ON A THEOREM OF DAVENPORT AND SCHMIDT

NICKOLAS ANDERSEN AND WILLIAM DUKE

To our wives Emily and Abbey

Abstract. This work is motivated by a paper of Davenport and Schmidt, which treats the question of when Dirichlet's theorems on the rational approximation of one or of two irrationals can be improved and if so, by how much. We consider a generalization of this question in the simplest case of a single irrational but in the context of the geometry of numbers in $\mathbb{R}^2$, with the sup-norm replaced by a more general one. Results include sharp bounds for how much improvement is possible under various conditions. The proofs use semi-regular continued fractions that are characterized by a certain best approximation property determined by the norm.

1. Introduction

In 1842 Dirichlet [13] applied the pigeonhole principle to give good approximations of real numbers by rationals. One form of his theorem in one dimension is the following.

Dirichlet Approximation Theorem. *For $\alpha \in \mathbb{R}$ and any $Q \in \mathbb{Z}^+$ there are $p, q \in \mathbb{Z}$ such that $1 \leq q \leq Q$ and $|p - q\alpha| < \frac{1}{Q}$.*

Davenport and Schmidt [10] considered those α for which an improvement of this result is possible, at least when we only require that Q be sufficiently large. More precisely, let $\delta(\alpha)$ be the largest number with the property that if $c > \delta(\alpha)$ then for *every* sufficiently large Q (depending only on α), there are integers $p, q \in \mathbb{Z}$ with $1 \leq q \leq Q$ and $Q|p - \alpha q| < c$, while if $c < \delta(\alpha)$ there are arbitrarily large Q for which no such p, q exist. If $\delta(\alpha) < 1$ then we say that an improvement on Dirichlet's theorem is possible for this α. Clearly $\delta(\alpha) = 0$ for rational α so we only consider irrational α.

An easy direct argument proves the fact, perhaps surprising at first, that any irrational α for which $\delta(\alpha) < 1$ must be *badly approximable*. For α to be badly approximable means that for some $c > 0$ we have $|\alpha - \frac{p}{q}| > \frac{c}{q^2}$ for all relatively prime integers p, q with $q > 0$. Davenport and Schmidt [10] gave another proof of this that also shows that, conversely, an improvement on Dirichlet's theorem is possible for every badly approximable number. They deduced this from a formula for $\delta(\alpha)$ given in terms of the regular continued fraction expansion of α. Recall that an irrational α has a unique infinite regular continued fraction

Date: July 13, 2020.

Supported by NSF grant DMS 1701638. The second author is also supported by the Simons Foundation: Award Number 554649.

expansion

$$\alpha = b_0 + \frac{1}{b_1+}\,\frac{1}{b_2+}\cdots \overset{\text{def}}{=} b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{\ddots}}}, \tag{1.1}$$

where the partial quotients b_n satisfy $b_0 = \lfloor \alpha \rfloor$ and $b_k \in \mathbb{Z}^+$ for $k \geq 1$. Also define $u_0 = \alpha - a_0$, $v_0 = 0$ while for $n \geq 1$ let

$$u_n = \frac{1}{b_{n+1}+}\,\frac{1}{b_{n+2}+}\cdots \quad \text{and} \quad v_n = \frac{1}{b_n+}\,\frac{1}{b_{n-1}+}\,\frac{1}{b_{n-2}+}\cdots\frac{1}{b_1}. \tag{1.2}$$

Theorem. *(Davenport-Schmidt [10]) For any irrational $\alpha \in \mathbb{R}$ we have that*

$$\delta(\alpha) = \limsup_{n\to\infty} \big(1 + u_n v_n\big)^{-1}. \tag{1.3}$$

An immediate consequence of (1.3) is that the irrational $\alpha \in \mathbb{R}$ for which Dirichlet's theorem can be improved are precisely those whose continued fraction have bounded partial quotients. This condition is well-known to be equivalent to α being badly approximable [51, p. 22]. Real quadratic irrationalities are precisely those whose regular continued fraction expansions are eventually periodic, so they are badly approximable. On the other hand, they are the only known examples that are algebraic. A continued fraction discovered by Euler [15] provides an explicit example of an irrational (in fact transcendental) number that is not badly approximable, namely

$$\frac{e-1}{e+1} = \frac{1}{2+}\,\frac{1}{6+}\,\frac{1}{10+}\,\frac{1}{14+}\cdots. \tag{1.4}$$

By a well-known result of Khintchine [25, Thm 29] badly approximable numbers, although uncountable, are rare in the sense of measure theory. Thus we have the following.

