.22) in a shorter form:

$$D(H, K) = \sum_{j=1}^{K-1} \left(\left(\frac{j}{K} \right) \right) \left(\left(\frac{jH}{K} \right) \right), \quad (2.26)$$

where, as usual, we assume that H and $K \geq 1$ are relatively prime integers. Notice that extending the summation in (2.26) from 1 to K makes no difference (just adds a zero to the sum).

Now we are ready to formulate and prove an important special case of Proposition 2.1.

Lemma 2.5. *We have*

$$M_\alpha(q_r) = \frac{-a_1 + a_2 - a_3 \pm \cdots + (-1)^{r-1} a_{r-1}}{12} + O(1), \quad (2.27)$$

where $\alpha = [a_1, a_2, a_3, \dots]$ and $p_r/q_r = [a_1, a_2, \dots, a_{r-1}]$ is the r th convergent of α . The implicit error term $O(1)$ is less than 5 for all α and r .

Proof. We recall (2.21) with $N = q_r$:

$$M_\alpha(q_r) = \left(\frac{q_r + 1}{q_r} - \frac{1}{2} \right) \sum_{k=1}^{q_r} \left(\{k\alpha\} - \frac{1}{2} \right) - \sum_{k=1}^{q_r} \left(\frac{k}{q_r} - \frac{1}{2} \right) \left(\{k\alpha\} - \frac{1}{2} \right). \quad (2.28)$$

First we focus on the following subsum of (2.28):

$$S^* = \sum_{k=1}^{q_r} \left(\frac{k}{q_r} - \frac{1}{2} \right) \left(\{k\alpha\} - \frac{1}{2} \right) = \sum_{k=1}^{q_r} \left(\left(\frac{k}{q_r} \right) \right) ((k\alpha)). \quad (2.29)$$

We compare S^* to the Dedekind sum

$$D(p_r, q_r) = \sum_{k=1}^{q_r} \left(\left(\frac{k}{q_r} \right) \right) \left(\left(\frac{kp_r}{q_r} \right) \right), \quad (2.30)$$

where p_r/q_r is the r th convergent of α .

We recall the well-known fact from diophantine approximation that

$$\left| \alpha - \frac{p_r}{q_r} \right| < \frac{1}{q_r^2},$$

which implies that the inequality

$$\left| k\alpha - \frac{kp_r}{q_r} \right| < \frac{k}{q_r^2} \leq \frac{1}{q_r} \quad (2.31)$$

holds for all $1 \leq k \leq q_r$. By (2.31) we have

$$|S^* - D(p_r, q_r)| < 1. \quad (2.32)$$

On the other hand, by Lemma 2.4,

$$\left| D(p_r, q_r) - \frac{a_1 - a_2 + a_3 \mp \cdots + (-1)^r a_{r-1}}{12} \right| \leq \frac{1}{4}. \quad (2.33)$$

Combining (2.32) and (2.33) we have

$$\left| S^* - \frac{a_1 - a_2 + a_3 \mp \cdots + (-1)^r a_{r-1}}{12} \right| \leq \frac{1}{4} + 1 = \frac{5}{4}. \quad (2.34)$$

Another application of (2.31) gives

$$\begin{aligned} \left| \sum_{k=1}^{q_r-1} (\{k\alpha\} - 1/2) \right| &\leq \left| \sum_{j=1}^{q_r-1} \left(\frac{j}{q_r} \pm \frac{1}{q_r} - \frac{1}{2} \right) \right| \leq \\ &\leq \left| \sum_{j=1}^{q_r-1} \left(\frac{j}{q_r} - \frac{1}{2} \right) \right| + q_r \frac{1}{q_r} = 0 + 1 = 1. \end{aligned} \quad (2.35)$$

Applying (2.34) and (2.35) in (2.28), we conclude that

$$\begin{aligned} &\left| M_\alpha(q_r) - \frac{a_1 - a_2 + a_3 \mp \dots + (-1)^r a_{r-1}}{12} \right| \leq \\ &\leq \frac{5}{4} + \left| \frac{q_r + 1}{q_r} - \frac{1}{2} \right| + \left| \frac{q_r + 1}{q_r} - \frac{1}{2} \right| \left| \{q_r \alpha\} - \frac{1}{2} \right| \leq \frac{5}{4} + 2 \left| \frac{q_r + 1}{q_r} - \frac{1}{2} \right| < 5, \end{aligned}$$

and Lemma 2.5 follows. \square

The last step is to derive the general Proposition 2.1 from the special case Lemma 2.5. There are many ways to reduce the general case to Lemma 2.5; see, e.g., Beck [Be4]. Here we follow a nice idea of Schoissengeier [Scho], involving telescoping sums, which seems to be the best treatment of the general case.

Let $N \geq 1$ be an arbitrary integer. Consider the Ostrowski expansion of N [see (1.54)]:

$$N = \sum_{i=1}^r b_i q_i, \text{ where } 0 \leq b_i \leq a_i \text{ and} \quad (2.36)$$

$b_i = a_i$ implies $b_{i-1} = 0$ ("Extra Rule"). Here a_i is the i th partial quotient of the continued fraction of $\alpha = [a_1, a_2, a_3, \dots]$ and $p_i/q_i = [a_1, \dots, a_{i-1}]$ is the i th convergent of α .

We are motivated by the following telescoping sum equation:

$$\begin{aligned} &\sum_{i=1}^N \frac{N+1-i}{N} \left(\left(\frac{ip_r}{q_r} \right) \right) = \\ &= \frac{1}{N} \sum_{k=1}^r \left(\sum_{i=1}^{N_k} (N_k + 1 - i) \left(\left(\frac{ip_k}{q_k} \right) \right) - \sum_{j=1}^{N_{k-1}} (N_{k-1} + 1 - j) \left(\left(\frac{jp_{k-1}}{q_{k-1}} \right) \right) \right), \end{aligned} \quad (2.37)$$

where N_k is the k th partial sum of (2.36): $N_k = \sum_{i=1}^k b_i q_i$.

We are going to evaluate the terms of the telescoping sum (2.37). The next lemma, clearly motivated by Eq. (2.37), can be considered as a generalization, or new version, of Lemma 2.5. The idea is to involve the Dedekind sum $D(p_k, q_k)$, just like we did in the proof of Lemma 2.5.

Lemma 2.6. *If $N_j = \sum_{i=1}^j b_i q_i$ then*

$$\begin{aligned} & \sum_{i=1}^{N_k} (N_k + 1 - i) \left(\left(\frac{ip_k}{q_k} \right) \right) - \sum_{j=1}^{N_{k-1}} (N_{k-1} + 1 - j) \left(\left(\frac{jp_{k-1}}{q_{k-1}} \right) \right) = \\ & = -b_k q_k D(p_k, q_k) + \frac{b_{k-1}}{4} (1 + (-1)^k) (2N_{k-1} + 1 - (b_{k-1} + 1)q_{k-1}) + \\ & \quad + (-1)^{k+1} \frac{N_{k-1}(N_{k-1} + 1)(N_{k-1} + 2)}{6q_k q_{k-1}}. \end{aligned} \quad (2.38)$$

Proof of Lemma 2.6. We basically repeat the proof of Lemma 2.5. Write

$$\sum_{i=1}^{N_k} (N_k + 1 - i) \left(\left(\frac{ip_k}{q_k} \right) \right) = \sum_1 + \sum_2, \quad (2.39)$$

where

$$\sum_1 = \sum_{i=1}^{b_k q_k} (N_k + 1 - i) \left(\left(\frac{ip_k}{q_k} \right) \right)$$

and

$$\sum_2 = \sum_{i=b_k q_k + 1}^{N_k} (N_k + 1 - i) \left(\left(\frac{ip_k}{q_k} \right) \right).$$

We evaluate \sum_1 first. Since $((x)) = 0$ if x is an integer, we take out the i 's that are divisible by q_k :

$$\begin{aligned} \sum_1 &= \sum_{t=0}^{b_k-1} \sum_{i=tq_k+1}^{(t+1)q_k-1} (N_k + 1 - i) \left(\left(\frac{ip_k}{q_k} \right) \right) = \\ &= \sum_{t=0}^{b_k-1} \sum_{j=1}^{q_k-1} (N_k + 1 - tq_k - j) \left(\left(\frac{jp_k}{q_k} \right) \right) = \\ &= -b_k \sum_{j=1}^{q_k-1} j \left(\left(\frac{jp_k}{q_k} \right) \right), \end{aligned} \quad (2.40)$$

since

$$\sum_{j=1}^{K-1} \left(\left(\frac{jH}{K} \right) \right) = 0.$$

Thus by (2.40),

$$\sum_1 = -b_k q_k \sum_{j=1}^{q_k-1} \left(\frac{j}{q_k} - \frac{1}{2} \right) \left(\left(\frac{j p_k}{q_k} \right) \right) = -b_k q_k D(p_k, q_k), \quad (2.41)$$

justifying the first term on the right-hand side of (2.38).

Next we evaluate $\sum_2 - \sum_3$, where \sum_2 is the second term in (2.39) and \sum_3 is the negative term on the left-hand side of (2.38):

$$\sum_3 = \sum_{j=1}^{N_{k-1}} (N_{k-1} + 1 - j) \left(\left(\frac{j p_{k-1}}{q_{k-1}} \right) \right). \quad (2.42)$$

We recall the well-known fact from the theory of continued fraction:

$$\frac{p_k}{q_k} = \frac{p_{k-1}}{q_{k-1}} + \frac{(-1)^{k-1}}{q_{k-1} q_k}, \quad (2.43)$$

and so, if $j \leq N_{k-1}$ then

$$\left(\left(\frac{j p_k}{q_k} \right) \right) = \left(\left(\frac{j p_{k-1}}{q_{k-1}} + \frac{(-1)^{k-1} j}{q_{k-1} q_k} \right) \right) = \left(\left(\frac{j p_{k-1}}{q_{k-1}} \right) \right) + \frac{(-1)^{k-1} j}{q_{k-1} q_k}, \quad (2.44)$$

when j is not divisible by q_{k-1} , and

$$\left(\left(\frac{j p_k}{q_k} \right) \right) = \left(\left(\frac{j p_{k-1}}{q_{k-1}} \right) \right) + \frac{(-1)^{k-1} j}{q_{k-1} q_k} + \frac{1 + (-1)^{k-1}}{2}, \quad (2.45)$$

when j is divisible by q_{k-1} . Thus we can rewrite \sum_2 [see (2.39)] in the form

$$\begin{aligned} & \sum_{i=b_k q_k + 1}^{N_k} (N_k + 1 - i) \left(\left(\frac{i p_k}{q_k} \right) \right) = \\ &= \sum_{j=1}^{N_{k-1}} (N_k - b_k q_k + 1 - j) \left(\left(\frac{j p_{k-1}}{q_{k-1}} \right) \right) = \\ &= \sum_{j=1}^{N_{k-1}} (N_k + 1 - j) \left(\left(\frac{j p_{k-1}}{q_{k-1}} \right) \right), \end{aligned}$$

and applying (2.44) and (2.45) we have [note that \sum_3 is defined in (2.42)]

$$\begin{aligned} \sum_2 = \sum_3 + \frac{(-1)^{k-1} j}{q_{k-1} q_k} \sum_{j=1}^{N_{k-1}} (N_k + 1 - j) j + \\ + b_{k-1} \frac{1 + (-1)^{k-1}}{2} \left(N_{k-1} + 1 - \frac{(b_{k-1} + 1) q_{k-1}}{2} \right). \end{aligned} \quad (2.46)$$

Combining (2.41), (2.42), and (2.46), Lemma 2.6 follows. \square

By using Lemma 2.6, we are ready to complete the proof of Proposition 2.1. Let's return to (2.36). First we extend the definition of $N_k = \sum_{i=1}^k b_i q_i$ for all $k > r$ in the trivial way: put $b_i = 0$ for $i > r$. We sum up both sides of Lemma 2.6 as $k = 1, 2, 3, \dots$; the left-hand side of (2.38) gives

$$\sum_{k=1}^r (N + 1 - k) ((k\alpha)), \quad (2.47)$$

and the right hand side of (2.38) gives

$$\begin{aligned} \sum_1^* + \sum_2^* + \sum_3^* \text{ where} \quad (2.48) \\ \sum_1^* = - \sum_{i=1}^r b_i q_i D(p_i, q_i), \\ \sum_2^* = \sum_{j=1}^r \frac{b_j}{4} (1 + (-1)^{j+1}) (2N_j + 1 - (b_j + 1) q_j), \\ \sum_3^* = \sum_{j=1}^{\infty} (-1)^j \frac{N_j (N_j + 1) (N_j + 2)}{6 q_j q_{j+1}} = \\ = \sum_{j=1}^r (-1)^j \frac{N_j (N_j + 1) (N_j + 2)}{6 q_j q_{j+1}} + \frac{N(N + 1)(N + 2)}{6} \left(\alpha - \frac{p_{r+1}}{q_{r+1}} \right), \end{aligned}$$

where in the last step we used (2.43) and the fact $p_i/q_i \rightarrow \alpha$ as $i \rightarrow \infty$.

First we evaluate \sum_1^* . By Lemma 2.4,

$$\begin{aligned}
 \sum_{i=1}^r b_i q_i D(p_i, q_i) &= \sum_{i=1}^r b_i q_i \left(\frac{a_1 - a_2 \pm \cdots + (-1)^i a_{i-1}}{12} + \frac{\theta}{4} \right) = \\
 &= \sum_{j=1}^r \frac{(-1)^j a_{j-1}}{12} (N - N_{j-1}) + \frac{\theta N}{4} = \\
 &= N \left(\sum_{j=1}^r \frac{(-1)^j a_{j-1}}{12} + \sum_{j=1}^r \frac{(-1)^{j-1} a_{j-1}}{12} \cdot \frac{N_{j-1}}{N} + \frac{\theta}{4} \right), \tag{2.49}
 \end{aligned}$$

where $|\theta_i| < 1$ and $|\theta| < 1$ are appropriate constants. Since the sequence $N_j = \sum_{i=1}^j b_i q_i$ increases at least exponentially fast, an upper bound like

$$\sum_{i=1}^k N_i \leq 4N_{k+1} \tag{2.50}$$

is trivial. Combining (2.49) and (2.50),

$$\sum_{i=1}^r b_i q_i D(p_i, q_i) = N \left(\frac{a_1 - a_2 \pm \cdots + (-1)^r a_{r-1}}{12} + \theta' \left(\max_{1 \leq j \leq r} a_j \right) + \theta'' \right), \tag{2.51}$$

where $|\theta'| \leq 4$ and $|\theta''| \leq 1/4$.

Next we estimate \sum_2^* from above:

$$\sum_2^* \leq \frac{1}{2} \sum_{i=1}^r b_i N_i \leq \frac{1}{2} \left(\max_{1 \leq j \leq r} a_j \right) \sum_{i=1}^r N_i \leq 3N \left(\max_{1 \leq j \leq r} a_j \right), \tag{2.52}$$

where in the last step we used (2.50).

Finally, we estimate \sum_3^* from above. Since

$$N_j = \sum_{i=1}^j b_i q_i \quad \text{and} \quad q_{j+1} \geq a_j q_j \geq b_j q_j,$$

we have

$$\left| \sum_{j=1}^r (-1)^j \frac{N_j (N_j + 1) (N_j + 2)}{6q_j q_{j+1}} \right| \leq \sum_{j=1}^r (b_j + 1)^2 q_j \leq 2N \left(\max_{1 \leq j \leq r} a_j \right). \tag{2.53}$$

We also have

$$\frac{N(N+1)(N+2)}{6} \cdot \left| \alpha - \frac{p_{r+1}}{q_{r+1}} \right| \leq \frac{N^3}{3q_{r+1}^2} \leq \frac{N}{3}. \quad (2.54)$$

Combining (2.47), (2.48), (2.51)–(2.54), we obtain

$$\begin{aligned} M_\alpha(N) &= \frac{1}{N} \sum_{k=1}^r (N+1-k)((k\alpha)) = \\ &= -\frac{a_1 - a_2 \pm \cdots + (-1)^r a_{r-1}}{12} + \theta \left(\max_{1 \leq j \leq r} a_j \right), \end{aligned} \quad (2.55)$$

where $|\theta| < 10$. Equation (2.55) completes the proof of Proposition 2.1. \square

Note that our original proof of Proposition 2.1 was a much longer, brute force deduction from Ostrowski's formula (1.55) (see [Be2, Be3]). Later Schoissengeier [Scho] pointed out the connection with Dedekind sums and some related results of Knuth [Kn1], which made the proof substantially shorter. The proof above follows the Schoissengeier–Knuth approach.

2.1.4 Proposition 2.1 and Some Works of Hardy and Littlewood

It is interesting to note that, a few weeks after we completed our proof of Proposition 2.1 (November 1995), we accidentally noticed the following technical lemma in Hardy–Littlewood [Ha-Li2].

“Lemma 14”: *If $\alpha = [a_0; a_1, a_2, \dots]$ then*

$$M_\alpha(N) = \frac{1}{12} \sum_{i=1}^l (-1)^k \left(\alpha_i + \frac{1}{\alpha_i} \right) + O \left(\left(\max_{1 \leq i \leq l} a_i \right)^2 \right), \quad (2.56)$$

where l is the least index such that $q_l \geq N$, and

$$\alpha_i = a_i + \frac{1}{a_{i+1} + \frac{1}{a_{i+2} + \cdots}} = [a_i; a_{i+1}, a_{i+2}, \dots].$$

By using the trivial identity $\alpha_i = a_i + \frac{1}{\alpha_{i+1}}$, the alternating sum in “Lemma 14” becomes

$$\begin{aligned} & -\left(\alpha_1 + \frac{1}{\alpha_1}\right) + \left(\alpha_2 + \frac{1}{\alpha_2}\right) - \left(\alpha_3 + \frac{1}{\alpha_3}\right) \pm \cdots \\ & = -a_1 + a_2 - a_3 \pm \cdots + (-1)^i a_i \pm \cdots \end{aligned} \quad (2.57)$$

The surprising conclusion is that from “Lemma 14” we can obtain a somewhat weaker version of Proposition 2.1 in *one line*. Note that (2.56) is weaker, because the error term $O((\max_{1 \leq i \leq l} a_i)^2)$ is the square of the linear error term $O(\max_{1 \leq i \leq l} a_i)$ in Proposition 2.1.

Note that Hardy and Littlewood proved their “Lemma 14” by using a different kind of reciprocity formula (namely, the reciprocity formula for the theta functions).

A related development is that, about 10 years later, in 1930, Hardy and Littlewood [Ha-Li3] studied the following (diophantine) series:

$$\sum_{n=1}^{\infty} \frac{1}{n \sin(\pi n \alpha)} \quad (2.58)$$

and made a very interesting discovery. Though the terms of the series (2.58) do not tend to zero for any α , Hardy and Littlewood managed to prove the next best thing; namely, that for the special value $\alpha = \sqrt{2}$ the partial sums of (2.58) remain uniformly bounded, i.e.,

$$\sum_{n=1}^N \frac{1}{n \sin(\pi n \alpha)} = O(1). \quad (2.59)$$

In general, if $\alpha = \sqrt{a^2 + 1}$, a is *odd*, then the partial sums are similarly $O(1)$.

On the other hand, Hardy and Littlewood noticed that for $\alpha = \sqrt{6}/2 - 1$ the N th partial sum is $c \log N + O(1)$ with $c \neq 0$.

What is going on here? The proof of the “ $O(1)$ -theorem” for $\alpha = \sqrt{a^2 + 1}$, a is *odd*, was so complicated, mysterious, and *ad hoc* that in his *Introduction to the Collected Papers of G.H. Hardy*, Vol. 1, Davenport listed the “real understanding” of this paper as a major research problem in diophantine approximation.

Now here is our “real understanding”: the “ $O(1)$ -theorem” of Hardy and Littlewood is a simple corollary of Proposition 2.1. Indeed, all that we need is the simple identity

$$\sum_{n=1}^N \frac{1}{n \sin(\pi n \alpha)} = 4\pi M_{\alpha/2}(N) - 2\pi M_{\alpha}(N) + O(\max_{1 \leq i \leq l} a_i), \quad (2.60)$$

where l is the last index such that $q_l \leq N$.

Equation (2.60) is an easy consequence of two facts. The first one is (2.12):

$$M_\alpha(N) = -\frac{1}{2\pi} \sum_{n=1}^N \frac{1}{n \tan(\pi n \alpha)} + O\left(\max_{1 \leq i \leq k} a_i\right)$$

where k is the last index for which $q_k \leq N$, and the second fact is a simple trigonometric identity:

$$\frac{1}{\tan(\beta)} - \frac{1}{\tan(2\beta)} = \frac{2 \cos^2(\beta) - \cos(2\beta)}{2 \sin(\beta) \cos(\beta)} = \frac{1}{\sin(2\beta)}.$$

It seems very likely that Hardy and Littlewood overlooked the simple application of Proposition 2.1 via (2.60) (the weaker error term (2.56) would be fine here). This is why they had to develop a complicated *ad hoc* method in [Ha-Li3].

We will return to the Hardy–Littlewood series $\sum_n 1/n \sin(\pi n \alpha)$ in Sect. 2.3.

2.2 Computing the Expectation in General (II)

2.2.1 The Expectation in Theorem 1.1

Next we switch from the saw-tooth function $((x))$ to the characteristic function

$$\chi_\rho(x) = \begin{cases} 1, & \text{if } 0 \leq x < \rho; \\ 0, & \text{if } \rho \leq x < 1, \end{cases} \quad (2.61)$$

of the interval $[0, \rho)$, where $0 < \rho < 1$, and extend it periodically modulo 1. Then we get the simple equation

$$\chi_\rho(x) - \rho = ((x - \rho)) - ((x)). \quad (2.62)$$

The sum

$$\sum_{k=1}^n \chi_\rho(k\alpha)$$

is the counting function for the irrational rotation: it counts the integers k in $1 \leq k \leq n$ for which $k\alpha \in [0, \rho)$ modulo 1. Theorem 1.1 is about this counting function. Therefore, to prove Theorem 1.1, we have to determine the corresponding expectation: by (2.62) we need to evaluate the generalized Dedekind sum

$$D(H, K; c) = \sum_{j=1}^{K-1} \left(\left(\frac{j}{K} \right) \right) \left(\left(\frac{jH+c}{K} \right) \right), \quad (2.63)$$

where c , the “shift constant,” is an arbitrary real number (by (2.62) we use $c = -\rho$ or $c = 1 - \rho$; it doesn’t matter which one).

The following lemma, a reciprocity law due to Dieter [Di], describes the connection between the ordinary Dedekind sum and its generalization (2.63). For later application, we have to include a proof.