Corollary. *The set of real irrationals for which Dirichlet's theorem can be improved is uncountable and has Lebesgue measure zero.*

Another consequence of the formula (1.3) is a bound for how much the Dirichlet theorem can be improved when it can be improved at all.[1]

Corollary. *The smallest value of $\delta(\alpha)$ is given by*

$$\delta(\alpha) = \tfrac{1}{10}(\sqrt{5}+5) = 0.723607\ldots, \tag{1.5}$$

when $\alpha = \frac{1}{2}(1+\sqrt{5})$.

2. Improving the Minkowski approximation theorem

Davenport and Schmidt used their theorem as a starting point to obtain results that pertain to the Dirichlet theorems about approximating two numbers simultaneously and later to simultaneous approximation of n numbers [11] (see also [50]). In this paper we will consider a different kind of generalization of Dirichlet's results, one that was conceived of by Hermite and Minkowski.

[1]For further results about the set of values of $\delta(\alpha)$ see [23] and the references therein. See also our §12.

Let $F : \mathbb{R}^2 \to \mathbb{R}$ be a fixed norm on $\mathbb{R}^2$ and $\mathcal{B}$ its unit ball. Define the stretched norm F_t for $t > 0$ by

$$F_t(x, y) = F(t^{-1}x, ty). \tag{2.1}$$

The following generalization of Dirichlet's theorem follows from the work of Minkowski. Although it was not stated directly by him, for the purposes of this paper we will refer to it as the *Minkowski approximation theorem* (in two dimensions).

Minkowski Approximation Theorem. *For a fixed norm F on $\mathbb{R}^2$ let $\Delta = \Delta_F$ be the minimal area of a parallelogram with one vertex at the origin and the other three on the boundary of $\mathcal{B}$. Fix $\alpha \in \mathbb{R}$. Then for any real $t \geq 1$ there exist integers p, q with $q > 0$ such that*

$$\Delta\, F_t^2(q, p - \alpha q) \leq 1. \tag{2.2}$$

Note that for this result we are not restricting t to be an integer. It is not hard to see that for the sup-norm the Minkowski approximation theorem implies Dirichlet's theorem. In this case $\Delta = 1$.

The idea of generalizing Dirichlet's theorem to other norms goes back at least to Hermite [19]. He applied (2.2) for the Euclidean norm, for which $\Delta = \frac{\sqrt{3}}{2}$, together with the inequality between arithmetic and geometric means. The resulting inequality implies that for any irrational α there are infinitely many integers p, q with $q > 0$ such that

$$q|p - \alpha q| < \tfrac{1}{\sqrt{3}}, \tag{2.3}$$

improving upon the corresponding upper bound 1 given by Dirichlet's theorem. Later Minkowski [33, 36] showed that (2.2) with the 1-norm given by $F(x, y) = |x| + |y|$ and for which $\Delta = \frac{1}{2}$, implies (2.3) with $\frac{1}{\sqrt{3}}$ replaced by $\frac{1}{2}$.

Given these results of Hermite and Minkowski, it is natural to study the generalization for any norm of the quantity $\delta(\alpha)$ from the Davenport-Schmidt theorem. We want this generalization to measure to what extent the Minkowski approximation theorem (2.2) can be improved for a particular α. Hence for a fixed norm F, let $\delta_F(\alpha)$ be the largest number with the property that if $c > \delta(\alpha)$ then *for every sufficiently large t* there are $p, q \in \mathbb{Z}$ with $q > 0$ such that

$$\Delta\, F_t^2(q, p - \alpha q) < c,$$

while for $c < \delta_F(\alpha)$ there are arbitrarily large t for which no such p, q exist. For a given norm we say that the Minkowski approximation theorem can be improved for irrational $\alpha \in \mathbb{R}$ if $\delta_F(\alpha) < 1$. A straightforward argument shows that when F is the sup-norm, $\delta_F = \delta$ for δ in the Davenport–Schmidt theorem.

We have only been able to obtain satisfactory results about δ_F if we make the assumption that for all $(x, y) \in \mathbb{R}^2$ the norm F satisfies

$$F(x, y) = F(|x|, |y|). \tag{2.4}$$

We also require that the norm F satisfies

$$F(0, \pm 1) = F(\pm 1, 0) = 1. \tag{2.5}$$

Definition 1. *Say that a norm F is strongly symmetric if it satisfies (2.4) and (2.5).*

The most important strongly symmetric norms are the p-norms. For $(x, y) \in \mathbb{R}^2$ and a fixed $1 \leq p < \infty$ the p-norm is defined by

$$F^{\langle p \rangle}(x, y) = (|x|^p + |y|^p)^{\frac{1}{p}},$$

while $F^{\langle \infty \rangle}(x, y) = \sup\{|x|, |y|\}$. Denote the corresponding $\mathcal{B}$ by $\mathcal{B}^p$, Δ by Δ_p and δ by δ_p. Other interesting examples are the two unique strongly symmetric norms whose unit balls are regular octagons: $\mathcal{B}^{\text{oct}_1}$ and $\mathcal{B}^{\text{oct}_2}$ (see Figure 1).