Lemma 2.7. *Let $1 \leq H < K$ be relatively prime integers, and let $0 < c < K$ be a real number. Then*

$$D(H, K; c) + D(K, H; c) = D(H, K) + D(K, H) + \frac{\lfloor c \rfloor \lceil c \rceil}{2HK} - \frac{1}{2} \lfloor c/H \rfloor + \frac{1}{4} E(H, c), \quad (2.64)$$

where

$$E(H, c) = \begin{cases} 0, & \text{if } c \not\equiv 0 \pmod{H}; \\ 1, & \text{if } c \equiv 0 \pmod{H}. \end{cases} \quad (2.65')$$

Proof. First assume that c is a natural number; we prove (2.64) by induction on c . Clearly

$$\left(\left(\frac{jH+c+1}{K} \right) \right) = \left(\left(\frac{jH+c}{K} \right) \right) + \frac{1}{K} - \frac{1}{2} \delta \left(\frac{jH+c}{K} \right) + \frac{1}{2} \delta \left(\frac{jH+c+1}{K} \right), \quad (2.66)$$

where in this section we use the notation $\delta(x) = 1$ if x is an integer and 0 otherwise (“Kronecker delta”). By (2.63) and (2.66),

$$\begin{aligned} D(H, K; c+1) &= \sum_{j=1}^{K-1} \left(\left(\frac{j}{K} \right) \right) \left(\left(\frac{jH+c}{K} \right) \right) + \frac{1}{K} \sum_{j=1}^{K-1} \left(\left(\frac{j}{K} \right) \right) \\ &\quad - \frac{1}{2} \sum_{j=1}^{K-1} \left(\left(\frac{j}{K} \right) \right) \left(\delta \left(\frac{jH+c}{K} \right) + \delta \left(\frac{jH+c+1}{K} \right) \right). \end{aligned} \quad (2.67)$$

Since $1 \leq H < K$ are relatively prime, there exist two integers h' and k' such that

$$Hh' + Kk' = 1. \quad (2.68)$$

If

$$j \equiv -h'c \pmod{K} \text{ then } jH + c \equiv 0 \pmod{K},$$

and because the saw-tooth function $((x))$ is odd, we can rewrite (2.67) as follows:

$$D(H, K; c + 1) = D(H, K; c) + \frac{1}{2} \left(\left(\frac{h'c}{K} \right) \right) + \frac{1}{2} \left(\left(\frac{h'(c+1)}{K} \right) \right).$$

It follows by induction on c that

$$D(H, K; c) = D(H, K; 0) + \sum_{j=1}^{c-1} \left(\left(\frac{h'j}{K} \right) \right) + \frac{1}{2} \left(\left(\frac{h'c}{K} \right) \right). \quad (2.69)$$

For every j with $1 \leq j \leq K-1$ [see (2.68)]

$$\begin{aligned} \left(\left(\frac{h'j}{K} \right) \right) &= \left(\left(\frac{j - k'Kj}{HK} \right) \right) = - \left(\left(\frac{k'Kj - j}{HK} \right) \right) = \\ &= - \left(\left(\frac{k'j}{H} \right) \right) + \frac{j}{HK} - \frac{1}{2} \delta \left(\frac{k'j}{H} \right). \end{aligned} \quad (2.70)$$

Adding (2.69) to itself with H and K interchanged, and using (2.70), we have

$$D(H, K; c) + D(K, H; c) = D(H, K) + D(K, H) + S,$$

where

$$S = \sum_{j=1}^{i-1} \left(\frac{j}{HK} - \frac{1}{2} \delta \left(\frac{k'j}{H} \right) \right) + \frac{c}{2HK} - \frac{1}{4} \delta \left(\frac{k'c}{H} \right). \quad (2.71)$$

The evaluation of the last line in (2.71) is easy: we have

$$S = \frac{c^2}{2HK} - \frac{1}{2} \left\lfloor \frac{c}{H} \right\rfloor + \frac{1}{4} \delta \left(\frac{c}{H} \right). \quad (2.72)$$

Equations (2.71) and (2.72) complete the proof when c is any integer.

For an arbitrary real number c we use the identity

$$D(H, K; c + \theta) = D(H, K; c) + \frac{1}{2} \left(\left(\frac{h'\theta}{K} \right) \right), \quad (2.73)$$

where $c \geq 0$ is an integer and $0 < \theta < 1$ [h' is defined by (2.68)]. The proof of (2.73) is easy:

$$\begin{aligned} & \sum_{j=1}^{K-1} \left(\left(\frac{j}{K} \right) \right) \left(\left(\frac{jH + c + \theta}{K} \right) \right) = \\ &= \sum_{j=1}^{K-1} \left(\left(\frac{j}{K} \right) \right) \left(\left(\left(\frac{jH + c}{K} \right) \right) + \frac{\theta}{K} - \frac{1}{2} \delta \left(\frac{jK + c}{K} \right) \right) = \\ &= D(H, K; c) + 0 - \frac{1}{2} \left(\left(\frac{-h'c}{K} \right) \right), \end{aligned}$$

because $-h'Hc + c \equiv 0 \pmod{K}$, and (2.73) follows.

When $0 < \theta < 1$, Eqs. (2.73) and (2.70) imply that

$$\begin{aligned} D(H, K; c + \theta) + D(K, H; c + \theta) &= D(H, K; c) + D(K, H; c) + \\ &+ \frac{c}{2HK} - \frac{1}{4} \delta \left(\frac{c}{H} \right). \end{aligned}$$

This completes the proof of Lemma 2.7. \square

Lemma 2.7 leads to the following analog of Lemma 2.4; see Knuth [Kn1]. Again we need the proof.

Lemma 2.8. *Let $1 \leq H < K$ be relatively prime integers and let $0 < c < K$ be a real number. Let*

$$\frac{H}{K} = \frac{1}{a_1 + \frac{1}{a_2 + \dots}} = [a_1, a_2, a_3, \dots, a_\ell],$$

then

$$\begin{aligned} D(H, K; c) - D(H, K) &= \frac{-b_1 + b_2 - b_3 \pm \dots + (-1)^\ell b_\ell}{2} + \\ &+ \frac{c_0^2}{2KH} - \frac{c_1^2}{2HH_1} + \frac{c_2^2}{2H_1H_2} \mp \dots + (-1)^{\ell-1} \frac{c_{\ell-1}^2}{2H_{\ell-2}H_{\ell-1}} + O(1), \end{aligned} \quad (2.74)$$

where the terms b_i, c_i, H_i in (2.74) are determined by two Euclidean algorithms as follows. Let $H_{-1} = K, H_0 = H$, and define H_i by the first Euclidean algorithm

$$K = a_1 H + H_1, \quad H = a_2 H_1 + H_2, \quad H_1 = a_3 H_2 + H_4, \dots, \quad H_{\ell-2} = a_\ell H_{\ell-1}, \quad (2.75)$$

where $H_{\ell-1} = \gcd(H, K) = 1$ (\gcd denotes the greatest common divisor); then by using (2.75), we define the integers b_i and the real numbers c_i via the second Euclidean algorithm

$$c = c_0 = b_1 H_0 + c_1, \quad c_1 = b_2 H_1 + c_2, \quad c_2 = b_3 H_2 + c_3, \dots, \quad c_{\ell-1} = b_\ell H_{\ell-1} + c_\ell, \quad (2.76)$$

where $0 \leq c_1 < H_0$, $0 \leq c_2 < H_1$, \dots , and $0 \leq c_\ell < 1$ (note that $H_\ell = 0$). The error term $O(1)$ in (2.74) has absolute value ≤ 1 .

Proof. First assume that c is an integer; then $c_\ell = 0$. Write

$$\Delta(h, k; c) = D(h, k; c) - D(h, k)$$

and

$$F(h, k, c) = \frac{c^2}{2hk} - \frac{1}{2} \left\lfloor \frac{c}{h} \right\rfloor + \frac{1}{4} \delta \left(\frac{c}{h} \right),$$

then by Lemma 2.7,

$$\begin{aligned} \Delta(h, k; c) &= F(h, k, c) - \Delta(k, h; c) = \\ &= F(h, k, c) - \Delta(k \pmod{h}, h; c \pmod{h}). \end{aligned} \quad (2.77)$$

Combining the Euclidean algorithms (2.75) and (2.76) with (2.77), we have

$$\Delta(H_j, H_{j-1}; c_j) = F(H_j, H_{j-1}, c_j) - \Delta(H_{j+1}, H_j; c_{j+1}) \quad (2.78)$$

for $j = 0, 1, 2, \dots, \ell - 1$. Write

$$F_j = F(H_j, H_{j-1}, c_j),$$

then by repeated application of (2.78), we have

$$\begin{aligned} \Delta(H, K; c) &= F_0 - F_1 + F_2 - F_3 \pm \dots + (-1)^{\ell-1} F_{\ell-1} = \\ &= \sum_{j=0}^{\ell-1} (-1)^j \left(\frac{c_j^2}{2hk} - \frac{1}{2} b_{j+1} + \frac{1}{4} \delta \left(\frac{c_j}{H_j} \right) \right) = \\ &= \frac{-b_1 + b_2 - b_3 \pm \dots + (-1)^\ell b_\ell}{2} + \sum_{j=0}^{\ell-1} (-1)^j \frac{c_j^2}{2H_{j-1}H_j} + \frac{(-1)^{\ell-1}}{4}. \end{aligned} \quad (2.79)$$

Equation (2.79) proves Lemma 2.8 if c is an integer.

If c is not an integer then we simply apply (2.73). □

2.2.2 An Analog of Proposition 2.1

Let $0 < \alpha < 1$ be any irrational and let $0 < \rho < 1$ be any rational number. To prove Theorem 1.1 about the irrational rotation, first we need to know the average (“expectation”)

$$M_\alpha(\rho; N) = \frac{1}{N} \sum_{n=1}^N S_\alpha(\rho; n), \quad (2.80)$$

where

$$S_\alpha(\rho; n) = \sum_{k=1}^n (\chi_\rho(k\alpha) - \rho) \quad (2.81)$$

and the characteristic function $\chi_\rho(x)$ is defined in (2.61).

By using (2.62) we have

$$S_\alpha(\rho; n) = \sum_{k=1}^n (((k\alpha - \rho)) - ((k\alpha))),$$

and

$$M_\alpha(\rho; N) = \frac{1}{N} \sum_{n=1}^N (N + 1 - k) (((k\alpha - \rho)) - ((k\alpha))).$$

Repeating the proof of Proposition 2.1 with some natural modifications, we obtain the following analogous result.

Proposition 2.9. *For any irrational $\alpha > 0$, any real number $0 < \rho < 1$, and any integer $N \geq 1$,*

$$M_\alpha(\rho; N) = \frac{b_1 - b_2 + b_3 \mp \cdots + (-1)^{\ell-1} b_\ell}{2} - \frac{c_0^2}{2KH} + \frac{c_1^2}{2HH_1} - \frac{c_2^2}{2H_1H_2} \pm \cdots + (-1)^\ell \frac{c_{\ell-1}^2}{2H_{\ell-2}H_{\ell-1}} + \theta \cdot \max_{1 \leq j \leq \ell} b_j, \quad (2.82)$$

where $|\theta| < 10$, $\alpha = [a_1, a_2, \dots]$, the index $\ell = \ell(\alpha, N)$ is defined as the last integer j such that $q_j \leq N$, where p_j/q_j is the j -th convergent of α , and finally the terms b_i , c_i , H_i in (2.82) are determined by the two Euclidean algorithms (2.75) and (2.76) with $c = c_0 = (1 - \rho)K$, $K = q_\ell$, $H = p_\ell$ (i.e., $H/K = p_\ell/q_\ell$). \square

Next we show some illustrations.

Example 2.10. First let $\rho = 1/2$. We begin with $\alpha = \sqrt{2}$, and evaluate $M_{\sqrt{2}}(1/2; N)$, i.e., the corresponding expectation in Theorem 1.1. The continued fraction $\sqrt{2} - 1 = [2, 2, 2, \dots] = [\bar{2}]$ gives that $2 = a_1 = a_2 = a_3 = \dots$ in (2.75). Next we compute b_i, c_i, H_i in (2.76) as follows:

$$c = c_0 = (1 - \rho)K = \frac{1}{2}(2H + H_1) = H + \frac{1}{2}H_1,$$

implying $b_1 = 1$, and

$$c_1 = \frac{1}{2}H_1 = 0 \cdot H + \frac{1}{2}H_1, \text{ implying } b_2 = 0, \text{ and}$$

$$c_2 = \frac{1}{2}H_1 = \frac{1}{2}(2H_2H + H_3) = H_2 + \frac{1}{2}H_3, \text{ implying } b_3 = 1,$$

and so on. Thus we obtain the periodic sequences

$$b_1 = 1, b_2 = 0, b_3 = 1, b_4 = 0, \dots, b_i = \frac{1}{2}(1 + (-1)^{i-1});$$

$$c_0 = \frac{1}{2}K, c_1 = c_2 = \frac{1}{2}H_1, c_3 = c_4 = \frac{1}{2}H_3, c_5 = c_6 = \frac{1}{2}H_5, \dots$$

Hence we have

$$\frac{b_1 - b_2 + b_3 - b_4 \pm \dots}{2} = \frac{1 - 0 + 1 - 0 + 1 - 0 + \dots}{2} \quad (2.83)$$

and

$$\begin{aligned} & -\frac{c_0^2}{2KH} + \frac{c_1^2}{2HH_1} - \frac{c_2^2}{2H_1H_2} \pm \dots = \\ & = -\frac{K}{8H} - \frac{H_1}{8} \left(\frac{1}{H_2} - \frac{1}{H} \right) - \frac{H_3}{8} \left(\frac{1}{H_4} - \frac{1}{H_2} \right) - \frac{H_5}{8} \left(\frac{1}{H_6} - \frac{1}{H_4} \right) - \dots \end{aligned} \quad (2.84)$$

Since

$$\begin{aligned} \frac{H_{2i+1}}{8} \left(\frac{1}{H_{2i+2}} - \frac{1}{H_{2i}} \right) &= \frac{H_{2i+1}}{8} \cdot \frac{H_{2i} - H_{2i+2}}{H_{2i+2}H_{2i}} = \frac{H_{2i+1}}{8} \cdot \frac{2H_{2i+1}}{H_{2i+2}H_{2i}} = \\ &= \frac{H_{2i+1}^2}{4H_{2i+2}H_{2i}} = \frac{1}{4} + \text{exponentially small}, \end{aligned} \quad (2.85)$$

applying (2.83)–(2.85) in Proposition 2.9, by (2.82) we have

$$M_{\sqrt{2}}\left(\frac{1}{2}; N\right) = \left(\frac{1-0}{2} - \frac{1}{4}\right) \cdot \frac{1}{2} \cdot \frac{\log N}{\log(1+\sqrt{2})} + O(1),$$

where in the last step we used the fact that [see (2.79)]

$$q_\ell = \frac{(1+\sqrt{2})^\ell - (1-\sqrt{2})^\ell}{2\sqrt{2}} = N \text{ implies } \ell = \frac{\log N}{\log(1+\sqrt{2})} + O(1).$$

Thus we obtain

$$M_{\sqrt{2}}\left(\frac{1}{2}; N\right) = \frac{1}{8} \cdot \frac{\log N}{\log(1+\sqrt{2})} + O(1), \quad (2.86)$$

which proves (1.32).

In the special case $\rho = 1/2$ we have the *ad hoc* identity

$$\chi_{1/2}(x) - \frac{1}{2} = ((2x)) - 2((x)), \quad (2.87)$$

which gives the equation [see (2.62) and (2.80)]

$$M_\alpha\left(\frac{1}{2}; N\right) = M_{2\alpha}(N) - 2M_\alpha(N). \quad (2.88)$$

By using (2.88), we can easily double-check (2.86). What it means is that we apply Proposition 2.1 for both $\alpha = \sqrt{2} = [\bar{2}]$ and

$$2\alpha = 2\sqrt{2} = \sqrt{8} = [2; 1, 4, 1, 4, 1, 4, \dots] = [2; \overline{1, 4}].$$

The length of the period of $\alpha = \sqrt{2}$ is odd, so the corresponding alternating sum in Proposition 2.1 cancels out. Thus we have

$$\begin{aligned} M_{\sqrt{2}}\left(\frac{1}{2}; N\right) &= M_{2\sqrt{2}}(N) = \frac{-1 + 4 - 1 + 4 - 1 + 4 \mp \dots}{12} + O(1) = \\ &= \frac{1}{12} \cdot \frac{-1 + 4}{2} \cdot \frac{\log N}{\log(1+\sqrt{2})} + O(1) = \frac{1}{8} \cdot \frac{\log N}{\log(1+\sqrt{2})} + O(1), \end{aligned} \quad (2.89)$$

which gives back (2.86). In Eq. (2.89) we used the fact that the $(2i)$ th convergent p_{2i}/q_{2i} of $\sqrt{8}$ satisfies the equation

$$p_{2i} \pm q_{2i} \sqrt{8} = (3 \pm \sqrt{8})^i$$

(due to the fact that the least positive solution of $x^2 - 8y^2 = \pm 1$ is $x = 3, y = 1$), which implies

$$q_{2i} = \frac{1}{2\sqrt{8}} \left((3 + \sqrt{8})^i - (3 - \sqrt{8})^i \right) \approx (3 + \sqrt{8})^i = (1 + \sqrt{2})^{2i}.$$

The *ad hoc* equation (2.88) gives a shortcut for $\rho = 1/2$ with any quadratic irrational α . For example, if $\alpha = \sqrt{3} = [1; \overline{1, 2}]$ then

$$2\alpha = 2\sqrt{3} = \sqrt{12} = [3; \overline{2, 6}].$$

Thus by (2.88) and Proposition 2.1,

$$\begin{aligned} M_{\sqrt{3}}\left(\frac{1}{2}; N\right) &= M_{2\sqrt{3}}(N) - 2M_{\sqrt{3}}(N) = \\ &= \frac{1}{12} \left(\frac{-2+6}{2} \cdot \frac{\log N}{\log(2+\sqrt{3})} - 2 \cdot \frac{-1+2}{2} \cdot \frac{2\log N}{\log(2+\sqrt{3})} \right) + O(1) = O(1), \end{aligned} \quad (2.90)$$

since the $(2i)$ th convergent p_{2i}/q_{2i} of $\sqrt{3}$ satisfies the equation

$$p_{2i} \pm q_{2i}\sqrt{3} = (2 \pm \sqrt{3})^i,$$

which implies

$$q_{2i} = \frac{1}{2\sqrt{3}} \left((2 + \sqrt{3})^i - (2 - \sqrt{3})^i \right) \approx (2 + \sqrt{3})^i;$$

similarly, the i th convergent denominator for $2\sqrt{3}$ is about $(2 + \sqrt{3})^i$ (because the least positive solution of $x^2 - 12y^2 = \pm 1$ is $x = 7, y = 2$, and $7 + 2\sqrt{12} = (2 + \sqrt{3})^2$).

Next consider the golden ratio $\alpha = (\sqrt{5}+1)/2$. Then $\alpha = [1; \overline{1}]$ and $2\alpha = [3; \overline{4}]$. Since the length of the period is odd for both continued fractions, by (2.88) and Proposition 2.1,

$$M_{(\sqrt{5}+1)/2}\left(\frac{1}{2}; N\right) = O(1). \quad (2.91)$$

The last example in this section is $\alpha = \sqrt{7}$ (again $\rho = 1/2$). We need the following facts: $\sqrt{7} = [2; \overline{1, 1, 1, 4}]$, $\sqrt{28} = [5; \overline{3, 2, 3, 10}]$, the least positive solutions of $x^2 - 7y^2 = \pm 1$ and $x^2 - 28y^2 = \pm 1$ are, respectively, $x = 8, y = 3$ and $x = 127, y = 24$ with the relation $127 + 24\sqrt{28} = (8 + 3\sqrt{7})^2$. Combining these facts with (2.88) and Proposition 2.1, we have

$$\begin{aligned}
M_{\sqrt{7}}\left(\frac{1}{2}; N\right) &= M_{2\sqrt{7}}(N) - 2M_{\sqrt{7}}(N) = \\
&= \frac{\log N}{12} \left(\frac{-3 + 2 - 3 + 10}{\log(127 + 24\sqrt{28})} - 2 \frac{-1 + 1 - 1 + 4}{\log(8 + 3\sqrt{7})} \right) \\
&\quad + O(1) = -\frac{\log N}{4 \log(8 + 3\sqrt{7})} + O(1). \tag{2.92}
\end{aligned}$$

Next we discuss examples where $\rho \neq 1/2$.