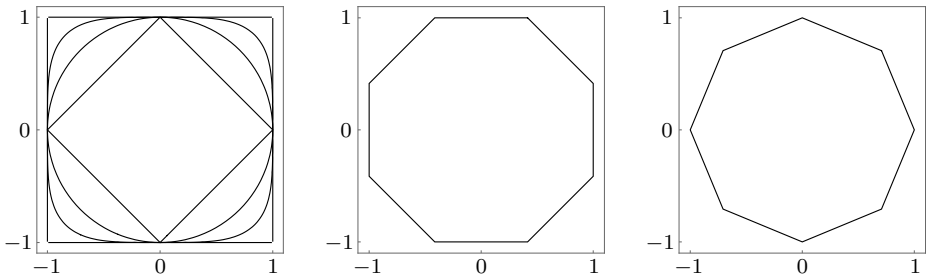

FIGURE 1. $\mathcal{B}^p$ for $p = 1, 2, 4, \infty$ and $\mathcal{B}^{\text{oct}_1}$ and $\mathcal{B}^{\text{oct}_2}$.

Our first result generalizes the first corollary of the theorem of Davenport and Schmidt. It shows that for a strongly symmetric norm the set of irrationals for which the Minkowski approximation theorem can be improved, while uncountable, is small in the sense of measure theory.

Theorem 1. *Fix a strongly symmetric norm F. Then the set of all real irrationals for which Minkowski's approximation theorem can be improved is uncountable and has Lebesgue measure zero.*

Next we have a uniform lower bound for $\delta_F(\alpha)$ for any strongly symmetric norm and any irrational α.

Theorem 2. *For any strongly symmetric norm F and any irrational $\alpha \in \mathbb{R}$ we have that*

$$\delta_F(\alpha) \geq \tfrac{1}{2}. \tag{2.6}$$

Equality in (2.6) can hold for the 1-norm. This follows from the next result since $\Delta_1 = \frac{1}{2}$. For simplicity say that an irrational $\alpha \in \mathbb{R}$ is *well approximable* if it is not badly approximable.

Theorem 3. *For any strongly symmetric norm F the smallest value of $\delta_F(\alpha)$ for a well approximable α is Δ.*

We will see in the proof of Theorem 3 that $\delta_p(\alpha) = \Delta$ for any α whose regular continued fraction has partial quotients that are eventually strictly increasing, for example $\alpha = \frac{e-1}{e+1}$ from (1.4). For the p-norm we can go further and identify the smallest value of $\delta_p(\alpha)$ for any irrational α.

Theorem 4. *For the p-norm the smallest value of $\delta_p(\alpha)$ for an irrational α is Δ_p when $1 \leq p \leq 2$ and is*

$$\tfrac{\Delta_p}{10}\left(\sqrt{5}+5\right)\left(\left(\tfrac{1}{2}(\sqrt{5}-1)\right)^p + 1\right)^{2/p}, \tag{2.7}$$

when $2 < p \leq \infty$. The value in (2.7) is attained when $\alpha = \frac{-1+\sqrt{5}}{2}$.

The value of Δ_p is given below in (4.2). See Figure 2 for graphs of Δ_p and the minimum value of δ_p. It is not the case that the Minkowski approximation theorem can always be improved for each badly approximable irrational, not even each real quadratic irrational. For example, we show at the end of §8 that

$$\delta_2\big(\tfrac{1}{2}(-1+\sqrt{3})\big) = 1. \tag{2.8}$$

Finding the norm or norms with the largest minimum value of $\delta_F(\alpha)$ among all strongly symmetric norms seems an interesting problem. The 2-norm has the largest minimum value $\frac{\sqrt{3}}{2} = 0.866025\ldots$ of δ_p among all p-norms. It can be shown that the minimum value of $\delta_F(\alpha)$ for both of the octagonal norms is $\frac{1}{8}\big(3\sqrt{2}+2\big) = 0.78033\ldots$ (see the end of §10). Among all of the examples we have considered, the 2-norm provides the largest minimum (see Figure 2).

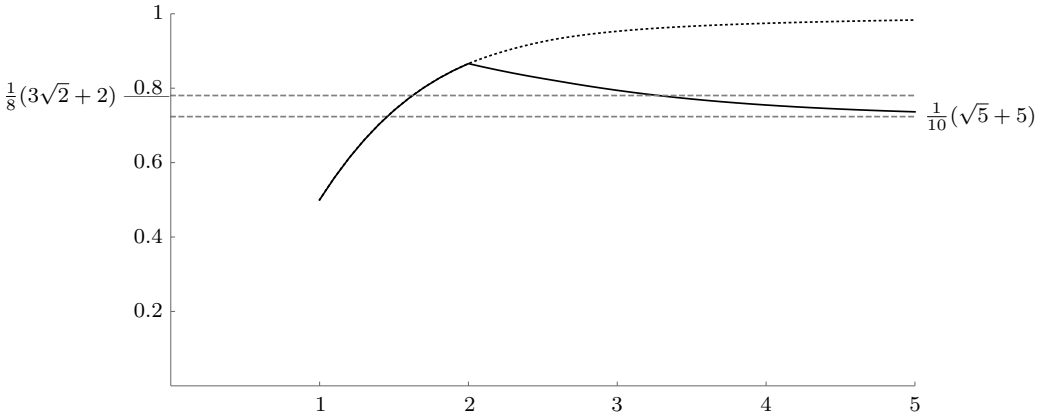

FIGURE 2. The minimum of δ_p for $p \geq 1$. The dotted line is Δ_p, the minimum of $\delta_p(\alpha)$ for well approximable α.

Remarks: Results like ours involving continuously varying norms belong to the "parametric geometry of numbers," an area that has recently seen a revival of activity stimulated by Schmidt and Summerer [52, 53]. See also [49] and its references.

The sequence (u_n, v_n) from (1.2) represents a trajectory of the dynamical system on $\Omega_0 = [0,1) \times [0,1]$ determined by the extended continued fraction map $T : \Omega_0 \to \Omega_0$ given by $T(0,v) = (0,0)$ and for $u > 0$ by

$$T(u,v) = \left(\frac{1}{u} - \left\lfloor \frac{1}{u} \right\rfloor, \frac{1}{v + \lfloor \frac{1}{u} \rfloor}\right). \tag{2.9}$$

It has an invariant measure ω with density function

$$\frac{1}{\log 2}\frac{1}{(1+uv)^2}. \tag{2.10}$$

The ergodicity of this system (see [41, 42]), which is the natural extension of the usual continued fraction dynamical system, can be used to give a different proof that Dirichlet's theorem cannot be improved for almost all real irrationals. An argument of [24], given as Lemma 5.3.11 of [9], allows one to conclude an almost all result for the special trajectories (1.2). Our proof of Theorem 1 proceeds along similar lines except that the trajectories of our dynamical system are determined by certain semi-regular continued fractions, which admit ± 1 as partial numerators. The continued fractions we need are examples of $\mathcal{S}$-expansions, which have

a well developed metrical theory again based on the ergodic theorem. For some remarks on the connection between these dynamical systems and the geodesic flow on $\mathrm{SL}(2,\mathbb{Z})\backslash\mathrm{SL}(2,\mathbb{R})$ see §12.

The second corollary of the Theorem of Davenport and Schmidt and our generalization, Theorem 4, require for their proofs information about *all*, rather than almost all trajectories. Other aspects of the continued fractions we use are needed, including a best approximation property given in terms of the norm, in order to be able to analyze in detail each individual trajectory.

3. The continued fraction associated to a norm

We want to give a generalization of the formula (1.3) of Davenport and Schmidt and for that we require, as previously mentioned, certain infinite *semi-regular* continued fraction expansions. Such a continued fraction has the form

$$a_0 + \frac{\varepsilon_1}{a_1+}\,\frac{\varepsilon_2}{a_2+}\,\frac{\varepsilon_3}{a_3+}\cdots, \quad \varepsilon_m = \pm 1,\, a_m \in \mathbb{Z} \tag{3.1}$$

where $a_m > 0$ and $a_m + \varepsilon_{m+1} \geq 1$ for all $m \geq 1$ and $a_m + \varepsilon_{m+1} \geq 2$ for infinitely many m. For any $m \geq 0$ the m^{th} convergent of this continued fraction

$$\frac{p_m}{q_m} = a_0 + \frac{\varepsilon_1}{a_1+}\,\frac{\varepsilon_2}{a_2+}\cdots\frac{\varepsilon_m}{a_m}$$

uniquely defines relatively prime integers p_m, q_m with $q_m > 0$, where $p_0 = a_0$ and $q_0 = 1$. Tietze ([57], see also [45, p. 135]) showed that there is an irrational α to which such a continued fraction converges, meaning that $\alpha = \lim_{m\to\infty} \frac{p_m}{q_m}$.