Example 2.11. Next let $\rho = 1/3$ and $\alpha = \sqrt{2}$. Then $\sqrt{2} = [1; \bar{2}]$ gives that $2 = a_1 = a_2 = a_3 = \dots$ in (2.75). We compute b_i, c_i, H_i in (2.76) as follows:

$$c = c_0 = (1 - \rho)K = \frac{2}{3}K = \frac{2}{3}(2H + H_1) = H + \frac{1}{3}H + \frac{2}{3}H_1,$$

implying $b_1 = 1$, and similarly

$$c_1 = \frac{1}{3}H + \frac{2}{3}H_1 = \frac{1}{3}(2H_1 + H_2) + \frac{2}{3}H_1 = H_1 + \frac{1}{3}H_1 + \frac{1}{3}H_2, \text{ implying } b_2 = 1, \text{ and}$$

$$c_2 = \frac{1}{3}H_1 + \frac{1}{3}H_2 = \frac{1}{3}(2H_2 + H_3) + \frac{1}{3}H_3 = H_2 + \frac{1}{3}H_3, \text{ implying } b_3 = 1, \text{ and}$$

$$c_3 = \frac{1}{3}H_3 = 0 \cdot H_3 + \frac{1}{3}H_3, \text{ implying } b_4 = 0, \text{ and}$$

$$c_4 = \frac{1}{3}H_3 = \frac{1}{3}(2H_4 + H_5) = 0 \cdot H_4 + \frac{2}{3}H_4 + \frac{1}{3}H_5, \text{ implying } b_5 = 0, \text{ and}$$

$$c_5 = \frac{2}{3}H_4 + \frac{1}{3}H_5 = \frac{2}{3}(2H_5 + H_6) + \frac{1}{3}H_5 = H_5 + \frac{2}{3}H_5 + \frac{2}{3}H_6, \text{ implying } b_6 = 1, \text{ and}$$

$$c_6 = \frac{2}{3}H_5 + \frac{2}{3}H_5 = \frac{2}{3}(2H_6 + H_7) + \frac{2}{3}H_6 = 2H_6 + \frac{2}{3}H_7, \text{ implying } b_7 = 2, \text{ and}$$

$$c_7 = \frac{3}{3}H_7 = 0 \cdot H_7 + \frac{2}{3}H_7, \text{ implying } b_8 = 0, \text{ and}$$

$$c_8 = \frac{2}{3}H_7 = \frac{2}{3}(2H_8 + H_9) = H_8 + \frac{1}{3}H_8 + \frac{2}{3}H_9, \text{ implying } b_9 = 1, \text{ and so on,}$$

back to the beginning. Thus we get the periodic sequence for b_1, b_2, b_3, \dots :

$$1, 1, 1, 0, 0, 1, 2, 0, \quad 1, 1, 1, 0, 0, 1, 2, 0, \quad 1, 1, 1, 0, 0, 1, 2, 0, \quad \dots$$

Therefore, we obtain

$$\begin{aligned} & \frac{b_1 - b_2 + b_3 - b_4 \pm \dots}{2} = \\ &= \frac{1}{2} \cdot \frac{1 - 1 + 1 - 0 + 0 - 1 + 2 - 0}{8} \cdot \frac{\log N}{\log(1 + \sqrt{2})} + O(1), \end{aligned} \quad (2.93)$$

and

$$\begin{aligned} & -\frac{c_0^2}{2KH} + \frac{c_1^2}{2HH_1} - \frac{c_2^2}{2H_1H_2} \pm \dots = \\ &= \frac{1}{18} \left(-\frac{(2K)^2}{KH} + \frac{(H + 2H_1)^2}{HH_1} - \frac{(H_1 + H_2)^2}{H_1H_2} + \frac{H_3^2}{H_2H_3} \right) \frac{\log N}{8 \log(1 + \sqrt{2})} + \\ &+ \frac{1}{18} \left(-\frac{H_3^2}{H_3H_4} + \frac{(2H_4 + H_5)^2}{H_4H_5} - \frac{(2H_5 + 2H_6)^2}{H_5H_6} + \frac{(2H_7)^2}{H_6H_7} \right) \frac{\log N}{8 \log(1 + \sqrt{2})} + O(1). \end{aligned} \quad (2.94)$$

Since by (2.75)

$$\frac{H_i - H_{i+2}}{H_{2i+1}} = a_{i+2} = 2,$$

we can rewrite (2.94) as follows:

$$\begin{aligned} \text{sum}(2.94) &= \frac{1}{18} \left(-\frac{4(K - H_1)}{H} + 4 + \frac{H - H_2}{H_1} - 2 - \frac{H_1 - H_3}{H_2} - \frac{H_3 - H_5}{H_4} + \right. \\ &\quad \left. + \frac{4(H_4 - H_6)}{H_5} - 8 - \frac{4(H_5 - H_7)}{H_6} \right) = \\ &= \frac{1}{18} (-8 + 4 + 2 - 2 - 2 - 2 + 4 + 8 - 8 - 8) = -\frac{2}{3}, \end{aligned}$$

implying

$$\text{sum}(2.94) = -\frac{\log N}{12 \log(1 + \sqrt{2})} + O(1). \quad (2.95)$$

Applying (2.93)–(2.95) in (2.82), we have

$$\begin{aligned} M_{\sqrt{2}}\left(\frac{1}{3}; N\right) &= \left(\frac{1}{8} - \frac{1}{12}\right) \frac{\log N}{\log(1 + \sqrt{2})} + O(1) = \\ &= \frac{\log N}{24 \log(1 + \sqrt{2})} + O(1). \end{aligned} \quad (2.96)$$

Next let $\rho = 2/3$ and $\alpha = \sqrt{2}$, then a similar calculation gives the same answer:

$$M_{\sqrt{2}}\left(\frac{2}{3}; N\right) = \frac{\log N}{24 \log(1 + \sqrt{2})} + O(1). \quad (2.97)$$

We can easily double-check (2.96) and (2.97) by using the *ad hoc* equation

$$\left(\chi_{1/3}(x) - \frac{1}{3}\right) + \left(\chi_{2/3}(x) - \frac{2}{3}\right) = ((3x)) - 3((x)), \quad (2.98)$$

which leads to [see (2.62) and (2.80)]

$$M_{\alpha}\left(\frac{1}{3}; N\right) + M_{\alpha}\left(\frac{2}{3}; N\right) = M_{3\alpha}(N) - 3M_{\alpha}(N). \quad (2.99)$$

Notice that (2.98) and (2.99) is an analog of (2.87) and (2.88).

We have $3\sqrt{2} = \sqrt{18} = [4; \overline{4, 8}]$, and so by Proposition 2.1,

$$M_{3\sqrt{2}}(N) = \frac{1}{12} \cdot \frac{-4 + 8}{2} \cdot \frac{\log N}{2 \log(1 + \sqrt{2})} + O(1), \quad (2.100)$$

because the least positive solution of $x^2 - 18y^2 = \pm 1$ is $x = 17, y = 4$, and so the $(2i)$ th convergent p_{2i}/q_{2i} of $\sqrt{18}$ satisfies the equation

$$p_{2i} \pm q_{2i} \sqrt{18} = (17 \pm 4\sqrt{18})^i,$$

which implies

$$q_{2i} \approx (17 + 4\sqrt{18})^i = (1 + \sqrt{2})^{4i}.$$

Since the length of the period of $\sqrt{2}$ is odd, by (2.99) and (2.100),

$$\begin{aligned} M_{\sqrt{2}}\left(\frac{1}{3}; N\right) + M_{\sqrt{2}}\left(\frac{2}{3}; N\right) &= M_{3\sqrt{2}}(N) - 3M_{\sqrt{2}}(N) = \\ &= \frac{\log N}{12 \log(1 + \sqrt{2})} + O(1), \end{aligned}$$

which is in agreement with (2.96) and (2.97).

Example 2.12. Let $\rho = 1/4$ and $\alpha = (\sqrt{5} + 1)/2 = [1; \bar{1}]$. Then $1 = a_1 = a_2 = a_3 = \dots$ in (2.75),

$$c = c_0 = (1 - \rho)K = \frac{3}{4}K = \frac{3}{4}(H + H_1) = H + \frac{1}{4}(3H_1 - H), \quad (2.101)$$

implying $b_1 = 1$. Note that $3H_1 > H$, since H/H_1 is very close to the golden ratio $\alpha = (\sqrt{5} + 1)/2 < 3$. We have

$$3H_1 - H = 3H_1 - (H_1 + H_2) = 2H_1 - H_2 = 2(H_2 + H_3) - H_2 = H_2 + 2H_3, \quad (2.102)$$

and so

$$\begin{aligned} c_1 &= \frac{1}{4}H_2 + \frac{1}{2}H_3 = 0 \cdot H_1 + c_2 = 0 \cdot H_2 + c_3, \text{ implying } b_2 = b_3 = 0, \text{ and} \\ c_3 &= \frac{1}{4}H_2 + \frac{1}{2}H_3 = \frac{1}{4}(H_3 + H_4) + \frac{1}{2}H_3 = \frac{3}{4}H_3 + \frac{1}{4}H_4 < H_3, \text{ implying } b_4 = 0, \text{ and} \\ c_4 &= \frac{3}{4}H_3 + \frac{1}{4}H_4 = \frac{3}{4}(H_4 + H_5) + \frac{1}{4}H_4 = H_4 + \frac{3}{4}H_5, \text{ implying } b_5 = 1, \text{ and} \\ c_5 &= \frac{3}{4}H_5 = 0 \cdot H_5 + \frac{3}{4}H_5, \text{ implying } b_6 = 0, \text{ and} \\ c_6 &= \frac{3}{4}H_5 = \frac{3}{4}(H_6 + H_7) = H_6 + \frac{1}{4}(3H_7 - H_6), \end{aligned}$$

which is the same as the beginning. Thus we get the periodic sequence for b_1, b_2, b_3, \dots :

$$1, 0, 0, 0, 1, 0, \quad 1, 0, 0, 0, 1, 0, \quad 1, 0, 0, 0, 1, 0, \quad \dots,$$

implying

$$\begin{aligned} &\frac{b_1 - b_2 + b_3 - b_4 \pm \dots}{2} = \\ &= \frac{1}{2} \cdot \frac{1 - 0 + 0 - 0 + 1 - 0}{6} \cdot \frac{\log N}{\log \frac{\sqrt{5}+1}{2}} + O(1), \end{aligned} \quad (2.103)$$

and

$$-\frac{c_0^2}{2KH} + \frac{c_1^2}{2HH_1} - \frac{c_2^2}{2H_1H_2} \pm \dots = \frac{1}{32} \cdot S_0 \cdot \frac{\log N}{6 \log \frac{\sqrt{5}+1}{2}} + O(1), \quad (2.104)$$

where

$$S_0 = -\frac{9K^2}{KH} + (H_2 + 2H_3)^2 \left(\frac{1}{HH_1} - \frac{1}{H_1H_2} + \frac{1}{H_2H_3} \right) - \frac{(3H_3 + H_4)^2}{H_3H_4} + \frac{9H_5^2}{H_4H_5}.$$

The critical sum S_0 in the middle of (2.104) equals (with $\alpha = (\sqrt{5} + 1)/2$)

$$S_0 = -9\alpha + (\alpha + 2)^2 (\alpha^{-5} - \alpha^{-3} + \alpha^{-1}) - \frac{(3\alpha + 1)^2}{\alpha} + 9\alpha^{-1}, \quad (2.105)$$

and using the simple facts $\alpha^2 = 1 + \alpha$ and $\alpha^{-2} = 1 - \alpha^{-1}$, it is easy to evaluate (2.105): $S_0 = -24$. Returning to (2.104), we have

$$\text{sum}(2.104) = \frac{1}{32} \cdot (-24) \cdot \frac{\log N}{6 \log \frac{\sqrt{5}+1}{2}} + O(1). \quad (2.106)$$

Applying (2.103)–(2.106) in (2.82), we have

$$\begin{aligned} M_{(\sqrt{5}+1)/2} \left(\frac{1}{4}; N \right) &= \left(1 - \frac{24}{32} \right) \frac{\log N}{6 \log \frac{\sqrt{5}+1}{2}} + O(1) = \\ &= \frac{\log N}{24 \log \frac{\sqrt{5}+1}{2}} + O(1). \end{aligned} \quad (2.107)$$

2.2.3 Periodicity in Proposition 2.9

Let's return to Proposition 2.9 and Eq. (2.82). The periodicity of b_1, b_2, b_3, \dots in the examples above was not an accident: we prove that if the sequence a_1, a_2, a_3, \dots is periodic and c/K is a rational number, then b_1, b_2, b_3, \dots is also periodic (but the length of the period is not necessarily the same).

Indeed, write $c/K = s/t$ where $1 \leq s < t$ are relatively prime integers. Then by (2.75) and (2.76),

$$c = c_0 = \frac{s}{t}K = \frac{s}{t}(a_1H + H_1) = b_1H + c_1,$$

where $(\lfloor x \rfloor)$ and $\{x\}$ denote the lower integral part and the fractional part of x)

$$b_1 = \left\lfloor \frac{sa_1}{t} \right\rfloor \quad \text{and} \quad c_1 = \left\{ \frac{sa_1}{t} \right\} H + \frac{s}{t} H_1 = \frac{s_1}{t} H + \frac{s}{t} H_1,$$

and here we *assume* that $c_1 < H$.

Similarly,

$$c_1 = \frac{s_1}{t} H + \frac{s}{t} H_1 = \frac{s_1}{t} (a_2 H_1 + H_2) + \frac{s}{t} H_1 = b_2 H_1 + c_2,$$

where

$$b_2 = \left\lfloor \frac{s_1 a_2 + s}{t} \right\rfloor \quad \text{and} \quad c_2 = \left\{ \frac{s_1 a_2 + s}{t} \right\} H + \frac{s_1}{t} H_2 = \frac{s_2 H_1 + s_1 H_2}{t},$$

and again we *assume* that $c_2 < H_1$.

Repeating this argument, for every $i \geq 0$ we have

$$c_i = \frac{s_i H_{i-1} + s_{i-1} H_i}{t}, \quad (2.108)$$

where $0 \leq s_i, s_{i-1} < t$ are integers, and we always *assume* that $c_i < H_{i-1}$.

The periodicity of a_i means that

$$a_i = a_{i+L} \quad \text{holds for (say) } M_1 \leq i \leq M_2, \quad (2.109)$$

and here we assume that $(M_2 - M_1)/L$ is a very large integer. Consider now the sequence with gap L [see (2.109)]:

$$c_{M_1}, c_{M_1+L}, c_{M_1+2L}, c_{M_1+3L}, \dots, c_{M_2};$$

by (2.108) we have

$$c_{M_1+jL} = \frac{s'_j H_{M_1+jL-1} + s''_j H_{M_1+jL}}{t} < H_{M_1+jL-1}, \quad (2.110)$$

where $0 \leq s'_j, s''_j < t$ are integers. If $(M_2 - M_1)/L$ is larger than t^2 , then by the Pigeonhole Principle there is a repetition among the pairs (s'_j, s''_j) , $j = 0, 1, 2, \dots$, and the first repetition implies the *periodicity* of the sequence b_1, b_2, b_3, \dots in the rest of the interval $M_1 \leq i \leq M_2$ [see (2.109)]. Of course, we cannot predict the length of the period, but it is certainly less than $L(t^2 + 1)$.

Warning! It may happen that our assumption

$$c_i = \frac{s_i H_{i-1} + s_{i-1} H_i}{t} < H_{i-1}, \quad 0 \leq s_i, s_{i-1} < t,$$

in (2.108) is violated; for example, see Eq. (2.101) in Example 2.12 (where $\alpha = (\sqrt{5} + 1)/2$ and $\rho = 1/4$):

$$c_0 = \frac{3}{4}(H + H_1) > H,$$

since H/H_1 is very close to $\alpha = (\sqrt{5} + 1)/2 < 3$. This is why we *cannot* write

$$c_0 = 0 \cdot H + c_1 \text{ with } c_1 = \frac{3}{4}(H + H_1),$$

instead we have to use

$$c_0 = H + \frac{3H_1 - H}{4} = H + c_1,$$

where in c_1 we face a negative(!) coefficient:

$$0 < c_1 = \left(-\frac{1}{4}\right)H + \frac{3}{4}H_1 < H. \quad (2.111)$$

For $\alpha = (\sqrt{5} + 1)/2 < 3$ we can use the *ad hoc* fact [see (2.102)]

$$3H_1 - H = H_2 + 2H_3, \quad (2.112)$$

which simply eliminates the “negativity problem” in (2.111).

Next we show that this trick *always* works; we can always eliminate the “negativity problem.” To prove this, assume that for some i we have—just like in (2.110)—the reverse of (2.108):

$$c_i = \frac{s_i H_{i-1} + s_{i-1} H_i}{t} > H_{i-1} \quad 0 \leq s_i, s_{i-1} < t. \quad (2.113)$$

Then we rewrite (2.113) in the form

$$c_i = H_{i-1} + c'_i \text{ where } c'_i = \frac{s_{i-1} H_i - (t - s_i) H_{i-1}}{t}$$

and $0 \leq c'_i < H_{i-1}$. In (2.75) we have the recurrence formula $H_{i-1} = a_{i+1} H_i + H_{i+1}$, so with $r_i = t - s_i$,

$$s_{i-1} H_i - r_i H_{i-1} = s_{i-1} H_i - r_i (a_{i+1} H_i + H_{i+1}) = s_{i-1}^* H_i - r_i H_{i+1},$$

where $s_{i-1}^* = s_{i-1} - r_i a_{i+1} \geq 1$.

Case 1: $s_{i-1}^* \geq r_i$.

By using $H_i = a_{i+2}H_{i+1} + H_{i+2}$, we have the following analog of (2.112):

$$\begin{aligned} s_{i-1}^* H_i - r_i H_{i+1} &= s_{i-1}^* (a_{i+2} H_{i+1} + H_{i+2}) - r_i H_{i+1} = \\ &= (s_{i-1}^* a_{i+2} - r_i) H_{i+1} + s_{i-1}^* H_{i+2}, \end{aligned} \quad (2.114)$$

which eliminates the “negativity problem.”

Case 2: $s_{i-1}^* < r_i$.

Then again we use (2.114):

$$s_{i-1}^* H_i - r_i H_{i+1} = (s_{i-1}^* a_{i+2} - r_i) H_{i+1} + s_{i-1}^* H_{i+2}. \quad (2.115)$$

If $(s_{i-1}^* a_{i+2} - r_i)$ is positive, then we are done; if it is negative, then clearly $r_{i+2} = |s_{i-1}^* a_{i+2} - r_i| < s_{i-1}^*$, and we can rewrite (2.115) in the form

$$s_{i-1}^* H_i - r_i H_{i+1} = s_{i-1}^* H_{i+2} - r_{i+2} H_{i+1} \quad \text{where } r_i > r_{i+2} \geq 0. \quad (2.116)$$

The decreasing property in (2.116) guarantees that, repeating this argument less than t times, the negative coefficient eventually disappears [i.e., turns into a positive coefficient like in (2.112)]. In other words, in both cases we can eliminate the “negativity problem.”

By getting rid of the “negativity problem,” we are safe to say that the Pigeonhole Principle argument above always works. As a consequence, we obtain the *periodicity* of b_1, b_2, b_3, \dots . Combining this periodicity with Lemma 2.7 and Proposition 2.9 [see Eq. (2.82)], we have

Proposition 2.13. *If α is a quadratic irrational and $0 < \rho < 1$ is a rational number, then there is a constant $c = c(\alpha, \rho)$ such that*

$$M_\alpha(\rho, N) = c \cdot \log N + O(1) \quad (2.117)$$

holds for every integer $N \geq 2$.

2.3 Fourier Series and a Problem of Hardy and Littlewood (I)

It is a standard exercise in every Fourier analysis course to compute the Fourier coefficients of the sawtooth function

$$((x)) = - \sum_{j=1}^{\infty} \frac{\sin(2\pi jx)}{\pi j}, \quad (2.118)$$

where $((x)) = \{x\} - 1/2$ if x is not an integer and 0 otherwise. We want to apply (2.118) in both

$$S_\alpha(n) = \sum_{k=1}^n ((k\alpha)) \quad \text{and} \quad M_\alpha(N) = \frac{1}{N} \sum_{n=1}^N S_\alpha(n) = \frac{1}{N} \sum_{n=1}^N (N+1-k)((k\alpha)),$$

but we have to be a little bit careful, since the Fourier series in (2.118) is not absolutely convergent. Instead of (2.118) we actually use a finite version with a small error term. First we recall Abel's transformation ("discrete integration by parts"):

$$\begin{aligned} \sum_{j=1}^m a_j b_j &= a_1(b_1 - b_2) + (a_1 + a_2)(b_2 - b_3) + \\ &+ (a_1 + a_2 + a_3)(b_3 - b_4) + \dots + (a_1 + \dots + a_{m-1})(b_{m-1} - b_m) + (a_1 + \dots + a_m)b_m. \end{aligned} \quad (2.119)$$

We also need the well-known summation formula

$$\sum_{j=1}^m \sin(j\beta) = \frac{\cos(\beta/2) - \cos((2m+1)\beta/2)}{2 \sin(\beta/2)}, \quad (2.120)$$

which implies the useful upper bound

$$\left| \sum_{j=1}^m \sin(j\beta) \right| \leq \frac{1}{|\sin(\beta/2)|}. \quad (2.121)$$

The pointwise convergence of the Fourier series in (2.118) follows from (2.119) and (2.121), and the equality of the two sides in (2.118) follows from Fejér's well-known theorem in Fourier analysis.

By (2.119) and (2.121), for any $T \geq 1$,

$$\left| ((x)) + \sum_{j=1}^T \frac{\sin(2\pi jx)}{\pi j} \right| \leq \frac{2}{\pi T |\sin(\pi x)|} < \frac{1}{T \|x\|}, \quad (2.122)$$

where $\|x\|$ denotes, as usual, the distance of x from the nearest integer. It follows that

$$\left| S_\alpha(n) + \sum_{j=1}^T \sum_{k=1}^n \frac{\sin(2\pi j k \alpha)}{\pi j} \right| < \frac{1}{T} \sum_{k=1}^n \frac{1}{\|k\alpha\|}. \quad (2.123)$$

2.3.1 Badly Approximable Numbers

We need to estimate the diophantine sum

$$\sum_{k=1}^n \frac{1}{\|k\alpha\|}$$

from above for the class of quadratic irrational α . Our argument below—a standard application of the Pigeonhole Principle—will work even for a larger class of reals, called *badly approximable* numbers. A real number α is called badly approximable, if there is a positive constant $c_0 = c_0(\alpha) > 0$ such that

$$k\|k\alpha\| \geq c_0 > 0 \text{ holds for all integers } k \geq 1.$$

One can easily characterize this class in terms of the continued fraction: α is badly approximable if and only if the sequence a_1, a_2, a_3, \dots of partial quotients in $\alpha = [a_0; a_1, a_2, a_3, \dots]$ is bounded, i.e., there is a threshold $M_0 = M_0(\alpha) < \infty$ such that $a_k \leq M_0$ holds for all $k \geq 1$. The well-known fact from diophantine approximation

$$\alpha = \frac{p_i}{q_i} + \frac{(-1)^{i+1}}{q_i(q_{i+1} + \theta q_i)},$$

where $p_i/q_i = [a_0; a_1, \dots, a_{i-1}]$ is the i th convergent of α , $q_{i+1} = a_i q_i + q_{i-1}$, and $0 < \theta = \theta(i) < 1$, implies that c_0 and M_0 are basically reciprocals of each other (apart from an absolute constant factor). Note that every quadratic irrational is badly approximable, since periodicity implies boundedness.

Lemma 2.14. *Assume that α is badly approximable, and $k\|k\alpha\| \geq c_0 > 0$ holds for all integers $k \geq 1$. Then for any integer n ,*

$$\sum_{k=1}^n \frac{1}{\|k\alpha\|} \leq \frac{4}{c_0} n \left(\log \left(\frac{n}{c_0} \right) / \log 2 \right).$$

In general, for any $m > 2$ we have

$$\sum_{\substack{n < k \leq 2n: \\ k\|k\alpha\| < m}} \frac{1}{\|k\alpha\|} = O(n \log m).$$

Proof. What we do is a routine application of the Pigeonhole Principle. To prove the first part, we define the set

$$A_j = \left\{ 1 \leq k \leq n : \frac{2^{j-1}}{n} c_0 \leq \|k\alpha\| < \frac{2^j}{n} c_0 \right\}.$$

Of course A_j is empty if

$$\frac{2^{j-1}}{n} c_0 > \frac{1}{2}. \quad (2.124)$$

We claim that the set A_j has at most 2^{j+1} elements. Indeed, if $|A_j| > 2^{j+1}$ then by the Pigeonhole Principle there exist $1 \leq k_1 < k_2 \leq n$ such that $k_i \in A_j$, $i = 1, 2$, and

$$|\{k_1\alpha\} - \{k_2\alpha\}| < \frac{c_0}{n}.$$

By choosing $\ell = k_2 - k_1$, we have $\|\ell\alpha\| < c_0/n$, which contradicts the hypothesis $\ell\|\ell\alpha\| \geq c_0 > 0$. Thus we have

$$\begin{aligned} \sum_{k=1}^n \frac{1}{\|k\alpha\|} &= \sum_{j \geq 1} \sum_{k \in A_j} \frac{1}{\|k\alpha\|} \leq \\ &\leq \sum_{j \geq 1} \frac{n}{2^{j-1}c_0} |A_j| \leq \frac{1}{c_0} \sum_{j \geq 1} \frac{n}{2^{j-1}} 2^{j+1} = \\ &= \frac{4n}{c_0} \sum_{j \geq 1: 2^j \leq n/c_0} 1 < \frac{4n}{c_0} \log \frac{n}{c_0} / \log 2, \end{aligned}$$

where at the end we used (2.124). This proves the first part in Lemma 2.14.

The same Pigeonhole Principle argument proves the second part. \square

By Eqs. (2.120), (2.123) and by Lemma 2.14, we obtain

Lemma 2.15. *Assume that α is badly approximable, and $k\|k\alpha\| \geq c_0 > 0$ for all integers $k \geq 1$. Then for any n and T ,*

$$\begin{aligned} S_\alpha(n) &= \sum_{j=1}^T \frac{\cos((2n+1)\pi j\alpha) - \cos(\pi j\alpha)}{2\pi j \sin(\pi j\alpha)} + \\ &\quad + \theta_1 \frac{4n \log(n/c_0)}{\log 2 c_0 T} \end{aligned} \quad (2.125)$$

and

$$M_\alpha(N) = \frac{1}{N} \sum_{n=1}^N S_\alpha(n) = - \sum_{j=1}^T \frac{1}{2\pi j \tan(\pi j \alpha)} - \sum_{j=1}^T \frac{\sin(2\pi j \alpha) - \sin(2\pi(N+1)j \alpha)}{4\pi N j \sin^2(\pi j \alpha)} + \theta_2 \frac{4n \log(n/c_0)}{\log 2c_0 T}, \quad (2.126)$$

where $|\theta_1| < 1$ and $|\theta_2| < 1$. □

The only novelty in the proof of (2.126) is the use of the summation formula

$$\sum_{n=1}^N \cos(n\beta + \gamma) = \frac{\sin((N + \frac{1}{2})\beta + \gamma) - \sin(\frac{1}{2}\beta + \gamma)}{2 \sin(\beta/2)}, \quad (2.127)$$

instead of (2.120).

2.3.2 The Hardy–Littlewood Series

Now we return to the numerical series

$$\sum_{n=1}^{\infty} \frac{1}{n \sin(\pi n \alpha)}, \quad \alpha \text{ is irrational}, \quad (2.128)$$

briefly mentioned at the end of Sect. 2.1. First notice that the series (2.128) cannot be convergent, since the terms do not tend to zero for any α . Indeed, the inequality $\|n\alpha\| < 1/n$ holds for infinitely many values of n , for example, let $n = q_j$ where p_j/q_j is the j th convergent of α . The inequality $\|n\alpha\| < 1/n$ combined with the trivial fact $|\sin(\pi n \alpha)| \leq \pi \|n\alpha\|$ implies that (2.128) contains infinitely many terms that have absolute value $\geq 1/\pi$. Thus the convergence is out of the question. Nevertheless, Hardy and Littlewood made the very interesting discovery that for the special value $\alpha = \sqrt{2}$ the partial sums of (2.128) remain uniformly bounded, that is,

$$\sum_{n=1}^N \frac{1}{n \sin(\pi n \alpha)} = O(1). \quad (2.129)$$

Equation (2.129) represents a miraculous cancellation; we can consider it the next best thing to convergence.

Note that Hardy and Littlewood actually proved the slightly more general result that if $\alpha = \sqrt{a^2 + 1}$, a is odd, then the partial sums always remain bounded. On the

other hand, for many other quadratic irrationals the N th partial sum is $c \log N + O(1)$ with $c \neq 0$ (Hardy and Littlewood gave the example $\alpha = \sqrt{6}/2 - 1$).

What is going on here? We will give a very transparent proof of (2.129) by using the following improved version of (2.126).

Proposition 2.16. *If α is badly approximable, then for any N ,*

$$M_\alpha(N) = - \sum_{j=1}^N \frac{1}{2\pi j \tan(\pi j \alpha)} + O(1), \quad (2.130)$$

where the implicit constant $O(1) = O_\alpha(1)$ is independent of N .

We postpone the proof of Proposition 2.16 to the next section.

Besides Proposition 2.16, we also need the following simple trigonometric identity:

$$\frac{1}{\tan(\beta)} - \frac{1}{\tan(2\beta)} = \frac{2 \cos^2(\beta) - \cos(2\beta)}{2 \sin(\beta) \cos(\beta)} = \frac{1}{\sin(2\beta)}. \quad (2.131)$$

By using (2.131), we obtain

$$\sum_{n=1}^N \frac{1}{n \sin(\pi n \alpha)} = \sum_{n=1}^N \frac{1}{n \tan(\pi n \alpha / 2)} - \sum_{n=1}^N \frac{1}{n \tan(\pi n \alpha)},$$

and combining this with Proposition 2.16, we get the equation

$$\sum_{n=1}^N \frac{1}{n \sin(\pi n \alpha)} = 2\pi M_\alpha(N) - 2\pi M_{\alpha/2}(N) + O(1). \quad (2.132)$$

If α is a quadratic irrational, then $\alpha/2$ is also a quadratic irrational; therefore, combining Eq. (2.132) with Proposition 2.1, we obtain

Proposition 2.17. *If α is a quadratic irrational, then there is a constant $c^* = c^*(\alpha)$ such that*

$$\sum_{n=1}^N \frac{1}{n \sin(\pi n \alpha)} = c^* \cdot \log N + O(1), \quad (2.133)$$

where the constant factor $c^* = c^*(\alpha)$ can be determined by using (2.132) and Proposition 2.16.

Now we are in a position to understand why the constant factor $c^*(\alpha)$ in (2.133) equals 0 for $\alpha = \sqrt{2}$, and why in general it equals 0 for any $\alpha = \sqrt{m^2 + 1}$ where

$m \geq 1$ is an odd integer. The advantage of $\alpha = \sqrt{m^2 + 1}$ is that it has a particularly simple continued fraction: $\alpha = [m; 2m, 2m, 2m, \dots] = [m; \overline{2m}]$ and

Case 1: if m is odd, then $\alpha/2 = [(m-1)/2; \overline{1, 1, m-1}]$;

Case 2: if m is even, then $\alpha/2 = [m/2; \overline{4m, m}]$.

In Case 1 both α and $\alpha/2$ have periods of *odd* length, so by Proposition 2.1 and (2.132), the partial sums of the series (2.128) are $O(1)$.

On the other hand, in Case 2, $\alpha/2$ has a period of *even* length, so the partial sums of the series (2.128) have the form $c^*(\alpha) \log N + O(1)$ where $c^*(\alpha)$ is *never* zero. Now we clearly understand why in the “ $O(1)$ -theorem” of Hardy and Littlewood the condition “ m is odd” was necessary. Indeed, if $\alpha = \sqrt{m^2 + 1}$ and m is even, then there is no $O(1)$ -theorem: by (2.132) and Case 2 above,

$$\begin{aligned} \sum_{n=1}^N \frac{1}{n \sin(\pi n \alpha)} &= O(1) - 2\pi M_{\alpha/2}(N) = \\ &= \frac{2\pi}{12} \cdot \frac{4m-m}{2} \cdot \frac{\log N}{\log(m + \sqrt{m^2 + 1})} + O(1) = \\ &= \frac{\pi m}{4 \log(m + \sqrt{m^2 + 1})} \log N + O(1), \end{aligned}$$

since $x = m$ and $y = 1$ is the least solution of Pell’s equation $x^2 - (m^2 + 1)y^2 = \pm 1$.

In view of (2.132) it is natural to ask the following related question: How to compute the continued fraction for $\alpha/2$ from the continued fraction for α ? Well, if $\alpha = [a_0; a_1, a_2, a_3, \dots]$ then

$$\alpha/2 = [a_0/2; 2a_1, a_2/2, 2a_3, a_4/2, \dots, a_{2i}/2, 2a_{2i+1}, \dots]$$

if this formula *does* make sense, i.e., if a_{2i} is even for every $i \geq 0$. Under this “parity condition,” by using (2.132) and Proposition 2.16, it is very easy to characterize those quadratic irrationals for which the partial sums of the series (2.128) are $O(1)$.

Indeed, if the length s of the period $a_{j+1}, a_{j+2}, \dots, a_{j+s}$ of α is *odd*, then the necessary and sufficient condition for an “ $O(1)$ -theorem” is $\sum_{i=j+1}^{j+s} (-1)^i a_i = 0$.

On the other hand, if the length of the period is *even*, then there is no “ $O(1)$ -theorem” whatsoever.

For example, if $\alpha = \sqrt{41} = [6; 2, 2, 12, 2, 2, 12, \dots] = [6; \overline{2, 2, 12}]$ then the “parity condition” holds:

$$\frac{\alpha}{2} = \frac{\sqrt{41}}{2} = [3; 4, 1, 24, 1, 4, 6, 4, 1, 24, 1, 4, 6, \dots] = [3; \overline{4, 1, 24, 1, 4, 6}],$$

and by (2.132) we have

$$\begin{aligned}
 \sum_{n=1}^N \frac{1}{n \sin(\pi n \alpha)} &= O(1) - 2\pi M_{\alpha/2}(N) = \\
 &= \frac{2\pi}{12} \cdot \frac{4 - 1 + 24 - 1 + 4 - 6}{6} \cdot \frac{\log N}{3 \log(32 + 5\sqrt{41})} + O(1) = \\
 &= \frac{2\pi \log N}{9 \log(32 + 5\sqrt{41})} + O(1),
 \end{aligned}$$

since $x = 32$ and $y = 5$ is the least solution of Pell's equation $x^2 - 41y^2 = \pm 1$.

The general case, when the “parity condition” is violated, is technically more complicated and somewhat unpleasant. We guess that this technical difficulty was the reason why Hardy and Littlewood restricted their study to the very special quadratic irrationals $\alpha = \sqrt{m^2 + 1} = [m; \overline{2m}]$ having the simplest possible (“one digit period”) continued fraction.

How to obtain the continued fraction for $\alpha/2$ in general, assuming we know $\alpha = [a_0; a_1, a_2, a_3, \dots]$? There is an interesting general procedure to answer this question, even when the “parity condition” is violated. We learned it from Richard Bumby (Rutgers University), an expert in continued fractions, who claims that the procedure goes back to Hurwitz. What Hurwitz was really interested in was to find the continued fraction for $e/2$ and $2e$, based on the knowledge of Euler's classical solution for e :

$$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, \dots, 1, 2i, 1, \dots]. \quad (2.134)$$

2.3.3 Doubling and Halving in Continued Fractions

The procedure consists of three operations. The first two, H = “halving,” D = “doubling,” are perfectly natural; the third, S = “special operation,” is the tricky one. For example, to get the continued fraction for $e/2$, first we apply the “halving operation” H to the first “digit” 2 in (2.134): this gives 1, and next comes the “doubling operation” D applied to the second “digit” 1 in (2.134), and so on. There are nine rules.

1. $H(2n) = nD$ (i.e., D comes next)
2. $Dn = (2n)H$
3. $H(2n + 1) = n, 1S$
4. $Dn, 1 = (2n + 1)S$
5. $S(2n) = 1, n - 1, 1S$
6. $S(2n + 1) = 1, nD$

7. $S1, n = (2n + 1)H$
8. $S1, n, 1 = (2n + 2)S$
9. $S2 = 2S$

Note that rules 1 and 2 are obvious, but the rest of the rules require a little bit of work with continued fraction. For example, to prove rule 3, we may proceed as follows ($n \geq 1, m \geq 1$ are integers and $x > 1$ is a real):

$$\begin{aligned} \frac{(2n+1) + \frac{1}{m+\frac{1}{x}}}{2} &= n + \frac{1}{2} + \frac{1}{2(m+\frac{1}{x})} = n + \frac{m+1+\frac{1}{x}}{2m+\frac{2}{x}} = \\ &= n + \frac{1}{\frac{2m+\frac{2}{x}}{m+1+\frac{1}{x}}} = n + \frac{1}{1 + \frac{m-1+\frac{1}{x}}{m+1+\frac{1}{x}}} = n + \frac{1}{1 + \frac{1}{\frac{m+1+\frac{1}{x}}{m-1+\frac{1}{x}}}} = n + \frac{1}{1 + \frac{1}{1 + \frac{1}{\frac{m-1+\frac{1}{x}}{2}}}}. \end{aligned}$$

Assume now that $m = 2k + 1$ where $k \geq 1$ is an integer, then

$$\frac{(2n+1) + \frac{1}{m+\frac{1}{x}}}{2} = n + \frac{1}{1 + \frac{1}{1 + \frac{1}{k + \frac{1}{2x}}}},$$

which proves the combination of rules 3 and 6. Similar argument proves the rest of the cases—we leave the details to the reader.

We illustrate the application of these rules by determining the continued fractions of $e/2$ and $2e$ (first published by Hurwitz).

To get $e/2$ we proceed on the “digits” in (2.134); we start with the “halving operation” applied on 2 (the first “digit” of e):

- H2 \implies rule 1 \implies 1 (D comes next)
- D1 \implies rule 2 \implies 2 (H comes next)
- H2 \implies rule 1 \implies 1 (D comes next)
- D1,1 \implies rule 4 \implies 3 (S comes next)
- S4 \implies rule 5 \implies 1,1,1 (S comes next)
- S1,1 \implies rule 7 \implies 3 (H comes next)
- H6 \implies rule 1 \implies 3 (D comes next)
- D1,1 \implies rule 4 \implies 3 (S comes next)
- S8 \implies rule 5 \implies 1,3,1 (S comes next)
- S1,1 \implies rule 7 \implies 3 (H comes next)
- H(10) \implies rule 1 \implies 5 (D comes next)
- D1,1 \implies rule 4 \implies 3 (S comes next)
- S(12) \implies rule 5 \implies 1,5,1 (S comes next)
- S1,1 \implies rule 7 \implies 3 (H comes next)

and so on. We applied the following rules:

1, 3, 1, 4, 5, 7, 1, 4, 5, 7, 1, 4, 5, 7, 1, 4, 5, 7, ...

This sequence shows periodicity; the period is 1,4,5,7, and we obtain

$$e/2 = [1; 2, 1, 3, 1, 1, 1, 3, 3, 3, 1, 3, 1, 3, 5, 3, 1, 5, 1, 3, \dots]. \quad (2.135)$$

It is easy to recognize the linear pattern in (2.135):

$$e/2 = [1; 2, 1, 3, 1, 1, 1, 3, 3, 3, 1, 3, 1, 3, 5, 3, 1, 5, 1, 3, \dots, 2i + 1, 3, 1, 2i + 1, 1, 3, \dots].$$

Similarly, to get $2e$ we proceed on the “digits” in (2.134), but of course here we start with the “doubling operation” applied on 2:

D2,1 \implies rule 4 \implies 5 (S comes next)
 S2 \implies rule 9 \implies 2 (S comes next)
 S1,1 \implies rule 7 \implies 3 (H comes next)
 H4 \implies rule 1 \implies 2 (D comes next)
 D1,1 \implies rule 4 \implies 3 (S comes next)
 S6 \implies rule 5 \implies 1,2,1 (S comes next)
 S1,1 \implies rule 7 \implies 3 (H comes next)
 H8 \implies rule 1 \implies 4 (D comes next)
 D1,1 \implies rule 4 \implies 3 (S comes next)
 S(10) \implies rule 5 \implies 1,4,1 (S comes next)
 S1,1 \implies rule 7 \implies 3 (H comes next)
 H(12) \implies rule 1 \implies 6 (D comes next)
 D1,1 \implies rule 4 \implies 3 (S comes next)
 S(14) \implies rule 5 \implies 1,6,1 (S comes next)

and so on. We applied the following rules:

$$4, 9, 7, 1, 4, 5, 7, 1, 4, 5, 7, 1, 4, 5, 7, 1, 4, 5, \dots$$

This sequence shows periodicity with the *same* period as for $e/2$, and we obtain

$$2e = [5; 2, 3, 2, 3, 1, 2, 1, 3, 4, 3, 1, 4, 1, 3, 6, 3, 1, 6, 1, \dots].$$

It is easy to recognize the linear pattern here:

$$2e = [5; 2, 3, 2, 3, 1, 2, 1, 3, 4, 3, 1, 4, 1, 3, 6, 3, 1, 6, 1, \dots, 2i, 3, 1, 2i, 1, 3, \dots]. \quad (2.136)$$

2.3.4 A Geometric Interpretation

We conclude Sect. 2.3 with the interesting observation that the partial sums of the Hardy–Littlewood series [see (2.128)]

$$\sum_{n=1}^N \frac{1}{n \sin(\pi n \alpha)}, \quad \alpha \text{ is irrational}, \quad (2.137)$$

have a nice geometric meaning: the partial sums represent the “average error” in yet another natural lattice point counting problem. To justify this claim, we go back to Sect. 1.2, where we counted lattice points inside the axes-parallel right triangle bounded with the lines $y = \alpha x$, $y = 0$, $x = n$ (we excluded the lattice points on the boundary). Here we slightly modify the problem: let $0 < \rho < 1$, we shift the line $y = \alpha x$ to the parallel line $y = \alpha(x - \rho)$ passing through the point $(\rho, 0)$ —this point is the left corner of our new triangle; the lines $y = 0$, $x = n$ remain unchanged. In other words, we just shift the left corner of the right triangle from the origin $(0, 0)$ to $(\rho, 0)$. Counting the lattice points inside the new triangle vertically, we obtain the following sum [an analog of (1.47)]:

$$\begin{aligned} & \lfloor \alpha - \rho\alpha \rfloor + \lfloor 2\alpha - \rho\alpha \rfloor + \lfloor 3\alpha - \rho\alpha \rfloor + \cdots + \lfloor (n-1)\alpha - \rho\alpha \rfloor = \\ &= \sum_{k=1}^{n-1} \left(k\alpha - \rho\alpha - \frac{1}{2} - \left(\{k\alpha\} - \frac{1}{2} \right) \right) = \\ &= E_{\alpha,\rho}^*(n-1) - S_{\alpha,\rho}^*(n-1), \end{aligned} \quad (2.138)$$

where

$$E_{\alpha,\rho}^*(m) = \alpha \binom{m+1}{2} - m \left(\rho\alpha + \frac{1}{2} \right)$$

and

$$S_{\alpha,\rho}^*(m) = \sum_{k=1}^m ((k\alpha - \rho\alpha)).$$

Just like in Sect. 1.2, we consider $E_{\alpha,\rho}^*(n-1)$ the “expectation,” and $S_{\alpha,\rho}^*(n-1)$ is the “error term” (i.e., the deviation from the expected value). By using the Fourier series of the sawtooth function [see (2.118)], we have

$$((x - \rho\alpha)) = - \sum_{j=1}^{\infty} \frac{\sin(2\pi j(x - \rho\alpha))}{\pi j},$$

and so we have the (formal) equation

$$\begin{aligned} S_{\alpha,\rho}^*(m) &= - \sum_{j=1}^{\infty} \frac{1}{\pi j} \sum_{k=1}^m \sin(2\pi j(k\alpha - \rho\alpha)) = \\ &= - \sum_{j=1}^{\infty} \frac{1}{\pi j} \cdot \frac{\cos(2\pi j\alpha(\frac{1}{2} - \rho)) - \cos(2\pi j\alpha(\frac{m+1}{2} - \rho))}{2 \sin(\pi j\alpha)}. \end{aligned}$$

Now we choose $\rho = 1/2$, that is, the left corner of our right triangle is the point $(1/2, 0)$ (instead of the origin). Then

$$S_{\alpha, 1/2}^*(m) = \sum_{j=1}^{\infty} \frac{\cos(2\pi m j \alpha) - 1}{2\pi j \sin(\pi j \alpha)}, \quad (2.139)$$

implying that in the average

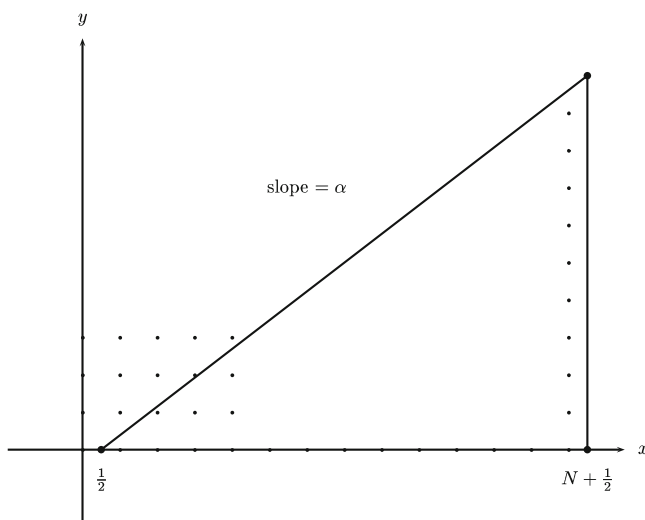
$$M_{\alpha, 1/2}^*(N) = \frac{1}{N} \sum_{m=1}^N S_{\alpha, 1/2}^*(m)$$

we have the new factor $\sin(\pi j \alpha)$ in the denominator instead of $\tan(\pi j \alpha)$ that we have in $M_{\alpha}(N)$; see (2.125), (2.126) and (2.139). Now assume that α is badly approximable; then the proof of Proposition 2.16 can be easily adapted for the similar $M_{\alpha, 1/2}^*(N)$, and it gives the following analog of (2.130):

$$M_{\alpha, 1/2}^*(N) = - \sum_{j=1}^N \frac{1}{2\pi j \sin(\pi j \alpha)} + O(1), \quad (2.140)$$

where the implicit constant $O(1) = O_{\alpha}(1)$ is independent of N .

Comparing (2.137) to (2.140), we see the geometric interpretation of the initial segment of the Hardy–Littlewood series. It represents the “average error” in a lattice point counting problem. Namely, counting lattice points in axes-parallel right triangles of slope α (where α is badly approximable), bounded by the horizontal axis, where the left corner is the fixed half-integer point $(1/2, 0)$; see the picture below.



2.4 Fourier Series and a Problem of Hardy and Littlewood (II)

The whole section is devoted to the **proof of Proposition 2.16**. By using Lemma 2.15 with the choice $T \geq N \log N$, we have

$$M_\alpha(N) = - \sum_{j=1}^N \frac{1}{2\pi j \tan(\pi j \alpha)} - S_1 - S_2 + O(1), \quad (2.141)$$

where

$$S_1 = \sum_{j=1}^N \frac{\sin(2\pi j \alpha) - \sin(2\pi(N+1)j \alpha)}{4\pi N j \sin^2(\pi j \alpha)} \quad (2.142)$$

and

$$S_2 = \sum_{j=N+1}^T \frac{1}{2\pi j} \left(\frac{1}{\tan(\pi j \alpha)} + \frac{\sin(2\pi j \alpha) - \sin(2\pi(N+1)j \alpha)}{2\pi N \sin^2(\pi j \alpha)} \right). \quad (2.143)$$

Since the irrational rotation is uniformly distributed, we have the “plausible” approximation

$$\frac{1}{M_2 - M_1} \sum_{M_1 \leq k < M_2} f(k\alpha) \approx \int_0^1 f(x) dx, \quad (2.144)$$

where $f(x)$ is a “nice” periodic function with period one. We can make the “plausible” approximation (2.144) precise by using the so-called Koksma’s inequality.