The continued fraction we need is characterized by a best approximation property stated in terms of the given strongly symmetric norm.

Definition 2. *Say that a rational number $\frac{p}{q}$ where $q > 0$ is a best approximation to α with respect to the norm F if there is a $t > 1$ depending only on $\frac{p}{q}$ such that*

$$F_t(q, p - \alpha q) < F_t(s, r - \alpha s)$$

for all rational $\frac{r}{s} \neq \frac{p}{q}$.

In the case of the sup-norm Definition 2 is equivalent to the usual one that states that a rational number $\frac{p}{q}$ with $q > 0$ is a best approximation to an irrational α if for all rational numbers $\frac{r}{s} \neq \frac{p}{q}$ with $0 < s \leq q$ we have

$$|p - \alpha q| < |r - \alpha s|$$

(see Lemma 6.1 below).

Theorem 5. *Fix a strongly symmetric norm F. Every irrational $\alpha \in \mathbb{R}$ has a unique semi-regular continued fraction expansion whose convergents are precisely the best approximations to α with respect to F.*

We will refer to this continued fraction as the *F-continued fraction of α* and, for the p-norm, as the *p-continued fraction of α*. For the sup-norm the ∞-continued fraction is closely related to, but not always equal to, the regular continued fraction. Suppose that

$$\alpha = b_0 + \frac{1}{b_1+}\,\frac{1}{b_2+}\,\frac{1}{b_3+}\cdots \tag{3.2}$$

is the regular continued fraction of an irrational α. Recall that Lagrange showed ([28], see also [45, §15]) that every best approximation in the usual sense is a convergent of the regular continued fraction of α and that every convergent, except possibly b_0, is a best approximation to α. In view of Theorem 5, (3.2) coincides with the ∞-continued fraction of α if and only if $b_1 > 1$. If $b_1 = 1$ the ∞-continued fraction of α is

$$\alpha = b_0 + 1 + \frac{-1}{b_2 + 1+} \, \frac{1}{b_3+} \cdots . \tag{3.3}$$

This is an example of a singularization, which has the effect of contracting the regular continued fraction by removing $b_1 = 1$ and the convergent b_0, which is *not* a best approximation to α in this case. This well-known exceptional case does not occur for the ∞-continued fraction. With this one possible exception, however, the convergents of the ∞-continued fraction and those of the regular continued fraction coincide.

For any $1 \le p < \infty$, the Minkowski Approximation Theorem and the inequality between arithmetic and geometric means immediately gives that a necessary condition for a regular convergent $\frac{r_n}{s_n}$ of an irrational α to be a convergent of the p-continued fraction is that

$$s_n |r_n - \alpha s_n| \le (4^{1/p} \Delta_p)^{-1}. \tag{3.4}$$

For $p = 1$, when the right hand side is $\frac{1}{2}$, Minkowski [34] showed that (3.4) is also sufficient.

Just as the formula (1.3) is given in terms of the sequence u_n, v_n coming from the regular continued fraction, our generalization will be given in terms of a sequence μ_m, ν_m determined by our continued fraction $\alpha = a_0 + \frac{\varepsilon_1}{a_1+} \, \frac{\varepsilon_2}{a_2+} \, \frac{\varepsilon_3}{a_3+} \cdots$. Namely, for a fixed norm we define $\mu_0 = \alpha$ and $\nu_0 = 0$, while for $m \ge 1$ we let

$$\mu_m = \frac{\varepsilon_{m+1}}{a_{m+1}+} \, \frac{\varepsilon_{m+2}}{a_{m+2}+} \cdots \quad \text{and} \quad \nu_m = \frac{1}{a_m+} \, \frac{\varepsilon_m}{a_{m-1}+} \, \frac{\varepsilon_{m-1}}{a_{m-2}+} \cdots \frac{\varepsilon_2}{a_1}. \tag{3.5}$$

For a general strongly symmetric norm we will express $\delta_F(\alpha)$ in terms of these numbers μ_m, ν_m in §7 below. For the p-norm the formula is completely explicit and we give it here. For p with $1 \le p < \infty$ let

$$D_p(u, v) = \frac{1}{1 + uv} \left(\frac{(1 - |u|^p v^p)^2}{(1 - |u|^p)(1 - v^p)} \right)^{\frac{1}{p}}, \tag{3.6}$$

while when $p = \infty$ set $D_\infty(u, v) = \lim_{p \to \infty} D_p(u, v) = (1 + uv)^{-1}$.

Theorem 6. *Fix $1 \le p \le \infty$. For any irrational α whose p-continued fraction is (3.1) we have that*

$$\delta_p(\alpha) = \limsup_{m \to \infty} \Delta_p \, D_p\big(\mu_m, \nu_m\big),$$

where μ_m, ν_m are given above in (3.5).

The F-continued fraction of an irrational α for any strongly symmetric norm is an example of an $\mathcal{S}$-expansion. Their theory has been developed by Kraaikamp [27] and others (see also [4], [9] and [22]). Recall the definition of T and ω from (2.9) and (2.10). A Borel set $\mathcal{S} \subset \Omega_0$ is called a singularization area if $\omega(\partial \mathcal{S}) = 0$ and if

(i) $\mathcal{S} \subseteq [\frac{1}{2}, 1) \times [0, 1]$ and
(ii) $T\mathcal{S} \cap \mathcal{S} \subseteq \{(\beta, \beta)\}$, where $\beta = \frac{1}{2}(-1 + \sqrt{5})$.

The $\mathcal{S}$-expansion of an irrational α is obtained from the regular continued fraction (1.1) by changing

$$\alpha = \cdots \frac{1}{b_n+} \frac{1}{1+} \frac{1}{b_{n+2}+} \cdots \quad \text{into} \quad \alpha = \cdots \frac{1}{(b_n+1)+} \frac{-1}{(b_{n+2}+1)+} \cdots$$

for each n such that $(u_n, v_n) \in \mathcal{S}$. Note that $(u_n, v_n) \in \mathcal{S}$ implies that $b_{n+1} = 1$ by (i). Also (ii) implies that this procedure is unambiguous. The result is a unique semi-regular continued fraction (see Section 4 of [27] for more details) for α whose convergents are precisely those regular convergents $\frac{r_n}{s_n}$ where $n \geq 0$ is such that $(u_n, v_n) \notin \mathcal{S}$. For example, the ∞-continued fraction discussed above is the $\mathcal{S}$ expansion for $\mathcal{S} = [\frac{1}{2}, 1) \times \{0\}$.

Theorem 7. *Fix a strongly symmetric norm F. There exists a singularization area $\mathcal{S}$ so that the F-continued fraction of any irrational α is the $\mathcal{S}$-expansion of α.*

As usual, we denote $\mathcal{S}$ by $\mathcal{S}_p$ in the case of the p-norm (see Figure 3).

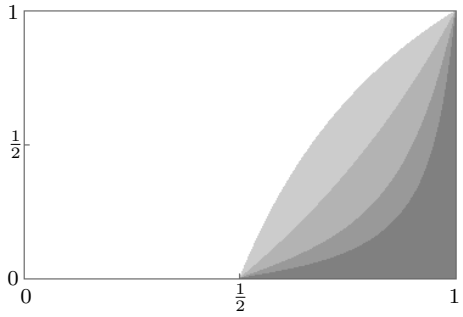

FIGURE 3. The $\mathcal{S}_p$ regions for $p = 1, 2, 3, 4$.

Remarks: At the beginning of the paper [33], Minkowski states without proof several of the main properties of the p-continued fraction for any p, including the best approximation property. Our proof of Theorem 5, which allows for F to be any strongly symmetric norm, was strongly influenced by his ideas. As previously mentioned, Minkowski [34] also gave the remarkable result that for the 1-continued fraction the necessary condition (3.4) is also sufficient. Unsurprisingly, this also follows from our arguments. In addition, the 2-continued fraction has actually been studied since the time of Hermite [18], especially by Humbert [20, 21]. It is also closely connected to the improper modular billiards studied in [1]. It was shown in [26] that the 1-continued fraction (known as Minkowski's diagonal continued fraction) is an $\mathcal{S}$-expansion. For related work on the 1-continued fraction see [40]. That the p-continued fraction for $p \neq 1, \infty$ is also an $\mathcal{S}$-expansion seems to be new.

In the next section we will review some basic facts from the geometry of numbers in the case we need, namely in two dimensions. Seven sections, each with the proof of one of our theorems, follow afterward. The theorems will be proven in the following order:

$$2 \to 5 \to 6 \to 7 \to 3 \to 4 \to 1.$$

Some concluding remarks are then given. Finally, an appendix contains a number of technical lemmas and their proofs that we will refer to as needed in the main body of the paper.