Lemma 2.18 (“Koksma’s inequality”). *Let $\mathcal{X} = \{x_1, \dots, x_n\}$ be an arbitrary n -element point set in the unit interval $[0, 1)$, then*

$$\left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \int_0^1 f(x) dx \right| \leq \frac{\Delta(\mathcal{X})}{n} \int_0^1 |f'(x)| dx,$$

where of course f' is the derivative of f (i.e., we assume that f is smooth), and

$$\Delta(\mathcal{X}) = \sup_{0 < y \leq 1} \left| \sum_{x_i \leq y} 1 - ny \right| \quad (2.145)$$

is the discrepancy of the set \mathcal{X} .

Notice that the *discrepancy* defined in (2.145) measures the deviation of $\mathcal{X} = \{x_1, \dots, x_n\}$ from the perfect uniform distribution in the unit interval. The integral

$$\int_0^1 |f'(x)| dx$$

is usually called the *variation* of f .

Proof of Lemma 2.18. Assume that the elements of \mathcal{X} are in increasing order: $0 \leq x_1 \leq x_2 \leq \dots \leq x_n \leq 1$. Using integration by parts, we have

$$\int_0^1 f(x) dx = f(1) - \int_0^1 x f'(x) dx. \quad (2.146)$$

The discrete analog of (2.146) is

$$\frac{1}{n} \sum_{i=1}^n f(x_i) = f(1) - \sum_{i=1}^n \frac{i}{n} (f(x_{i+1}) - f(x_i)), \quad (2.147)$$

where $x_{n+1} = 1$; Equation (2.147) is a routine application of Abel's transformation (2.119). Putting $x_0 = 0$, by (2.146) and (2.147),

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \int_0^1 f(x) dx \right| &= \left| \int_0^1 x f'(x) dx - \sum_{i=1}^n \frac{i}{n} (f(x_{i+1}) - f(x_i)) \right| \leq \\ &\leq \sum_{i=0}^n \int_{x_i}^{x_{i+1}} \left| x - \frac{i}{n} \right| |f'(x)| dx \leq \frac{\Delta(\mathcal{X})}{n} \int_0^1 |f'(x)| dx, \end{aligned}$$

and Lemma 2.18 follows. \square

It is easy to rescale Lemma 2.18 to any interval $[a, b]$: if $a \leq x_1 \leq x_2 \leq \dots \leq x_n \leq b$ then

$$\left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \frac{1}{b-a} \int_a^b f(x) dx \right| \leq \frac{\Delta}{n} \int_a^b |f'(x)| dx, \quad (2.148)$$

where

$$\Delta = \sup_{a < y \leq b} \left| \sum_{x_i \leq y} 1 - n \frac{y-a}{b-a} \right|, \quad (2.149)$$

an analog of the discrepancy in (2.145).

Let's return to the Discrepancy Lemma in Sect. 1.1 [see (1.22) and (1.23)]: it implies that the discrepancy of the irrational rotation $k\alpha \pmod{1}$, $1 \leq k \leq n$,

is $O(\log n)$, if α is badly approximable. Next we show that, for a given interval $I \subset [0, 1]$, we can replace the upper bound $O(\log n)$ for the discrepancy in I with $O(\log(n|I| + 2))$. This is a substantial improvement if $|I|$ is “close” to $1/n$.

Lemma 2.19. *If α is badly approximable then*

$$\mathcal{Z}_\alpha(n; I) = \sum_{\substack{1 \leq k \leq n: \\ k\alpha \in I \pmod{1}}} 1 = n|I| + O(\log(n|I| + 2)).$$

Proof. We repeat the argument in (1.15)–(1.21) with a twist at the end. Assume $q_{\ell-1} \leq n < q_\ell$; in view of (1.20) we can write

$$n = b_{\ell-1}q_{\ell-1} + b_{\ell-2}q_{\ell-2} + \dots + b_1q_1,$$

where $1 \leq b_{\ell-1} \leq a_{\ell-1}$, $0 \leq b_j \leq a_j$ for $2 \leq j < \ell - 1$, $0 \leq b_1 \leq a_1 - 1$, and

$$\sum_{i=1}^{j-1} b_i q_i < q_j \quad \text{for } 1 \leq j \leq \ell.$$

Let r be the largest index j such that $\|q_j \alpha\| > |I|$, and write $n = M + m$ where

$$M = b_{\ell-1}q_{\ell-1} + b_{\ell-2}q_{\ell-2} + \dots + b_r q_r \quad \text{and} \quad m = b_{r-1}q_{r-1} + b_{r-2}q_{r-2} + \dots + b_1 q_1 < q_r.$$

By (1.22) and (1.23),

$$|\mathcal{Z}_\alpha(M; I) - M|I|| \leq 3(b_{\ell-1} + b_{\ell-2} + \dots + b_r). \quad (2.150)$$

Notice that the end sequence $(M + j)\alpha \pmod{1}$, $1 \leq j \leq m$, of the irrational rotation contains at most one member in the interval I . Indeed, otherwise there exist $n_1 < n_2$ such that $1 \leq n_2 - n_1 < m < q_r$ with $n_i \alpha \in I \pmod{1}$, $i = 1, 2$, and so $\|(n_2 - n_1)\alpha\| \leq |I| < \|q_r \alpha\|$. But this contradicts the following well-known *minimum property* of the convergent denominators q_j of α : $\|p\alpha\| < \|q_j \alpha\|$ implies that $p > q_j$. Thus we have

$$\mathcal{Z}_\alpha(n; I) = \mathcal{Z}_\alpha(M; I) + O(1),$$

and so by (2.150),

$$\begin{aligned} |\mathcal{Z}_\alpha(n; I) - n|I|| &= |\mathcal{Z}_\alpha(M; I) - M|I| + O(1) - n|I|| \leq \\ &\leq |\mathcal{Z}_\alpha(M; I) - M|I|| + O(1) + n|I| = \\ &= \left(\max_{r \leq j < \ell} b_j \right) \cdot O(\ell - r) + m|I| = O(\ell - r) + O(1). \end{aligned} \quad (2.151)$$

In the last step we used that α is badly approximable, and also

$$m|I| < q_r|I| < q_r\|q_r\alpha\| = O(1),$$

where in the last step we used (1.9).

Again using the fact that α is badly approximable, we have

$$n \approx q_\ell \quad \text{and} \quad |I| \approx \|q_r\alpha\| \approx \frac{1}{q_r},$$

implying

$$\frac{q_\ell}{q_r} \approx n|I| \quad \text{and} \quad \ell - r = O(\log(n|I| + 2)). \quad (2.152)$$

Combining (2.151) and (2.152), Lemma 2.19 follows. \square

Now we are ready to estimate S_2 in (2.143). We define the set

$$A_k = \left\{ N < j \leq T : \frac{2^k}{N} \leq \|j\alpha\| < \frac{2^{k+1}}{N} \right\}. \quad (2.153)$$

Depending on whether $\{j\alpha\}$ is small or $1 - \{j\alpha\}$ is small, we split A_k into two parts: $A_k = A_k^+ \cup A_k^-$. More precisely, let $\|x\|^+ = \|x\|$ if the interval $(x - 1/2, x]$ contains an integer and 0 otherwise, and similarly let $\|x\|^- = \|x\|$ if the interval $(x, x + 1/2]$ contains an integer and 0 otherwise. Then $\|x\| = \|x\|^+ + \|x\|^-$, and write

$$A_k^+ = \left\{ N < j \leq T : \frac{2^k}{N} \leq \|j\alpha\|^+ < \frac{2^{k+1}}{N} \right\} \quad (2.154)$$

and

$$A_k^- = \left\{ N < j \leq T : \frac{2^k}{N} \leq \|j\alpha\|^- < \frac{2^{k+1}}{N} \right\}. \quad (2.155)$$

The proof of Proposition 2.16 proceeds in several steps: Step One, Step Two, and so on.

Step One: We estimate the sum

$$\sum_{j \in A_k} \frac{1}{j \tan(\pi j \alpha)} \quad \text{for every } k \geq 1, \quad (2.156)$$

where A_k is defined in (2.153). For technical reasons, we decompose A_k into several parts: for $1 \leq \ell \leq k^2$ let

$$A_{k,\ell} = \left\{ j \in A_k : N \cdot \left(1 + \frac{1}{k^2}\right)^{\ell-1} < j \leq N \cdot \left(1 + \frac{1}{k^2}\right)^\ell \right\}, \quad (2.157)$$

and for $\ell \geq k^2 + 1$ let

$$A_{k,\ell} = \{j \in A_k : N(k, \ell - 1) < j \leq N(k, \ell)\}, \quad (2.158)$$

where

$$N(k, k^2) = N \cdot \left(1 + \frac{1}{k^2}\right)^{k^2} \quad \text{and} \quad N(k, \ell) = N(k, \ell-1) \left(1 + \frac{1}{\ell}\right). \quad (2.159)$$

Again we split

$$A_{k,\ell} = A_{k,\ell}^+ \cup A_{k,\ell}^- \quad (2.160)$$

exactly the same way as we did in (2.153)–(2.155).

We estimate the sum

$$\sum_{j \in A_{k,\ell}} \frac{1}{j \tan(\pi j \alpha)}$$

by using Lemmas 2.18 and 2.19. The reason why we defined the “short” sets $A_{k,\ell}$ is that the factor j hardly changes in such a short set. We apply Eq. (2.148) with

$$a = \frac{2^k}{N}, \quad b = \frac{2^{k+1}}{N}, \quad f(x) = \frac{1}{\tan(\pi x)}, \quad (2.161)$$

and the finite point set \mathcal{X} in the interval $[a, b]$ is the following [see (2.160)]:

$$\mathcal{X} = \{j\alpha \pmod{1} : j \in A_{k,\ell}^+\}; \quad (2.162)$$

then we have

$$\left| \sum_{j \in A_{k,\ell}^+} \frac{1}{\tan(\pi j \alpha)} - \frac{|A_{k,\ell}^+|}{b-a} \int_a^b f(x) dx \right| \leq \Delta \int_a^b |f'(x)| dx. \quad (2.163)$$

Notice the difference between (2.156) and (2.163): in the latter factor j is missing from the denominator.

Write

$$E(k, \ell) = (N(k, \ell) - N(k, \ell - 1)) \frac{2^k}{N}, \quad (2.164)$$

where $N(k, \ell)$ is defined in (2.159) for $\ell \geq k^2$, and

$$N(k, \ell) = N \cdot \left(1 + \frac{1}{k^2}\right)^\ell \quad \text{for } 0 \leq \ell < k^2. \quad (2.165)$$

We may call $E(k, \ell)$ the “expectation,” because by Lemma 2.19,

$$\Delta = O(\log(E(k, \ell) + 2)), \quad (2.166)$$

and

$$\int_a^b |f'(x)| dx = |f(b) - f(a)|, \quad (2.167)$$

because $f(x) = (\tan(\pi x))^{-1}$ is monotonic in $a \leq x \leq b$ as long as $2^k \leq N/4$. Combining (2.161)–(2.167), we have

$$\sum_{j \in A_{k,\ell}^+} \frac{1}{\tan(\pi j \alpha)} = \frac{E(k, \ell)}{b - a} \int_a^b f(x) dx + O(N 2^{-k} \log(E(k, \ell) + 2)). \quad (2.168)$$

We repeat the same argument for $A_{k,\ell}^-$: the only difference is that $a_1 = 1 - 2^{k+1}N^{-1}$ and $b_1 = 1 - 2^k N^{-1}$ are the new endpoints instead of a, b in (2.161). Thus we have the analog of (2.168):

$$\sum_{j \in A_{k,\ell}^-} \frac{1}{\tan(\pi j \alpha)} = \frac{E(k, \ell)}{b_1 - a_1} \int_{a_1}^{b_1} f(x) dx + O(N 2^{-k} \log(E(k, \ell) + 2)). \quad (2.169)$$

Since $\tan(x)$ is an odd function,

$$\int_{a_1}^{b_1} f(x) dx + \int_a^b f(x) dx = 0,$$

and by (2.168) and (2.169),

$$\sum_{j \in A_{k,\ell}} \frac{1}{\tan(\pi j \alpha)} = O(N 2^{-k} \log(E(k, \ell) + 2)). \quad (2.170)$$

By (2.157)–(2.159), if $j_1, j_2 \in A_{k,\ell}$ then with $j_1 < j_2$ we have

$$1 < \frac{j_2}{j_1} \leq 1 + \frac{1}{k^2} \quad \text{if } \ell \leq k^2 \quad (2.171)$$

and

$$1 < \frac{j_2}{j_1} \leq 1 + \frac{1}{\ell} \text{ if } \ell > k^2. \quad (2.172)$$

By (2.170)–(2.172), we can control the effect of the extra factor of j in the denominator as follows:

$$\begin{aligned} \sum_{j \in A_{k,\ell}} \frac{1}{j \tan(\pi j \alpha)} &= O((N(k, \ell - 1))^{-1} N 2^{-k} \log(E(k, \ell) + 2)) + \\ &+ \min\{k^{-2}, \ell^{-2}\} \sum_{j \in A_{k,\ell}} \frac{1}{j |\tan(\pi j \alpha)|}. \end{aligned} \quad (2.173)$$

By the definition of $A_{k,\ell}$ [see (2.153)–(2.159)]

$$\sum_{j \in A_{k,\ell}} \frac{1}{j |\tan(\pi j \alpha)|} \leq \frac{|A(k, \ell)|}{N(k, \ell - 1)} \cdot N 2^{-k},$$

so by (2.173) we have

$$\sum_{j \in A_{k,\ell}} \frac{1}{j \tan(\pi j \alpha)} = O(H_{k,\ell}) \quad (2.174)$$

where

$$H_{k,\ell} = (N(k, \ell - 1))^{-1} N 2^{-k} (\log(E(k, \ell) + 2) + \min\{k^{-2}, \ell^{-2}\} \cdot |A_{k,\ell}|). \quad (2.175)$$

By (2.164)–(2.166),

$$\begin{aligned} \sum_{\ell \geq 1} H_{k,\ell} &= \sum_{1 \leq \ell \leq k^2} H_{k,\ell} + \sum_{k^2 < \ell} H_{k,\ell} = \\ &= \sum_{1 \leq \ell \leq k^2} 2^{-k} (O(k) + O(k^{-4} 2^k)) + \sum_{k^2 < \ell} 2^{-k} e^{-\ell/k^2} (O(\ell k^{-2} + k) + O(\ell^{-2} e^{\ell/k^2})) = \\ &= O(k^{-2}). \end{aligned} \quad (2.176)$$

Combining (2.175) and (2.176), for every $k \geq 1$ we have

$$\sum_{j \in A_k} \frac{1}{j \tan(\pi j \alpha)} = O(k^{-2}), \quad (2.177)$$

which completes Step One.

Adding up (2.177) for all $k = 1, 2, 3, \dots$ we have

$$\sum_{\substack{N < j \leq T: \\ \|j\alpha\| \geq 2/N}} \frac{1}{j \tan(\pi j\alpha)} = O\left(\sum_{k=1}^{\infty} k^{-2}\right) = O(1). \quad (2.178)$$

Of course, if $\|j\alpha\|$ is “around” $1/N$, then the method of Step One still works, for example,

$$\sum_{\substack{N < j \leq T: \\ 2/N > \|j\alpha\| \geq 1/16N}} \frac{1}{j \tan(\pi j\alpha)} = O(1), \quad (2.179)$$

but if $\|j\alpha\|$ is much smaller than $1/N$, then we switch to

Step Two: Let

$$B_k = \left\{ N < j \leq T : \frac{1}{2^k N} \geq \|j\alpha\| > \frac{1}{2^{k+1} N} \right\}, \quad (2.180)$$

then we estimate the sum [see (2.143)]

$$\sum_{j \in B_k} \left(\frac{1}{j \tan(\pi j\alpha)} + \frac{\sin(2\pi j\alpha) - \sin(2\pi(N+1)j\alpha)}{2jN \sin^2(\pi j\alpha)} \right)$$

for every $k \geq 4$. We repeat the argument of Step One with the new function

$$g(x) = \frac{1}{\tan(\pi x)} + \frac{\sin(2\pi x) - \sin(2\pi(N+1)x)}{2N \sin^2(\pi x)} \quad (2.181)$$

instead of $f(x) = 1/\tan(\pi x)$ [see (2.161)] that we used in Step One. Note that $g(x)$ is also odd (which is crucial for the cancellation part); and $g(x)$ is also monotonic at least in the interval $0 < x < 1/16N$; and

$$g(x) \approx \frac{2\pi}{3} N^2 x \text{ if } 0 < x < \frac{1}{16N}. \quad (2.182)$$

Applying the method of Step One with B_k and $g(x)$ instead of A_k and $f(x)$, and heavily relying on (2.182) (what we need is monotonicity: smaller $x = \|j\alpha\|$ leads to smaller $g(x)$), we obtain the following analog of (2.178):

$$\sum_{k \geq 4} \sum_{j \in B_k} \frac{g(j\alpha)}{j} = O(1). \quad (2.183)$$

This completes Step Two.

Let's return to S_2 in (2.143). In view of (2.178)–(2.183), the last step is

Step Three: We have to estimate the sum

$$\sum_{\substack{N < j \leq T: \\ \|j\alpha\| \geq 1/16N}} \frac{\sin(2\pi(N+1)j\alpha) - \sin(2\pi j\alpha)}{2jN \sin^2(\pi j\alpha)}. \quad (2.184)$$

Again we repeat the argument of Step One: this time with the function

$$h(x) = \frac{\sin(2\pi(N+1)x) - \sin(2\pi x)}{2N \sin^2(\pi x)}, \quad (2.185)$$

and as an analog of the set $A_{k,\ell}$ [see (2.157)–(2.159)], we introduce the new set $A_{k,\ell}^*$ defined as

$$\left\{ N \left(1 + \frac{1}{\log^2\left(\frac{T}{N} + 2\right)} \right)^{\ell-1} < j \leq N \left(1 + \frac{1}{\log^2\left(\frac{T}{N} + 2\right)} \right)^{\ell} : \frac{k}{16N} < \|j\alpha\| \leq \frac{k+1}{16N} \right\},$$

where $k = 1, 2, 3, \dots$ and $\ell = 1, 2, 3, \dots$. Similarly to Step One, we estimate the sum

$$\sum_{j \in A_{k,\ell}^*} \frac{h(j\alpha)}{j}$$

by combining Koksma's inequality [in fact, we use the form (2.148) and (2.149)] with Lemma 2.19 and taking advantage of the fact that the function $h(x)$ is odd (which gives the crucial cancellation); also we use the fact that the factor j hardly changes in the “short” set $A_{k,\ell}^*$. A simple calculation gives

$$\text{sum}(2.184) = O(1); \quad (2.186)$$

a key reason why (2.186) holds is that the square $\sin^2(\pi x)$ in the denominator of (2.185) implies the appearance of the convergent series $\sum_{k \geq 1} k^{-2} = O(1)$ (instead of the divergent harmonic series).

Summarizing, by (2.178)–(2.184) and (2.186) we have

$$S_2 \text{ in (2.143) } = O(1). \quad (2.187)$$

It remains to show that

$$S_1 \text{ in (2.142) } = O(1). \quad (2.188)$$

To prove (2.188) we don't need the sophisticated method of Step One; instead we can succeed by simply using the trivial upper bound

$$|S_1| \leq \sum_{k=1}^N \frac{1}{Nk\|k\alpha\|^2}. \quad (2.189)$$

By repeating the proof of Lemma 2.14 (Pigeonhole Principle), we obtain

$$\sum_{k=1}^N \frac{1}{Nk\|k\alpha\|^2} = O(1), \quad (2.190)$$

due to the fact that the square $\|k\alpha\|^2$ leads to the convergent series $\sum_{k \geq 1} k^{-2} = O(1)$ (instead of the divergent harmonic series). Combining (2.141)–(2.143) with (2.187)–(2.190), Proposition 2.16 follows.

□

The next section is a (very important) detour: it is a short essay about the paradigm of *determinism versus randomness*, providing a broader perspective for our main results, Theorems 1.1 and 1.2.

2.5 A Detour: The Giant Leap in Number Theory

2.5.1 Looking at the “Big Picture”

As we already said in the Preface, we did not choose the (otherwise catchy and quite fitting) subtitle *randomness of $\sqrt{2}$* to avoid misleading the reader. Our objective is *not* to prove the apparent “randomness” of the digit distribution of $\sqrt{2}$ (which, unfortunately, remains open). Nevertheless, this notorious and totally untouchable problem is a perfect illustration of what we like to call the “Giant Leap” in number theory.

Historically the first attempt to prove something *vaguely* similar to the apparent randomness of the digit distribution of $\sqrt{2}$ was a measure-theoretic result. About 100 years ago, in 1909 E. Borel proved that *almost every* real number is *normal* in all bases $b = 2, 3, \dots, 10, \dots$. Of course, *almost every* means “all but a set of Lebesgue measure zero,” and a real number is said to be *normal* in a particular base if every block of digits of any length occurs with the same density depending only on the length and the base. In particular, if the base is $b \geq 2$ and the length is $l \geq 1$ then the density is b^{-l} , that is, normality is an equidistribution property.

Unfortunately, the measure-theoretic approach says nothing about individual numbers such as $\sqrt{2}$ or π . This is why now, 100 years later, we still don't know any explicit example of a number that is normal in all bases (such a number is often called *absolutely normal*).

To be fair, we have to admit that there are some very indirectly defined numbers, such as the Chaitin's number—defined as the halting probability of a universal Turing machine—and the so-called Sierpinski's number (which gives a little bit of extra information beyond Borel's measure-theoretic existential proof), that are absolutely normal, but most mathematicians are not happy with them—they are not considered “properly explicit.” For example, the so-called Champernowne number, see below, is undoubtedly “properly explicit,” and perfectly satisfies everybody. The core problem is that we don't have a rigorous definition of “concrete example.” For example, Sierpinski, mainly a set theorist, has a very broad interpretation and considers everything “explicit” if it does not use the Axiom of Choice. Sierpinski's “explicit example” is the minimum of a bounded countable set of real numbers. For most number theorists this is some sort of cheating; they want something more explicit, something “similar” to the Champernowne number. We concede, at this point the discussion becomes very murky—so we just stop this inserted remark.

When we say we don't know any explicit example of an absolutely normal number, we mean that we don't have a rigorous mathematical proof. We have, however, a very convincing “experimental proof,” because there is an overwhelming numerical evidence that the famous special numbers, such as $\pi = 3.14\dots$, $e = 2.718\dots$, $\sqrt{2}$, $\sqrt{3}$, $\sqrt[3]{2}$, $\log 2$ (meaning the natural logarithm of 2), and $\log 3 / \log 2$ (meaning the base 2 logarithm of 3), are *all* absolutely normal.

We cannot help but insert here two historic remarks. One of the early (pure mathematical) experimentations with the electronic computer—in 1949 von Neumann and his group working on ENIAC, the first fully electronic computer—was to determine the first two thousand decimal digits of π and to carry out a statistical treatment of the digit distribution. The second remark is a prediction of the great Dutch mathematician L.E.J. Brouwer. Almost 100 years ago, well before the revolution of the electronic computer, Brouwer wanted to show an example of an “unsolvable” problem—or at least unsolvable in his lifetime—and he came up with the following question: In the decimal expression for π , do we ever come to a place where a thousand consecutive digits are all zero? The answer is still unknown (but of course we all expect a positive answer).

As illustration, here are the first 50 digits of π in bases 10 and 2:

$$\pi = 3.141592653589793238462643383279502884197169399375105820\dots$$

$$\pi = 11.00100100001111110110101010001000100001011010001100001\dots$$

And here are the first 50 digits of $\sqrt{2}$ in bases 10 and 2:

$$\sqrt{2} = 1.414213562373095048801688724209698078569671875376948\dots$$

$$\sqrt{2} = 1.011010100000100111100110011001111110011101111001100\dots$$

But much more is true—or seems to be true—here: according to Wolfram's book *A New Kind of Science* (especially Chap. 4), *every* single irrational special number

ever tried so far seems to be normal in all bases. This observation is supported by an enormous computational evidence. For example, the frequency of digit 7 among the first 10^n decimal digits of π is 8 %, 9.5 %, 9.7 %, 10.025 %, 9.980 %, 10.002 % as $n = 2, 3, 4, 5, 6, 7$ —the occurrence ratios for digit 7 seem to be converging to $\frac{1}{10}$.

The vaguely defined notion *special* number means a real number expressed in terms of standard mathematical functions. The rational numbers are trivial exceptions: they are eventually periodic in every base, and periodicity (i.e., the repetition of the same block) is the complete opposite of the equidistribution of the blocks.

Note that normality is much less than “randomness”: the number

$$0.123456789101112131415161718192021 \dots 99100101102 \dots$$

is normal in base 10 in spite of exhibiting a very clear and predictable anti-randomness pattern. The pattern is that the digits are those of all natural numbers in succession; this is called the Champernowne number. Is the Champernowne number normal in base 2 or base 3? No one knows.

Irrational special numbers seem to exhibit digit equidistribution (i.e., normality), and what is more, far beyond normality they all seem to exhibit “full-blown randomness,” including the trademark square root size fluctuation of the random walk (physicists call it the “square root law”). For example, a statistical analysis of the first 10 million decimal digits of π tells us something interesting. The frequencies of 0, 1, 2, ..., 9 differ from the expected number 10^6 by

$$-560, -667, 306, -36, 1093, 466, -663, 207, -186, 40.$$

Since the standard deviation of the corresponding binomial distribution $\sqrt{np(1-p)}$ with $n = 10^7$, $p = 1/10$ is 300, the fluctuations are close to what one would expect by the central limit theorem.

Among the first $2 \cdot 10^{11}$ (200 billion) decimal digits of π , the frequencies of 0, 1, 2, ..., 9 differ from the expected number $2 \cdot 10^{10}$ by

$$30841, -85289, 136978, 69393, -78309, -82947, -118485, 32406, 291044, \\ -130820;$$

the data are from Wolfram’s book, see p. 912. Now the standard deviation of the corresponding binomial distribution $\sqrt{np(1-p)}$ with $n = 2 \cdot 10^{11}$, $p = 1/10$ is roughly 135,000, and again the fluctuations are well predicted by the central limit theorem. We have similar data for $\sqrt{2}$. The decimal expansions of π and $\sqrt{2}$ seem to exhibit normality, or using an alternative probabilistic name: the law of large numbers, and what is much more, they also seem to exhibit the square root law, or perhaps even the delicate central limit theorem. (Note that these results, the law of large numbers, the square root law, and the central limit theorem, are the benchmarks of Probability Theory.)

Summarizing, we can say that for the “interesting” real numbers (or “special” numbers) the decimal expansion, and in general any base $b \geq 2$ expansion, either features a simple behavior (such as the periodicity for the rationals) or features full-blown “randomness” (which seems to be the case for all special irrationals ever tried). We refer to this striking phenomenon as the *Giant Leap*. What makes the Giant Leap so uniquely interesting is the sharp contrast between the overwhelming numerical evidence and the total lack of rigorous mathematical proof. We don’t even know whether or not each of the ten digits keeps occurring infinitely often in the decimal form of π (or $\sqrt{2}$, or e , etc.).

How come that these questions are mathematically untouchable? We are sure the reader’s first reaction is to turn to Probability Theory for help. But here is the big dilemma: the decimal expansion of π (or $\sqrt{2}$ or e) is an individual sequence, and traditional probability theory says nothing about the “randomness” of individual sequences. In fact, the basic idea of Kolmogorov’s axiomatic foundation for probability theory is to scrupulously avoid the notion of “individual random sequence,” and right now we simply do not have any workable, agreed-on definition of “randomness.”

Note that in the 1920s, before Kolmogorov’s axioms, von Mises made an attempt to come up with a definition, but his work remained incomplete and controversial (we can actually say that von Mises’s failure was a key motivation for Kolmogorov’s axiomatic approach). Von Mises’s basic idea was to express the apparent lack of successful gambling schemes in a formal definition for *random sequences*. Many years later Information Theory (Shannon) suggested the new idea to define randomness via inability to compress data. Combining Mises’s old idea with this new idea, people like Chaitin, Kolmogorov, Solomonoff, and Martin-Löf introduced and developed the notion of *algorithmic randomness*. An individual sequence of length n features *algorithmic randomness* if the program-size complexity (i.e., the length of the shortest program describing the sequence) is close to n (i.e., the length of the sequence). The intuitive meaning is that the sequence is “patternless”; we cannot really compress the information: we have to write down the *whole* sequence.

Notice that *algorithmic randomness* is an extremely restrictive notion. Any sequence generated by a simple program (i.e., every “long” sequence we know) can by definition never be algorithmically random. For example, we know very long initial segments of the decimal digits of $\sqrt{2}$ and π ; they are generated by simple programs. For $\sqrt{2}$ we have the ancient Babylonian Algorithm: let $a_0 = 1$ and define a sequence a_1, a_2, a_3, \dots inductively by letting

$$a_{n+1} = \frac{a_n + \frac{2}{a_n}}{2}, \quad n \geq 0. \quad (2.191)$$

The convergence $a_n \rightarrow \sqrt{2}$ is extremely rapid: the number of correct decimal digits doubles with each iteration. Since (2.191) is a very short program, the program-size complexity of the digit sequence of $\sqrt{2}$ is very low, so the *algorithmic randomness* of the digit sequence of $\sqrt{2}$ is also very low. This means the concept of *algorithmic randomness* is quite irrelevant in our quest for understanding the apparent randomness we clearly see in these digit sequences.

The message of von Mises's failure is that there is no "absolute randomness"; in each case one has to decide on a cutoff. For example, in this book we say "enough" and stop around the central limit theorem; this is where we draw the line in the infinite hierarchy of notions of randomness.

Most mathematicians would agree that "randomness up to the central limit theorem" is already a high, advanced level in the hierarchy.

For more readings about "randomness" and "random numbers," we recommend Chap. 3 in Knuth [Kn2].

In our search for finding further evidence supporting the Giant Leap, we switch now from the decimal expansion to the continued fraction. To represent a real number x as a continued fraction, first we take the integral part of x , then we take the reciprocal $1/\{x\}$ of the fractional part of x , write it as the sum of the integral part and the fractional part, then take the reciprocal of the fractional part, and keep repeating the process:

$$x = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \dots}}}, \quad (2.192)$$

or by using the space-saving notation, $x = [a_0; a_1, a_2, a_3, \dots]$. Note that continued fractions play a key role in diophantine approximation, in uniform distribution, and in the solution of some diophantine equations. Continued fractions provide another perfect illustration for the Giant Leap phenomenon. Indeed, for every "interesting" real number ever tried the continued fraction either has a simple behavior or it exhibits full-blown randomness.

Examples of Simple Behavior:

1. rational numbers have finite continued fraction;
2. quadratic irrationals, such as $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$, $\sqrt{6}$, $\sqrt{7}$, all have periodic continued fractions—here are a few examples:

$$\sqrt{2} = [1; 2, 2, 2, 2, 2, \dots],$$

$$\sqrt{3} = [1; 1, 2, 1, 2, 1, 2, \dots],$$

$$\sqrt{5} = [2; 4, 4, 4, 4, 4, \dots],$$

$$\sqrt{6} = [2; 2, 4, 2, 4, 2, 4, \dots],$$

$$\sqrt{7} = [2; 1, 1, 1, 4, 1, 1, 1, 4, 1, 1, 1, 4, \dots],$$

$$\frac{1 + \sqrt{5}}{2} = [1; 1, 1, 1, 1, 1, \dots],$$

where the last one, representing the golden ratio, has the simplest form. A more complicated example is

$$\sqrt{67} = [8; 5, 2, 1, 1, 7, 1, 1, 2, 5, 16, 5, 2, 1, 1, 7, 1, 1, 2, 5, 16, \dots],$$

where the period of $\sqrt{67}$ is the block $5, 2, 1, 1, 7, 1, 1, 2, 5, 16$ of length 10. Note that the length of the period of \sqrt{n} in general remains a big mystery. The maximum length of the period for \sqrt{n} can be asymptotically as large as (roughly) \sqrt{n} itself, or it can be very short like $\sqrt{65} = [8; 16, 16, 16, \dots]$, where the period has length one.

3. special number e and its “family”: we know from Euler that

$$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, \dots, 1, 2n, 1, \dots],$$

$$\sqrt{e} = [1; 1, 1, 1, 5, 1, 1, 9, 1, 1, 13, 1, \dots, 1, 4n + 1, 1, \dots],$$

$$e^2 = [7; 2, 1, 1, 3, 18, 5, 1, 1, 6, 30, \dots, 3n - 1, 1, 1, 3n, 12n + 6, \dots],$$

$$\sqrt[3]{e} = [1; 2, 1, 1, 8, 1, 1, 14, 1, 1, 20, 1, \dots, 1, 6n + 2, 1, \dots].$$

Notice that they all have a simple linear pattern. The list is in fact infinite, including all numbers of the form $e^{2/k}$ where $k \geq 1$ is an integer; for more about it, see, e.g., Lang [La]. By the way, the “simplest” member of the family is

$$\frac{e^2 - 1}{e^2 + 1} = [1, 3, 5, 7, 9, \dots, 2n + 1, \dots]$$

(when the integral part is zero, we often delete 0 and the semicolon from the beginning).

Examples of Random Behavior: The rest of the special numbers, including e^3 , all seem to exhibit full-blown randomness with a common limit distribution for the digits. Unlike the familiar decimal expansion, where we have ten possible digits, in the continued fraction the j th digit a_j (often called the j th partial quotient) can be any integer ≥ 1 , so equidistribution does not make any sense. The particular limit distribution for the continued fraction comes from the invariant measure of the relevant mapping

$$T : x \rightarrow \{1/x\}, \tag{2.193}$$

which maps the open unit interval $(0,1)$ onto itself. Note that T is *not* one-to-one: the inverse image of an interval (a, b) , where $0 < a < b < 1$, is the infinite union of disjoint intervals

$$\left(\frac{1}{1+b}, \frac{1}{1+a}\right), \left(\frac{1}{2+b}, \frac{1}{2+a}\right), \left(\frac{1}{3+b}, \frac{1}{3+a}\right), \dots; \quad (2.194)$$

each one of these intervals is mapped to the whole (a, b) by T .

If we define the measure of an interval (a, b) to be

$$m(a, b) = \frac{1}{\log 2} \int_a^b \frac{dx}{1+x} = \frac{1}{\log 2} \log \frac{1+b}{1+a}, \quad (2.195)$$

then one can easily check that this m -measure of the interval (a, b) equals the sum of the m -measures of the intervals in (2.194). We can extend (2.195) to any measurable set $A \subset (0, 1)$ by the integral

$$m(A) = \frac{1}{\log 2} \int_A \frac{dx}{1+x}, \quad (2.196)$$

where \log stands for the natural (base e) logarithm. Measure (2.195) and (2.196) was already known to Gauss (who, for number-theoretic reasons, carried out an extensive numerical experimentation on continued fractions). The key property of measure (2.195) and (2.196) is that it is preserved by the transformation T . By definition the first partial quotient a_1 of a real $x \in (0, 1)$ equals an integer $k \geq 1$ if and only if x falls into the interval $(\frac{1}{k+1}, \frac{1}{k})$, which has m -measure

$$\begin{aligned} \frac{1}{\log 2} \int_{1/(k+1)}^{1/k} \frac{dx}{1+x} &= \frac{1}{\log 2} \left(\log\left(1 + \frac{1}{k}\right) - \log\left(1 + \frac{1}{k+1}\right) \right) = \\ &= \frac{\log \frac{(k+1)^2}{k(k+2)}}{\log 2} = \frac{1}{\log 2} \log \left(1 + \frac{1}{k(k+2)} \right). \end{aligned} \quad (2.197)$$

A well-known theorem of Kusmin states that, for almost every $x \in (0, 1)$, the density with which an arbitrary integer $k \geq 1$ appears in the sequence a_1, a_2, a_3, \dots of partial quotients in (2.192) is exactly (2.197). For example, for almost every $x \in (0, 1)$, the density of the digit 1 is exactly

$$\frac{\log(4/3)}{\log 2} = 0.415\dots \approx 41.5\%. \quad (2.198)$$

It was realized later that both Borel's theorem and Kusmin's theorem are special cases of the very general Ergodic Theorem of Birkhoff. Note, however, that Birkhoff's general theorem doesn't give any error term; on the other hand, in Borel's theorem and also in Kusmin's theorem we can prove a basically square root size error term (the sharpest form of Borel's theorem is the well-known Law of the Iterated Logarithm).

Kusmin's theorem clearly fails for $x = e$ [where the frequency of the digit 1 is $2/3$, which differs from the 41.5 % in (2.198)] and fails for the quadratic irrationals (which are periodic). By contrast, higher roots (cube roots, fourth roots, etc.) never appear to show any simple pattern like what e or \sqrt{e} or e^2 does. Unlike "regularity," they all seem to show "randomness" with Kusmin's rescaling [see (2.197)].

For example, among the first million partial quotients in the continued fraction for the cube root of 2 the digit 1 appears 414,983 times, which is remarkably close to the 41.5 % in (2.198), i.e., Kusmin's limit (2.197) with $k = 1$.

The same remarkable fact holds for the special number π : among the first million partial quotients the digit 1 appears 414,526 times, again very close to 41.5 %.

These are striking numerical facts, but, unfortunately, we cannot prove any theorem—not even the most plausible conjecture. For example, we don't know for sure whether the sequence a_1, a_2, a_3, \dots of partial quotients for the cube root of 2 is bounded or not. What is worse, we don't know a single algebraic number of degree ≥ 3 for which the sequence a_1, a_2, a_3, \dots of partial quotients is unbounded. We don't know this in spite of the well-known conjecture (raised by Khinchin in the 1930s) claiming that a_1, a_2, a_3, \dots is unbounded for *every* single real algebraic number of degree ≥ 3 .

Summarizing, we can safely say that computer experimentation strongly supports the Giant Leap phenomenon for both the decimal (or any other base) expansion and the continued fraction expansion of special numbers: they either exhibit very simple behavior or they exhibit full-blown randomness. The only technical difference is in the *scaling*: in continued fractions the ordinary uniform Lebesgue measure in the unit interval $(0,1)$ has to be replaced by the nonuniform Gauss measure (2.195) and (2.196).

In spite of the overwhelming numerical evidence, we don't have the slightest idea how to prove the Giant Leap phenomenon. A good illustration of what contemporary mathematics can do versus the conjectured truth is the concrete special number $x = \sqrt[3]{2}$ and a brief discussion of the celebrated works of two Fields medal winners, K.F. Roth and A. Baker. We begin with recalling a classical result of Dirichlet: for every irrational α there are infinitely many rationals p/q such that

$$\left| \alpha - \frac{p}{q} \right| < \frac{1}{q^2}. \quad (2.199)$$

In the 1950s K.F. Roth completed a long line of research initiated by Thue and Siegel and proved the following basic theorem in diophantine approximation (he was awarded a Fields medal in 1958): for any real algebraic number of degree ≥ 3 , including the case $\alpha = \sqrt[3]{2}$, and for any $\varepsilon > 0$,

$$\left| \alpha - \frac{p}{q} \right| > \frac{c(\alpha, \varepsilon)}{q^{2+\varepsilon}}, \quad (2.200)$$

where $c = c(\alpha, \varepsilon) > 0$ is a constant (note that the case of quadratic irrationals is trivial). In view of (2.199) Roth's inequality (2.200) is nearly best possible (since

$\varepsilon > 0$ can be arbitrarily small), but a more delicate analysis reveals that there is plenty of room for improvement in (2.200). Indeed, (2.200) is equivalent to

$$q \cdot \|q\alpha\| > \frac{c(\alpha, \varepsilon)}{q^\varepsilon} \quad (2.201)$$

for every integer $q \geq 1$, where $\|x\|$ denotes the distance of a real x from the nearest integer. On the other hand, for every real algebraic number of degree ≥ 3 , including $\alpha = \sqrt[3]{2}$, computer experimentation seems to support the much stronger inequality

$$q \cdot \|q\alpha\| > \frac{c(\alpha, \varepsilon)}{\log q \cdot (\log \log q)^{1+\varepsilon}} \quad (2.202)$$

for every integer $q \geq 3$, and also that (2.202) is best possible in the sense that we cannot delete $\varepsilon > 0$. Notice that there is an exponential gap between (2.201) and (2.202).

By the way, (2.202) is certainly true for almost every real α ; the proof is easy.

A serious handicap of Roth's theorem (or Thue–Siegel–Roth theorem) is that the constant $c = c(\alpha, \varepsilon) > 0$ is ineffective: we cannot replace it with an explicit constant. The reason is that the proof technique (“Thue method”) is indirect—it involves a hypothetical assumption that there is a large “bad” q , which behaves wickedly, and the constant $c = c(\alpha, \varepsilon) > 0$ depends on the size of this “bad” q (q is finite, but in principle it can be arbitrarily large). Nevertheless effective results have been obtained by A. Baker in the 1960s (for which he was awarded the Fields medal in 1970). For example, in 1964 Baker proved the explicit result

$$q \cdot \|q\sqrt[3]{2}\| > \frac{10^{-6}}{q^{0.955}} \quad (2.203)$$

that holds for every integer $q \geq 1$. The point here is the effective constant 10^{-6} in the numerator and the exponent $0.955 < 1$ in the denominator (notice that (2.203) with 1 instead of 0.955 is trivial, since $\sqrt[3]{2}$ is a cubic number).

We have to admit, therefore, that there is a humiliating exponential gap between the apparent truth [i.e., conjecture (2.202)] and what contemporary mathematics can do: the ineffective (2.201) and the effective (2.203), due to two Fields' medalists. (Nevertheless, even a “weak” result like (2.203) has remarkable consequences in the theory of diophantine equations.)

Conjecture (2.202) for real algebraic numbers (of degree ≥ 3)—a special case of the vague Giant Leap phenomenon—features “randomness.” **Where does this pseudorandomness come from?** This is a fundamental open problem, and we are nowhere near to understand it (not to mention answering it). For more about this exciting general issue, see Wolfram [Wo] and Beck [Be6]].

With some exaggeration we may even include the celebrated Riemann Hypothesis as another example of the Giant Leap. In the history of mathematics the set of primes served the first example of what one would call a “random set.” The Riemann Hypothesis (arguably the most famous open problem in mathematics) is equivalent to a problem about the “randomness” of the primes in the following way. The starting point is Riemann’s remarkable Explicit Formula for the prime-counting function $\pi(x) = \sum_{p \leq x} 1$, which involves the nontrivial *zeros* of the Riemann zeta function. Instead of the original formula, nowadays it is customary to discuss a simplified version, due to von Mangoldt, where the plain prime-counting function $\pi(x)$ is replaced with a weighted version (“Mangoldt sum”)

$$\psi_0(x) = \sum_{1 \leq n \leq x} \Lambda(n), \quad (2.204)$$

where $\Lambda(n) = \log p$ if n is a power of p (p always stands for a prime) and $\Lambda(n) = 0$ if n is not a prime power. Riemann’s Explicit Formula in prime number theory goes as follows:

$$\psi_0(x) = x - \sum_{\rho} \frac{x^{\rho}}{\rho} + O(1), \quad (2.205)$$

where ρ runs through the nontrivial zeta-zeros (meaning the zeros in the vertical strip with real part between 0 and 1). Riemann described the number of the nontrivial zeta-zeros (say) in the vertical box where the imaginary part has absolute value $\leq T$ (T is “large”): the number is

$$\frac{1}{2\pi} T \log T - \frac{1 + \log(2\pi)}{2\pi} T + O(\log T). \quad (2.206)$$

In sharp contrast to the *number*, we can prove very little about the *location* of the nontrivial zeta-zeros. What we can prove is much, much less than the Riemann Hypothesis, which claims that the nontrivial zeta-zeros are all on the critical line (vertical line with $\Re z = 1/2$; we cannot even prove the existence of any zero-free strip between $0 < \Re z < 1$). Applying the Riemann Hypothesis to (2.205), we obtain

$$\psi_0(x) = x + O(x^{1/2+o(1)}), \quad (2.207)$$

or equivalently (via integration by parts)

$$\pi(x) = \int_2^x \frac{dt}{\log t} + O(x^{1/2+o(1)}). \quad (2.208)$$

The square root size error term $O(x^{1/2+o(1)})$ nicely fits the so-called random set simulation of the primes. By the Prime Number Theorem, the density of the primes at x is $\frac{1}{\log x}$. This motivates the following simulation (due to Cramer): starting from

$n = 3$, for every integer $n \geq 3$ we toss a “loaded n -coin” that shows Heads with probability $\frac{1}{\log n}$ and shows Tails with probability $1 - \frac{1}{\log n}$. Keeping n if the outcome of the trial is Heads and rejecting it if the outcome is Tails, we obtain a Random Subset of the natural numbers; we call the elements of this random set “random primes.” The expected number of “random primes” is exactly

$$\sum_{n=3}^x \frac{1}{\log n} = \int_2^x \frac{dt}{\log t} + O(1), \quad (2.209)$$

and the actual number of “random primes” $\leq x$ fluctuates around the expected number (2.209) with the usual square root size standard deviation $O(x^{1/2+o(1)})$. In other words, formula (2.208), which is equivalent to the Riemann Hypothesis, is in perfect harmony with the $O(x^{1/2+o(1)})$ size fluctuation of the Random Subset (i.e., the Monte Carlo simulation of the primes).

The converse is also true: if the Riemann Hypothesis fails then the fluctuation in (2.205) is much larger than the standard deviation $O(x^{1/2+o(1)})$. Indeed, if there is a nontrivial zeta-zero $\rho = \beta + i\gamma$ with $\beta \neq 1/2$, then $\rho^* = (1-\beta) + i\gamma$ is another zeta-zero (follows from a symmetry of the Functional Equation of the zeta function), and $\max\{\beta, 1-\beta\} = \alpha > 1/2$. Then in (2.205) the fluctuation around x is at least as large as $x^{\alpha-o(1)}$, and also the fluctuation of $\pi(x)$ around the logarithmic integral is at least as large as $x^{\alpha-o(1)}$, which is asymptotically much larger than the standard deviation $O(x^{1/2+o(1)})$ of the Random Subset (it is not too difficult to make this argument precise). In other words, the failure of the Riemann Hypothesis implies that the “random prime” model is grossly incorrect.