4. Geometry of numbers

As above let F be a fixed norm on $\mathbb{R}^2$. This means that for $P, P' \in \mathbb{R}^2$ we have

(i) $F(P) \geq 0$ and $F(P) = 0$ if and only if $P = (0,0)$
(ii) $F(tP) = |t|F(P)$ for $t \in \mathbb{R}$
(iii) $F(P+P') \leq F(P) + F(P')$.

The unit ball of the norm is

$$\mathcal{B} = \{P \in \mathbb{R}^2; F(P) < 1\}.$$

This $\mathcal{B}$ is open, bounded, convex and symmetric around 0 and every such body arises as the unit ball of some norm (see e.g. [54]). Denote by area($\mathcal{B}$) the Lebesgue measure of $\mathcal{B}$ on $\mathbb{R}^2$. It is convenient to define the stretched ball for $t > 0$

$$\mathcal{B}_t = \{(x,y) \in \mathbb{R}^2; F_t(x,y) < 1\}.$$

Let $L \subset \mathbb{R}^2$ be a (full) lattice. By the determinant of L, denoted $\det L$, we mean $|\det g|$ for any $g \in \mathrm{GL}(2,\mathbb{R})$ whose rows give a $\mathbb{Z}$-basis for L. The lattice L is *admissible* for $\mathcal{B}$ if $\mathcal{B}$ contains no other points of L than $(0,0)$. The following result is fundamental [36]:

Minkowski's First Convex Body Theorem. *If L is admissible for $\mathcal{B}$ then*

$$\operatorname{area} \mathcal{B} \leq 4 \det L.$$

The *critical determinant* of $\mathcal{B}$, denoted $\Delta(\mathcal{B})$ or simply Δ, is the infimum of all determinants of lattices admissible for $\mathcal{B}$. Building on work of Minkowski [35, 37], Mahler [29] proved that lattices with determinant Δ actually exist, and these are called *critical lattices.* Minkowski's first convex body theorem implies that

$$\Delta \geq \tfrac{1}{4}(\operatorname{area} \mathcal{B}). \tag{4.1}$$

This is sharp for the 1-norm and the sup-norm.

Apparently, if we wish to evaluate $\delta_F(\alpha)$ exactly we must also know Δ exactly. Finding the critical determinant of a given $\mathcal{B}$ is the main problem of the geometry of numbers in $\mathbb{R}^2$. Although the n-dimensional version of this problem is apparently intractable in general, here it is approachable. For a given critical lattice L for $\mathcal{B}$ the boundary of $\mathcal{B}$ must contain a $\mathbb{Z}$-basis $\{P, P'\}$ for L as well as their sum $P + P'$. Furthermore, the lattice generated by any pair of points P, P' with $P, P', P + P'$ on the boundary of $\mathcal{B}$ is admissible for $\mathcal{B}$ (see [7, Thm XI p. 160]). Therefore, as Minkowski already knew, computing Δ amounts to solving the (generally quite difficult) calculus problem of minimizing the area of a parallelogram with one vertex at the origin and the three others on the boundary of $\mathcal{B}$. This justifies our definition of Δ in the statement of the Minkowski approximation theorem.

Next we review what is known about the value of Δ_p for all p. Let

$$\Delta_p^{(0)} = (1 - 2^{-p})^{\frac{1}{p}} \quad \text{and} \quad \Delta_p^{(1)} = 2^{-\frac{2}{p}} \, \tfrac{1+\tau_p}{1-\tau_p},$$

where $0 < \tau_p < \frac{1}{2}$ satisfies $\tau_p^p + 1 = 2(1-\tau_p)^p$. A modification of a conjecture of Minkowski [37, p. 51–58] made by Davis [12] states that

$$\Delta_p = \min\{\Delta_p^{(0)}, \Delta_p^{(1)}\}. \tag{4.2}$$

Furthermore, there is a unique value $2.57 < \rho < 2.58$ so that $\Delta_p = \Delta_p^{(0)}$ when $2 \leq p \leq \rho$, while otherwise $\Delta_p = \Delta_p^{(1)}$. Many mathematicians obtained partial results, among them Mordell [39], Davis [12], Cohn [8], Watson [58, 59] and Malyshev [31]. Building on their

work, the proof of the full conjecture was finally completed by Glazunov, Golovanov and Malyshev [16].