Even if no one has a rigorous mathematical proof, everyone would agree that the Riemann Hypothesis is “true”—just like everyone would agree that π , e , $\sqrt{2}$ are all normal. Indeed, we have an overwhelming “computer science proof”: it cannot be an accident that the first billion zeta-zeros are all on the critical line. Since the Riemann Hypothesis is “true,” the Random Prime model predicts the fluctuations in the global distribution of primes very accurately.

The common feature of the digit sequences of special numbers and the set of primes is the “apparent randomness” and the (almost) total lack of rigorous proofs. Our main goal is to *prove* results, such as Theorems 1.1 and 1.2, which support the Giant Leap phenomenon. These results are admittedly modest first steps only. Our second goal is to challenge the reader to participate in the long-term research project of exploring this exciting mystery.

What we do here has some vague formal similarities to the Erdős–Kac theorem (about the number of prime divisors of typical integers) and other probabilistic results about multiplicative and additive number theoretic functions (see, e.g., Elliott’s book [El] or Kac [Ka]). However, in spite of the formal similarity, the two subjects are rather different.

2.6 Connection with Quadratic Fields (I)

After the philosophical detour of Sect. 2.5, now we return to the proofs of our central limit theorems (Theorems 1.1 and 1.2); in particular, to the computation of the *expectation* and the *variance*. In Sect. 2.4 we proved Proposition 2.16, which evaluates the mean value as follows:

$$M_\alpha(N) = -\frac{1}{2\pi} \sum_{n=1}^N \frac{1}{n \tan(\pi n \alpha)} + O(1), \quad (2.210)$$

assuming α is a badly approximable number. The following result is an alternative formula for $M_\alpha(N)$ in the special case when $\alpha = \sqrt{d}$, $d \equiv 3 \pmod{4}$ is a square-free positive integer. The necessary distinction between the cases $d \equiv 1$ or $3 \pmod{4}$ is one of the characteristic peculiarities of algebraic number theory—a subject that we are going to heavily use below.

Proposition 2.20. *Assume that d is a square-free positive integer with $d \equiv 3 \pmod{4}$, then*

$$M_{\sqrt{d}}(N) = \frac{\sqrt{d}}{\pi^2} \left(\sum_{\substack{(x,y) \neq (0,0): \\ \text{primary representations}}} \frac{1}{x^2 - dy^2} \right) \frac{\log N}{\log \eta_d} + O((\log \log N)^3), \quad (2.211)$$

where $\eta_d = u_0 + v_0\sqrt{d}$ comes from the least solution $x = u_0$, $y = v_0$ of Pell's equation $x^2 - dy^2 = 1$ (“least” means that $x_0 > 0$, $y_0 > 0$ and y_0 is least). The meaning of “primary representations” in (2.211) will be explained in the proof below.

Proof. First we give a precise definition of the infinite series

$$\sum_{\substack{(x,y) \neq (0,0): \\ \text{primary representations}}} \frac{1}{x^2 - dy^2} \quad (2.212)$$

in the middle of (2.211), and prove the convergence. Note that $x^2 - dy^2$ is the principal (binary quadratic) form of discriminant $4d$, and the theory of quadratic forms of discriminant $4d$ is equivalent to the theory of the real quadratic field $\mathbf{Q}(\sqrt{d})$. We assume that the reader is somewhat familiar with the simplest concepts and facts about quadratic forms and quadratic fields (see, for example the book [Za4]).

We recall the well-known fact that, given any integer $A \neq 0$, if the equation $x^2 - dy^2 = A$ has one integral solution (x, y) , then the equation has *infinitely* many integral solutions. Indeed, if $x_1^2 - dy_1^2 = A$ and $u^2 - dv^2 = 1$, then the product formula

$$(x_1 + y_1\sqrt{d})(u + v\sqrt{d}) = (x_1u + y_1vd) + (x_1v + y_1u)\sqrt{d} = x_2 + y_2\sqrt{d} \quad (2.213)$$

leads to a new solution $x_2 = x_1u + y_1vd$, $y_2 = x_1v + y_1u$ of the equation $x^2 - dy^2 = A$. Since Pell's equation $u^2 - dv^2 = 1$ has infinitely many solutions, generated by the least solution, product formula (2.213) gives rise to infinitely many solutions of $x^2 - dy^2 = A$. The two solutions, (x_1, y_1) and (x_2, y_2) , related by the product formula (2.213), are called *associates*—this defines an equivalence relation on the set of all solutions of $x^2 - dy^2 = A$. Let $R_d(A)$ denote the number of equivalence classes. Note that $R_d(A)$ is always finite and satisfies the inequality

$$R_d(A) \leq \tau(|A|), \quad (2.214)$$

where $\tau(n)$ is the divisor function, i.e., $\tau(n)$ is the number of (positive) divisors of n , including 1 and n itself. Inequality (2.214) is a classical result (it is in fact a corollary of an exact formula for $R_d(A)$, due to Dirichlet). Now we are ready to define the precise meaning of series (2.212):

$$\sum_{\substack{(x,y) \neq (0,0): \\ \text{primary representations}}} \frac{1}{x^2 - dy^2} = \sum_{A \neq 0} \frac{R_d(A)}{A} = \sum_{n=1}^{\infty} \frac{R_d(n) - R_d(-n)}{n}. \quad (2.215)$$

To prove the convergence in (2.215), we describe a definite way of selecting a representative solution from each equivalence class—we call these representatives the *primary solutions* of $x^2 - dy^2 = A$. First we take the conjugate of the product formula (2.213):

$$(x_1 - y_1\sqrt{d})(u - v\sqrt{d}) = x_2 - y_2\sqrt{d}, \quad (2.216)$$

and then take the ratio of (2.213) and (2.216):

$$\frac{x_1 + y_1\sqrt{d}}{x_1 - y_1\sqrt{d}} \cdot \frac{u + v\sqrt{d}}{u - v\sqrt{d}} = \frac{x_2 + y_2\sqrt{d}}{x_2 - y_2\sqrt{d}}. \quad (2.217)$$

We have $u + v\sqrt{d} = \pm\eta^m$ for some integer m (where $\eta = \eta_d$ is the fundamental unit), and so $u - v\sqrt{d} = \pm\eta^{-m}$. Returning to (2.217), we have

$$\frac{x_2 + y_2\sqrt{d}}{x_2 - y_2\sqrt{d}} = \frac{x_1 + y_1\sqrt{d}}{x_1 - y_1\sqrt{d}} \cdot \eta^{2m}. \quad (2.218)$$

In view of (2.218) there is just one choice of m (for a given x_1 and y_1) which will ensure that

$$1 < \frac{x_2 + y_2\sqrt{d}}{x_2 - y_2\sqrt{d}} \leq \eta_d^2. \quad (2.219)$$

Equation (2.219) does not change if we replace (x_2, y_2) with $(-x_2, -y_2)$, so we can further ensure that

$$x_2 - y_2\sqrt{d} > 0. \quad (2.220)$$

The particular solution $x = x_2, y = y_2$ of $x^2 - dy^2 = A$ that satisfies (2.219) and (2.220) will be called *primary*.

To prove the convergence in (2.215), we estimate the sums

$$\sum_{n=1}^N R_d(n) \quad \text{and} \quad \sum_{n=1}^N R_d(-n)$$

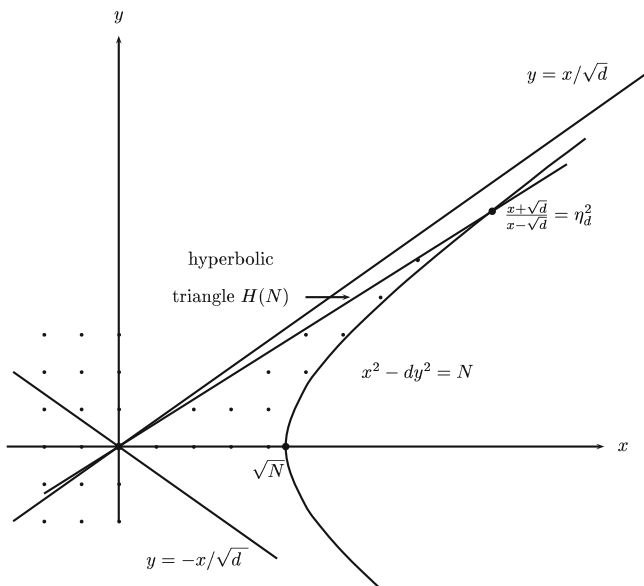
by employing a simple lattice point counting argument. (It is worthwhile to point out that the same lattice point counting argument is used in the proof of Dirichlet's class number formula for real quadratic fields $h(d) \log \eta_d = \sqrt{d} L(1, \chi_d)$.) We will show that

$$\sum_{n=1}^N R_d(n) = c_0(d)N + O(\sqrt{N}) \quad (2.221')$$

and

$$\sum_{n=1}^N R_d(-n) = c_0(d)N + O(\sqrt{N}) \quad (2.221'')$$

with the *same* constant factor $c_0(d)$ (which is of course independent of N).



To prove (2.221), we use (2.219) and (2.220), which tells us that the sum $\sum_{n=1}^N R_d(n)$ equals the number of lattice points $(x, y) \in \mathbb{Z}^2$ satisfying the three requirements:

$$0 < x^2 - dy^2 \leq N, \quad x - y\sqrt{d} > 0, \quad 1 < \frac{x + y\sqrt{d}}{x - y\sqrt{d}} \leq \eta_d^2. \quad (2.222)$$

The region defined by Eq. (2.222) is a sector of a hyperbola bounded by two half lines through the origin—we call it a “hyperbolic triangle,” and denote it with $H(N) = H_d(N)$; see the picture. The left corner of the “hyperbolic triangle” $H(N) = H_d(N)$ is the origin $(0, 0)$, the lower right corner is the point $(\sqrt{N}, 0)$, and the upper right corner is the intersection of the hyperbola $x^2 - dy^2 = N$ and the positive side of the line

$$\frac{x + y\sqrt{d}}{x - y\sqrt{d}} = \eta_d^2.$$

It is not too difficult to determine the area of $H(N)$: we have

$$\text{Area}(H_d(N)) = \frac{N}{2\sqrt{d}} \log \eta_d. \quad (2.223)$$

We outline the proof of (2.223). First we change the coordinates from x, y to u, v where $u = x - y\sqrt{d}$ and $v = x + y\sqrt{d}$ and compute the determinant

$$\frac{\partial(u, v)}{\partial(x, y)} = \begin{vmatrix} 1 & -\sqrt{d} \\ 1 & \sqrt{d} \end{vmatrix} = 2\sqrt{d}. \quad (2.224)$$

In the u, v -plane, the hyperbolic triangle $H(N)$ [defined by (2.222)] is given by

$$0 < uv \leq N, \quad u > 0, \quad u < v \leq u\eta^2.$$

These conditions are equivalent to

$$0 < u < \sqrt{N}, \quad u < v \leq \min\{u\eta^2, N/u\}. \quad (2.225)$$

Since $u\eta^2 < N/u$ is equivalent to $u < \sqrt{N}/\eta$, the area of (2.225) is

$$\int_0^{\sqrt{N}/\eta} (u\eta^2 - u) du + \int_{\sqrt{N}/\eta}^{\sqrt{N}} \left(\frac{N}{u} - u \right) du = N \log \eta.$$

This has to be divided by the determinant in (2.224) to obtain the area in the x, y -plane and this gives (2.223).

To estimate the number of lattice points inside the hyperbolic triangle $H(N)$, we use the general inequality (see Proposition 1.9)

$$\text{Area}(H) - O(\text{Perimeter}(H)) \leq |H \cap \mathbb{Z}^2| \leq \text{Area}(H) + O(\text{Perimeter}(H)) + 1. \quad (2.226)$$

The perimeter of the hyperbolic triangle $H(N)$ is $O(\sqrt{N})$. Indeed, the three vertices of $H(N)$ are $(0, 0)$, $(\sqrt{N}, 0)$, and (x_0, y_0) , where the point (x_0, y_0) satisfies both equations

$$x^2 - dy^2 = N, \quad \frac{x + y\sqrt{d}}{x - y\sqrt{d}} = \eta_d^2. \quad (2.227)$$

It follows from (2.227) that $x_0 + y_0\sqrt{d} = \sqrt{N}\eta_d$. The coordinates of the vertices of $H(N)$ are all in the range $O(\sqrt{N})$, implying that the perimeter of $H(N)$ is $O(\sqrt{N})$.

Applying (2.226) we have

$$\begin{aligned} \sum_{n=1}^N R_d(n) &= \text{Area}(H(N)) + O(\text{Perimeter}(H(N))) = \\ &= \frac{N}{2\sqrt{d}} \log \eta_d + O(\sqrt{N}). \end{aligned} \quad (2.228)$$

Repeating the same argument for $0 < dy^2 - x^2 \leq N$ instead of $0 < x^2 - dy^2 \leq N$, we obtain the same right-hand side:

$$\sum_{n=1}^N R_d(-n) = \frac{N}{2\sqrt{d}} \log \eta_d + O(\sqrt{N}), \quad (2.229)$$

proving (2.221') and (2.221'').

Taking the difference of (2.228) and (2.229), we have

$$\sum_{n=1}^N (R_d(n) - R_d(-n)) = O(\sqrt{N}). \quad (2.230)$$

Now it is easy to prove the convergence of the series in (2.215). Indeed, by using (2.230) and Abel's transformation (2.119), we have for any $1 < N < M$,

$$\begin{aligned} \sum_{n=N}^M \frac{R_d(n) - R_d(-n)}{n} &= \sum_{m=N}^{M-1} \frac{\sum_{n=N}^m (R_d(n) - R_d(-n))}{m(m+1)} \\ &+ \frac{1}{M} \sum_{n=N}^M (R_d(n) - R_d(-n)) = \sum_{m=N}^{M-1} \frac{O(\sqrt{m})}{m^2} + \frac{O(\sqrt{M})}{M} = O(N^{-1/2}). \end{aligned} \quad (2.231)$$

Equation (2.231) immediately implies the convergence of the infinite series in (2.215):

$$\sum_{\substack{(x,y) \neq (0,0): \\ \text{primary representations}}} \frac{1}{x^2 - dy^2} = \sum_{n=1}^{\infty} \frac{R_d(n) - R_d(-n)}{n} \text{ is convergent.} \quad (2.232)$$

If $x = w \geq 0$, $y = z \geq 0$ is a primary solution of $x^2 - dy^2 = A$ with $A > 0$, then by definition

$$A = w^2 - dz^2 = (w + z\sqrt{d})(w - z\sqrt{d}), \quad 1 < \frac{w + z\sqrt{d}}{w - z\sqrt{d}} \leq \eta_d^2,$$

implying

$$\sqrt{A} < w + z\sqrt{d} \leq \sqrt{A}\eta_d. \quad (2.233)$$

It follows from the product formula (2.213) that for every integer j , $(w + z\sqrt{d})\eta_d^j$ gives another solution of $x^2 - dy^2 = A$, and by (2.233) we have

$$(w + z\sqrt{d})\eta_d^j \leq (2 + o(1))N\sqrt{d} \iff j \leq \frac{\log(N/\sqrt{A})}{\log \eta_d} + O(1). \quad (2.234)$$

The same holds for $x^2 - dy^2 = A$ with $A < 0$, the only minor difference is that in (2.234) we have to replace \sqrt{A} with $\sqrt{|A|}$.

Thus by (2.234) we obtain the key formula:

$$\sum_{\substack{1 \leq y \leq N, 1 \leq x \leq N\sqrt{d}: \\ |x^2 - dy^2| \leq m}} \frac{1}{x^2 - dy^2} = \sum_{1 \leq A \leq m} \frac{R_d(A) - R_d(-A)}{A} \cdot \left(\frac{\log(N/\sqrt{A})}{\log \eta_d} + O(1) \right), \quad (2.235)$$

which holds for any $1 < m < N$. Equation (2.235) is the key to prove Proposition 2.20; in the application below we will use (2.235) with the choice $m \approx (\log N)^c$, where $c > 1$ is an absolute constant to be specified later.

We divide the left-hand side of (2.235) into two parts:

$$\sum_{\substack{1 \leq y \leq N, 1 \leq x \leq N\sqrt{d}: \\ |x^2 - dy^2| \leq m}} \frac{1}{x^2 - dy^2} = \sum_1 + \sum_2, \quad (2.236)$$

where

$$\sum_1 = \sum_{\substack{1 \leq y \leq N, 1 \leq x \leq N\sqrt{d}: \\ |x^2 - dy^2| \leq m, |x - y\sqrt{d}| < 1/2}} \frac{1}{x^2 - dy^2} \quad (2.237)$$

and

$$\sum_2 = \sum_{\substack{1 \leq y \leq N, 1 \leq x \leq N\sqrt{d}: \\ |x^2 - dy^2| \leq m, |x - y\sqrt{d}| \geq 1/2}} \frac{1}{x^2 - dy^2}. \quad (2.238)$$

First we show that

$$\sum_2 = O((\log m)^3). \quad (2.239)$$

To prove (2.239), notice that the conditions

$$|x^2 - dy^2| \leq m, \quad |x - y\sqrt{d}| \geq 1/2$$

in (2.238) clearly imply

$$0 < x + y\sqrt{d} \leq 2m. \quad (2.240)$$

Since the number of solutions of $x^2 - dy^2 = A$ with $x \geq 0, y \geq 0, x + y\sqrt{d} \leq 2m$ is estimated from above by $R_d(A) \cdot O(\log m)$, by (2.238) and (2.240) we have the following trivial upper bound on \sum_2 :

$$\sum_2 = O\left(\log m \sum_{A=1}^m \frac{R_d(A) - R_d(-A)}{A}\right). \quad (2.241)$$

We recall (2.214): $R_d(A) + R_d(-A) \leq \tau(A)$ where $\tau(n)$ is the divisor function (number of divisors of n) and using this in (2.241) we obtain

$$\sum_2 = O\left(\log m \sum_{k=1}^m \frac{\tau(k)}{k}\right). \quad (2.242)$$

We recall the following well-known fact about the divisor function (see, e.g., in [Ha-Wr]):

$$\sum_{k=1}^n \tau(k) = O(n \log n). \quad (2.243)$$

An application of (2.243) in formula (2.242), combined with the Abel's transformation (2.119), gives (2.239).

Next we study \sum_1 defined in (2.237). We are motivated by the vague approximation

$$\tan(\pi k \sqrt{d}) \approx \pi(k \sqrt{d} - \ell) = \pi(k \sqrt{d} - \ell) \frac{k \sqrt{d} + \ell}{k \sqrt{d} + \ell} \approx \frac{\pi}{2k \sqrt{d}} (dk^2 - \ell^2), \quad (2.244)$$

where $\ell = \ell(k, d)$ is the nearest integer to $k \sqrt{d}$. It is easy to make (2.244) precise by using the beginning of the Taylor series of $\tan(x)$: $\tan(x) = x + O(x^3)$; then a simple calculation gives the following precise equality:

$$\frac{1}{k \tan(\pi k \sqrt{d})} - \frac{2\sqrt{d}}{\pi(dk^2 - \ell^2)} = O(\|k \sqrt{d}\|/k) + O(1/k^2). \quad (2.245)$$

Thus we have [see (2.237) and (2.245)]

$$\begin{aligned} - \sum_{\substack{1 \leq k \leq N: \\ 2\sqrt{d}k \|k \sqrt{d}\| \leq m}} \frac{1}{k \tan(\pi k \sqrt{d})} &= \sum_{\substack{1 \leq k \leq N, 1 \leq \ell \leq N\sqrt{d}: \\ |\ell^2 - dk^2| \leq m, |\ell - k\sqrt{d}| < 1/2}} \frac{2\sqrt{d}}{\pi(\ell^2 - dk^2)} + \\ &+ O\left(\sum_{k \geq 1} k^{-2}\right) + O\left(\sum_{\substack{1 \leq k \leq N: \\ 2\sqrt{d}k \|k \sqrt{d}\| \leq m}} \|k \sqrt{d}\|/k\right) = \\ &= \frac{2\sqrt{d}}{\pi} \sum_1 + O(1) + O\left(\sum_{\substack{1 \leq k \leq N: \\ 2\sqrt{d}k \|k \sqrt{d}\| \leq m}} \|k \sqrt{d}\|/k\right), \end{aligned} \quad (2.246)$$

provided

$$m \text{ is a half-integer, i.e., } m = \text{integer} + \frac{1}{2}. \quad (2.247)$$

To explain the role of “half-integer m ” [see condition (2.247)] in (2.246), note that $|dk^2 - \ell^2| = (k \sqrt{d} + \ell)|k \sqrt{d} - \ell|$ is clearly an integer, and

$$\begin{aligned} 2\sqrt{d}k \|k \sqrt{d}\| &= 2\sqrt{d}k |k \sqrt{d} - \ell| = ((\sqrt{d}k + \ell) + (\sqrt{d}k - \ell))|k \sqrt{d} - \ell| = \\ &= |dk^2 - \ell^2| \pm (\sqrt{d}k - \ell)^2 = \text{integer} \pm (\sqrt{d}k - \ell)^2. \end{aligned} \quad (2.248)$$

Since ℓ is the nearest integer to $\sqrt{d}k$, $(\sqrt{d}k - \ell)^2 \leq 1/4$, and so by (2.247) and (2.248) with $m = m_1 + 1/2$, where m_1 is an integer, we have

$$2\sqrt{d}k \|k\sqrt{d}\| \leq m = m_1 + \frac{1}{2} \iff |dk^2 - \ell^2| \leq m_1. \quad (2.249)$$

It is easy to estimate the error term in (2.246):

$$\sum_{\substack{1 \leq k \leq N: \\ 2\sqrt{d}k \|k\sqrt{d}\| \leq m}} \|k\sqrt{d}\|/k \leq \sum_{1 \leq k \leq m} 1/k + \sum_{m < k \leq N} m/k^2 = O(\log m). \quad (2.250)$$

Next we apply the following general result, which holds for any badly approximable α (we will choose $\alpha = \sqrt{d}$).

Lemma 2.21. *If α is badly approximable, then for any $N \geq 2$ and $\mu \geq (\log N)^6$,*

$$M_\alpha(N) = -\frac{1}{2\pi} \sum_{\substack{1 \leq n \leq N: \\ n\|\alpha\| \leq \mu}} \frac{1}{n \tan(\pi n\alpha)} + O(1).$$

Here the error term $O(1)$ depends only on the upper bound on the partial quotients of the badly approximable α .

First we show how to use Lemma 2.21 to complete the proof of Proposition 2.20. We make the choice $\mu = (\log N)^6 + O(1)$; here I choose the constant $O(1)$ in such a way that

$$2\sqrt{d}\mu = m = \text{integer} + \frac{1}{2}. \quad (2.251)$$

Combining (2.246)–(2.251) with Lemma 2.21—where $\alpha = \sqrt{d}$ —we obtain

$$M_{\sqrt{d}}(N) = \frac{\sqrt{d}}{\pi^2} \sum_1 + O(\log \log N). \quad (2.252)$$

By (2.235)–(2.239) and (2.252),

$$M_{\sqrt{d}}(N) = \frac{\sqrt{d}}{\pi^2} \sum_{1 \leq A \leq m} \frac{R_d(A) - R_d(-A)}{A} \cdot \left(\frac{\log(N/\sqrt{A})}{\log \eta_d} + O(1) \right) + O((\log \log N)^3), \quad (2.253)$$

where by condition (2.251),

$$m = 2\sqrt{d}\mu = 2\sqrt{d}((\log N)^6 + O(1)) = \text{half-integer}. \quad (2.254)$$

Note that

$$\begin{aligned} & \sum_{1 \leq A \leq m} \frac{R_d(A) - R_d(-A)}{A} \cdot \left(\frac{\log(N/\sqrt{A})}{\log \eta_d} + O(1) \right) = \\ &= \frac{\log N}{\log \eta_d} \sum_{1 \leq A \leq m} \frac{R_d(A) - R_d(-A)}{A} + O \left(\sum_{1 \leq A \leq m} \frac{(R_d(A) - R_d(-A)) \log A}{A} \right). \end{aligned} \quad (2.255)$$

Again using (2.214), (2.243), and Abel's transformation (2.119), a routine calculation gives

$$\sum_{1 \leq A \leq m} \frac{(R_d(A) - R_d(-A)) \log A}{A} = O((\log m)^3). \quad (2.256)$$

Moreover, by (2.231) and (2.254),

$$\sum_{A > m} \frac{R_d(A) - R_d(-A)}{A} = O(m^{-1/2}) = O((\log N)^{-3}). \quad (2.257)$$

Combining (2.255)–(2.257), we have

$$\begin{aligned} & \sum_{1 \leq A \leq m} \frac{R_d(A) - R_d(-A)}{A} \cdot \left(\frac{\log(N/\sqrt{A})}{\log \eta_d} + O(1) \right) = \\ &= \frac{\log N}{\log \eta_d} \sum_{A=1}^{\infty} \frac{R_d(A) - R_d(-A)}{A} + O((\log \log N)^3). \end{aligned} \quad (2.258)$$

Finally, (2.253) and (2.258) imply Proposition 2.20.

It remains to give a

Proof of Lemma 2.21. We basically repeat the argument of Step One in the proof of Proposition 2.16 (see Sect. 2.4). This means, we are going to combine Koksma's inequality [in fact, we use the form (2.148) and (2.149)] with Lemma 2.19 and try to force the usual cancellation of the “positive and negative sides.” Since the notation “ $\|x\|=\text{small}$ ” does not tell us whether x is slightly less or slightly more than an integer, we will use the notation $\|x\|^+$ and $\|x\|^-$ introduced in Sect. 2.4, see the definition between (2.153) and (2.154). Let

$$A^+(M, p, q, r) = \left\{ M(1 - \frac{1}{r}) < k \leq M : \frac{p}{M} \leq \|k\alpha\|^+ < \frac{p}{M}(1 + \frac{1}{q}) \right\}$$

and

$$A^-(M, p, q, r) = \left\{ M(1 - \frac{1}{r}) < k \leq M : \frac{p}{M} \leq \|k\alpha\|^- < \frac{p}{M}(1 + \frac{1}{q}) \right\},$$

where $M \geq 2p$, $p \geq 2$, $q \geq 1$, $r \geq 1$ are real numbers (to be specified later).

We apply Lemma 2.18—in fact, we use Eq. (2.148)—with

$$a = \frac{p}{M}, \quad b = \frac{p}{N}(1 + \frac{1}{q}), \quad f(x) = \frac{1}{\tan(\pi x)},$$

and the finite point set in the interval $[a, b]$ is

$$\mathcal{X} = \{k\alpha \pmod{1} : k \in A^+(M, p, q, r)\};$$

then we have

$$\left| \sum_{k \in A^+(M, p, q, r)} \frac{1}{\tan(\pi k\alpha)} - \frac{|A^+(M, p, q, r)|}{b-a} \int_a^b f(x) dx \right| \leq \Delta \int_a^b |f'(x)| dx,$$

where by Lemma 2.19, $\Delta = O(\log p)$. Also, we have

$$\int_a^b |f'(x)| dx = |f(b) - f(a)| = O\left(\frac{M}{p}(1 + \frac{1}{q}) - \frac{M}{p}\right) = O\left(\frac{M}{pq}\right),$$

and again using Lemma 2.19—in fact, we use it twice: first for $n = M$, then for $n = M(1 - 1/r)$, and finally, take the difference—we can estimate the number of elements $|A^+(M, p, q, r)|$ of the set $A^+(M, p, q, r)$ as follows:

$$\begin{aligned} |A^+(M, p, q, r)| &= \frac{M}{r}(b-a) + O(\log(M(b-a) + 2)) = \\ &= \frac{M}{r}(b-a) + O(\log((p/q) + 2)). \end{aligned}$$

It follows that

$$\begin{aligned} &\sum_{k \in A^+(M, p, q, r)} \frac{1}{\tan(\pi k\alpha)} = \\ &= \frac{M}{r} \int_a^b f(x) dx + O(M \log p / pq) + O(\log((p/q) + 2)) \frac{1}{b-a} \int_a^b f(x) dx = \\ &= \frac{M}{r} \int_a^b f(x) dx + O(M(\log p) / pq) + O(M(\log((p/q) + 2)) / p). \end{aligned}$$

If $k_1, k_2 \in A^+(M, p, q, r)$ then $k_1/k_2 = 1 + O(1/r)$, and so we have

$$\begin{aligned} \sum_{k \in A^+(M, p, q, r)} \frac{1}{k \tan(\pi k \alpha)} &= \frac{1}{r} \int_a^b f(x) dx + \\ &+ O\left(\frac{1}{pq} \log p\right) + O\left(\frac{1}{p} \log((p/q) + 2)\right) + O(r^{-2}) \int_a^b f(x) dx. \end{aligned} \quad (2.259)$$

Note that

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^b \frac{dx}{\tan(\pi x)} \leq \\ &\leq \log(b/a) = \log(1 + 1/q) = O(1/q). \end{aligned} \quad (2.260)$$

Since we can repeat the argument for $A^-(M, p, q, r)$, by (2.259) and (2.260) we have for both $A^\pm(M, p, q, r)$

$$\begin{aligned} \sum_{k \in A^\delta(M, p, q, r)} \frac{1}{k |\tan(\pi k \alpha)|} &= \frac{1}{r} \int_a^b f(x) dx + \\ &+ O\left(\frac{1}{pq} \log p\right) + O\left(\frac{1}{p} \log((p/q) + 2)\right) + O(r^{-2} q^{-1}) \end{aligned} \quad (2.261)$$

holds for both “ $\delta = +$ ” and “ $\delta = -$ ”.

Applying (2.261) with $p_j = p(1 + 1/q)^j$, $j = 0, 1, 2, \dots$, we have for both $\|x\|^+$ and $\|x\|^-$, i.e., formally for both “ $\delta = +$ ” and “ $\delta = -$ ” (note that the value of parameter $q \geq 1$ will be specified later):

$$\begin{aligned} \sum_{\substack{M(1-1/r) < k \leq M: \\ \|k\alpha\|^\delta \geq p/M}} \frac{1}{k |\tan(\pi k \alpha)|} &= \frac{1}{r} \int_a^{1/2} f(x) dx + \\ &+ O\left(\frac{1}{pq} \log p\right) O(q \log M) + O(q \log M) O\left(\frac{1}{p} \log((p/q) + 2)\right) + O(q \log M) O(r^{-2} q^{-1}), \end{aligned} \quad (2.262)$$

since we can clearly stop at $j = O(q \log M)$.

What we *really* want to estimate is a slightly different variant of (2.262), where the condition $\|k\alpha\|^\delta \geq p/M$ is replaced by $k \|k\alpha\|^\delta \geq p$:

$$\sum_{\substack{M(1-1/r) < k \leq M: \\ k \|k\alpha\|^\delta \geq p}} \frac{1}{k |\tan(\pi k \alpha)|}. \quad (2.263)$$

Since $M(1 - 1/r) < k \leq M$, $\|k\alpha\|^\delta \geq p/M$ in (2.262) implies $k\|k\alpha\|^\delta \geq (1 - 1/r)p$. By changing p to $p' = (1 - 1/r)p$, $a = p/M$ changes to $a' = (1 - 1/r)p/M$, and this gives the additional error term

$$\frac{1}{r} \int_{a'}^a f(x) dx = \frac{1}{r} \int_{(1-1/r)p/M}^{p/M} \frac{dx}{\tan(\pi x)} \approx \frac{1}{r} \cdot \frac{p}{Mr} \cdot \frac{M}{p} = O(r^{-2}). \quad (2.264)$$

Thus, by using (2.262) and (2.264) in (2.263), we have (“ $\delta = +$ ” and “ $\delta = -$ ”)

$$\begin{aligned} & \sum_{\substack{M(1-1/r) < k \leq M: \\ k\|k\alpha\|^\delta \geq p}} \frac{1}{k|\tan(\pi k\alpha)|} = \frac{1}{r} \int_a^{1/2} f(x) dx + \\ & + O\left(\frac{\log M \cdot \log p}{p}\right) + O\left(\frac{q}{p} \log M \cdot \log((p/q) + 2)\right) + O(r^{-2} \cdot \log M) + o(r^{-2}). \end{aligned} \quad (2.265)$$

In (2.265) we take the difference for “ $\delta = +$ ” and “ $\delta = -$ ”:

$$\begin{aligned} & \sum_{\substack{M(1-1/r) < k \leq M: \\ k\|k\alpha\| \geq p}} \frac{1}{k \tan(\pi k\alpha)} = \\ & = O\left(\frac{\log M \cdot \log p}{p}\right) + O\left(\frac{q}{p} \log M \cdot \log((p/q) + 2)\right) + O(r^{-2} \cdot \log M) + O(r^{-2}). \end{aligned} \quad (2.266)$$

Next we choose $r = (\log M)^3$ and apply (2.266) with $M_j = M(1 - 1/r)^j$, $j = 0, 1, 2, \dots, r - 1$. Since $(1 - 1/r)^r = e^{-1} + o(1)$, (2.266) implies that for every M there is a constant times smaller $M^* = (1 + o(1))M/e$ such that

$$\begin{aligned} & \sum_{\substack{M^* < k \leq M: \\ k\|k\alpha\| \geq p}} \frac{1}{k \tan(\pi k\alpha)} = \\ & = O\left(\frac{(\log M)^4 \cdot \log p}{p}\right) + O\left(\frac{q}{p} (\log M)^4 \cdot \log((p/q) + 2)\right) + O((\log M)^{-2}). \end{aligned} \quad (2.267)$$

We use (2.267) repeatedly: with $M = N$, $M = (1 + o(1))Ne^{-1}$, $M = (1 + o(1))Ne^{-2}$, $M = (1 + o(1))Ne^{-3}$, and so on—at the end we obtain

$$\begin{aligned} & \sum_{\substack{1 \leq k \leq N: \\ k\|k\alpha\| \geq p}} \frac{1}{k \tan(\pi k\alpha)} = \\ & = O\left(\frac{(\log N)^5 \cdot \log p}{p}\right) + O\left(\frac{q}{p} (\log N)^5 \cdot \log((p/q) + 2)\right) + O((\log N)^{-1}). \end{aligned} \quad (2.268)$$

By choosing $q = 1$ and $p \geq (\log N)^6$ in (2.268), we conclude that

$$\sum_{\substack{1 \leq k \leq N: \\ k \|k\alpha\| \geq p}} \frac{1}{k \tan(\pi k \alpha)} = o(1).$$

Combining this with Proposition 2.16, Lemma 2.21 follows. \square

This completes the proof of Proposition 2.20. \square

2.6.1 A Detour: Another Class Number Formula

We recall that Proposition 2.20 is exactly Eq. (2.14) in Sect. 2.1, and it quickly leads to a proof of the elegant Hirzebruch–Meyer–Zagier class number formula (HMZ-formula, in short) as follows. Assume that $d = p \equiv 3 \pmod{4}$ is a prime > 3 , and the class number of the real quadratic field $\mathbf{Q}(\sqrt{d})$ is one, or using the traditional h -notation, $h(d) = h(p) = 1$. Then we have the equality

$$\sum_{\substack{(x,y) \neq (0,0): \\ \text{primary representations}}} \frac{1}{x^2 - py^2} = L(1, \chi^*), \quad (2.269)$$

where χ^* is the so-called norm-sign character and $L(1, \chi^*)$ is the corresponding L-function at $s = 1$.

More precisely, χ^* is a unique character with values ± 1 defined for all ideals in the ring of the algebraic integers of $\mathbf{Q}(\sqrt{d})$ (in fact, χ^* depends only on the narrow ideal class) and satisfies $\chi^*((a)) = \text{sign } \text{Norm}(a)$ for the principal ideals (a) . Notice that, in our special case $d = p$ with $h(p) = 1$, every ideal is principal.

The special L-function

$$L(s, \chi^*) = \sum_{A: \text{ideals}} \frac{\chi^*(A)}{\text{Norm}(A)^s}$$

has the product decomposition

$$L(s, \chi^*) = L(s, \chi_{-4}) L(s, \chi_{-p}) \quad (2.270)$$

where

$$L(s, \chi_{-4}) = \sum_{n=1}^{\infty} \frac{\chi_{-4}(n)}{n^s} \quad \text{and} \quad L(s, \chi_{-p}) = \sum_{n=1}^{\infty} \frac{\chi_{-p}(n)}{n^s}$$

are the (ordinary) L-functions of the complex quadratic fields $\mathbf{Q}(\sqrt{-4}) = \mathbf{Q}(\sqrt{-1})$ (“Gauss integers”) and $\mathbf{Q}(\sqrt{-p})$; the characters χ_{-4} and χ_{-p} are defined as follows: $\chi_{-4}(n) = \pm 1$ if $n \equiv \pm 1 \pmod{4}$ and $\chi_{-4}(n) = 0$ if n is even, and

$$\chi_{-p}(n) = \left(\frac{n}{p} \right)$$

is the usual Legendre symbol (i.e., the quadratic residue symbol). Note that (2.270) is “explained” by the elementary factorization $4p = (-4)(-p)$ of the discriminant of $x^2 - py^2$; for a precise proof, see, e.g., Zagier’s book [Za4].

In the special case $s = 1$ Eq. (2.270) gives

$$L(1, \chi^*) = L(1, \chi_{-4})L(1, \chi_{-p}), \quad (2.271)$$

and by Dirichlet’s class number formula,

$$L(1, \chi_{-4}) = \frac{\pi}{4} \quad \text{and} \quad L(1, \chi_{-p}) = \frac{\pi h(-p)}{\sqrt{p}}, \quad (2.272)$$

if $p > 3$.

Let a_1, a_2, \dots, a_{2s} be the period of the continued fraction for \sqrt{p} (since $p \equiv 3 \pmod{4}$ prime, the length of the period has to be even). (We have to exclude $p = 3$, because $\mathbf{Q}(\sqrt{-3})$ has too many automorphisms—a technical nuisance in algebraic number theory.) By Proposition 2.1,

$$\begin{aligned} M_{\sqrt{p}}(N) &= \frac{-a_1 + a_2 - a_3 \pm \dots + (-1)^\ell a_\ell}{12} + O(1) = \\ &= \frac{-a_1 + a_2 \mp \dots + a_{2s}}{12} \cdot \frac{\log N}{\log \eta} + O(1), \end{aligned} \quad (2.273)$$

where ℓ is the last index for which $q_\ell \leq N$ and η is the fundamental unit of $\mathbf{Q}(\sqrt{p})$ (in the last equation we heavily used the periodicity of the continued fraction of \sqrt{p}).

On the other hand, combining Proposition 2.20 with (2.269)–(2.273), we have

$$M_{\sqrt{p}}(N) = \frac{h(-p)}{4} \cdot \frac{\log N}{\log \eta} + O((\log \log N)^3). \quad (2.274)$$

Comparing (2.273) and (2.274), we obtain the beautiful equation

$$h(-p) = \frac{-a_1 + a_2 - a_3 \pm \dots + a_{2s}}{3}. \quad (2.275)$$

As far as we know this equation was discovered (or rediscovered) in the 1970s by Hirzebruch, and it is called the Hirzebruch or Hirzebruch–Meyer–Zagier class number formula.

Note that, among the primes $p \equiv 3 \pmod{4}$, the majority (in fact, about 80 %) seems to satisfy the requirement $h(p) = 1$ (i.e., the real quadratic field $\mathbf{Q}(\sqrt{p})$ has class number one)—at least this is what we can read out from the numerical tables. Unfortunately, despite the overwhelming computational evidence, nothing is proved here.

It is more than surprising that the “mean value” $M_{\sqrt{p}}(N)$, associated with the irrational rotation $k\sqrt{p} \pmod{1}$, $k = 1, 2, \dots, N$, is intimately bound up with the class number $h(-p)$ of the complex quadratic field $\mathbf{Q}(\sqrt{-p})$. This leads to the following question.

2.6.2 How to Compute the Class Number in General: The Complex Case

One way to do it is to use Dirichlet’s finite class number formula, which expresses the class number in terms of the Dirichlet character of the corresponding discriminant. The formula is the simplest when $-d = -p$, where $p \equiv 3 \pmod{4}$. We form the sum, say R , of all quadratic residues \pmod{p} , and the sum, say N , of all quadratic non-residues. Then $h(-p) = (N - R)/p$. For example, if $p = 7$, the quadratic residues are $1^2, 2^2$, and $3^2 \equiv 2 \pmod{7}$, and the quadratic non-residues are the remaining $3, 5, 6 \pmod{7}$. The formula gives

$$h(-7) = \frac{(3 + 5 + 6) - (1 + 4 + 2)}{7} = \frac{14 - 7}{7} = 1.$$

In the general case, the formula is the following: if $K = \mathbf{Q}(\sqrt{-d})$ is a complex quadratic field, then

$$h(-d) = -\frac{w(-d)}{2D} \sum_{k=1}^D \chi_{-D}(k)k,$$

where $-D(=-d \text{ or } -4d)$ is the discriminant of K , $\chi_{-D}(k)$ is the real character of K periodic modulo D (it is a product of certain Legendre symbols), and finally $w(-1) = 4$, $w(-3) = 6$, $w(-d) = 2$ for the rest (the number of roots of unity in the field). An equivalent form is

$$h(-d) = \frac{1}{2 - \chi_{-D}(2)} \sum_{0 < k < D/2} \chi_{-D}(k)$$

for all square-free $d \geq 2$.

An alternative—in fact, more efficient—way to compute the class number is to use “reduction theory.” There is an elegant reduction theory for positive definite quadratic forms (i.e., when the discriminant is negative; we denote it $(-D)$), which leads to a surprisingly simple algorithm to determine the class number $h(-D)$ of a complex quadratic field $\mathbf{Q}(\sqrt{-D})$. We summarize it in a nutshell. By using a finite sequence of simple unimodular substitutions of the form $x = y'$, $y = -x'$ and $x = x' \pm y'$, $y = y'$, any binary form can be transformed into another binary form $ax^2 + bxy + cy^2$, for which $|b| \leq a \leq c$. In fact, we can even force that either

$$-a < b \leq a < c \quad \text{or} \quad 0 \leq b \leq a = c.$$

Such a form is called a *reduced* form. It is an important theorem that there is one and only one reduced form equivalent to any given form. The number of reduced forms with discriminant $-D$ is the class number $h(-D)$.

For example, to calculate the class number when $-D = -7$, the inequality $b^2 \leq a^2 \leq ac$ and the fact $4ac - b^2 = D$ give $3b^2 \leq D$, i.e., $|b| \leq \sqrt{D/3} = \sqrt{7/3} < 2$. Since $4ac - b^2 = D = 7$ implies that b is odd, we have $b = \pm 1$. Now $4ac = 1 + 7 = 8$ gives $a = 1, c = 2$. The requirement $-a < b \leq a < c$ excludes the case $b = -1$, so there is only one reduced form of discriminant -7 —namely, $x^2 + xy + 2y^2$ —yielding $h(-7) = 1$.

A more complicated example is $-D = -23$. The inequality $|b| \leq \sqrt{D/3} = \sqrt{23/3} < 2$ and the fact $4ac = b^2 + 23$ imply that b is odd and $b = \pm 1$. Now $4ac = 1 + 23 = 24$ gives $a = 1, c = 6$ or $a = 2, c = 3$. The requirement $-a < b \leq a < c$ excludes the case $a = 1, b = -1, c = 6$, so there are three reduced forms of discriminant -23 —namely, $x^2 + xy + 6y^2$ and $2x^2 \pm xy + 3y^2$ —yielding $h(-23) = 3$.

Since $h(7) = h(23) = 1$ (i.e., the class numbers in the real cases are both one; we omit the proof), we can double-check the facts $h(-7) = 1$ and $h(-23)$ by using the HMZ-formula, see (2.275). Since $\sqrt{7} = [2; \overline{1, 1, 1, 4}]$ and $\sqrt{23} = [4; \overline{1, 3, 1, 8}]$, we have

$$h(-7) = \frac{-1 + 1 - 1 + 4}{3} = 1 \quad \text{and} \quad h(-23) = \frac{-1 + 3 - 1 + 8}{3} = 3.$$

We conclude this section with the remark that if α is an arbitrary quadratic irrational

$$\alpha = \frac{-B + \sqrt{D}}{2A}, \quad \text{that is, } \alpha \text{ is a root of } Ax^2 + Bx + C = 0, \text{ and } D = B^2 - 4AC > 0,$$

then we have the following analog of formula (2.211):

$$M_{\alpha}(N) = \frac{\sqrt{D}}{2\pi^2} \left(\sum_{\substack{(x,y) \neq (0,0): \\ \text{primary representations}}} \frac{1}{Ax^2 + Bxy + Cy^2} \right) \frac{\log N}{\log \eta} + \text{negligible}, \quad (2.276)$$

where η is the fundamental unit in $\mathbf{Q}(\sqrt{D})$.

The proof of (2.276) is the same as that of Proposition 2.20. The guiding intuition is that if $y\alpha$ is very close to an integer x , then

$$\|y\alpha\| \sqrt{D} y = \pm A(x - y\alpha)(y\alpha - y\alpha') \approx A(x - y\alpha)(x - y\alpha') = Ax^2 + Bxy + Cy^2,$$

where $\alpha' = (-B - \sqrt{D})/2A$ is the other root of $Ax^2 + Bx + C = 0$.

Probabilistic Diophantine Approximation
Randomness in Lattice Point Counting

Beck, J.

2014, XVI, 487 p. 22 illus., Hardcover

ISBN: 978-3-319-10740-0