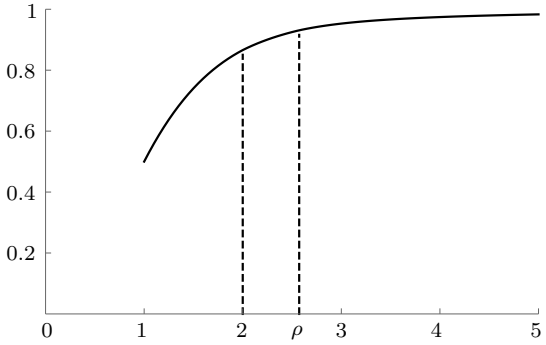

FIGURE 4. Δ_p for $1 \le p \le 5$

In the case of the p-norm parallelograms that minimize the area may be given explicitly. For $2 \le p \le \rho$ we may take the parallelogram with vertices at $0, P, P', P + P'$ where

$$P = (1, 0) \text{ and } P' = \left(\tfrac{1}{2}, \tfrac{1}{2}(2^p - 1)^{\frac{1}{p}}\right). \tag{4.3}$$

For $1 \le p \le 2$ or $\rho \le p \le \infty$ we may take

$$P = \left(2^{-\frac{1}{p}}(1 - \tau_p)^{-1}, -2^{-\frac{1}{p}}\tau_p(1 - \tau_p)^{-1}\right) \text{ and } P' = \left(2^{-\frac{1}{p}}, 2^{-\frac{1}{p}}\right) \tag{4.4}$$

where again $0 < \tau_p < \frac{1}{2}$ solves $\tau_p^p + 1 = 2(1 - \tau_p)^p$. Except when $p = 1, 2$ or ∞ these parallelograms are unique up to obvious symmetries. When $p = 1, 2$ or ∞ there are infinitely many essentially different minimizing parallelograms. They are easily parameterized. For example, when $p = 2$ all are obtained by rotating the standard hexagonal lattice coming from (4.3).

Minkowski's method can be restated as saying that $3\Delta(\mathcal{B})$ is the minimal area of an affinely regular symmetric hexagon inscribed in $\mathcal{B}$. A useful alternative due to Reinhardt [48] is that $4\Delta(\mathcal{B})$ is the minimum area of a symmetric convex circumscribed hexagon (see also [7, p. 239] or [17, Thm 2 p. 243]). Using this fact, that he also found independently, Mahler [30] computed $\Delta(\mathcal{B}^{\text{oct}_1}) = \sqrt{2} - \frac{1}{2}$ for the regular octagon from Figure 2. Thus we also have $\Delta(\mathcal{B}^{\text{oct}_2}) = \frac{1}{8}\left(3\sqrt{2} + 2\right)$, obtained by scaling.

5. A Minkowski-type algorithm

In this section we will prove Theorem 2. First we give a needed definition. A *minimal basis* for a lattice $L \subset \mathbb{R}^2$ with respect to a norm F is a $\mathbb{Z}$-basis $\{P, P'\}$ for L with the property that

$$F(P) = F(P') = \min_{P_0 \in L \setminus \{0\}} F(P_0).$$

For $\alpha \in \mathbb{R}$ let

$$L_\alpha = (1, -\alpha)\mathbb{Z} + (0, 1)\mathbb{Z}. \tag{5.1}$$

Obviously L_α has determinant one.

To prove Theorems 2 and 5 we require an algorithm that constructs a sequence of points $P_n \in L_\alpha$ and positive numbers t_m such that $\{P_{m-1}, P_m\}$ gives a minimal basis for L_α with

respect to F_{t_m}. We also want P_{m-1} to have the smallest norm $\|P_{m-1}\|_t$ among non-zero points in L_α for any $t \in (t_{m-1}, t_m)$. We will start with $P_{-1} = (0, 1)$ and $t_{-1} = 1$. Roughly speaking, given P_{m-1}, to find the new point P_m and the associated t_m, we simultaneously expand $\mathcal{B}$ in the x-direction while shrinking in the y-direction in a such a way that P_{m-1} remains on its boundary until we encounter P_m. We then repeat this procedure starting with P_m (see Figure 5). Our algorithm will produce pairs of lattice points in L_α that are linearly independent over $\mathbb{R}$ and on a ball for which L_α is admissible. First we need to know that they give a basis for L_α.

Lemma 5.1. *Let $\alpha \in \mathbb{R}$ be irrational and F be a fixed strongly symmetric norm. Suppose that $P, P' \in L_\alpha$ lie on the boundary of $\mathcal{B}_t$ for some $t > 0$ and are linearly independent over $\mathbb{R}$. If L_α is admissible for $\mathcal{B}_t$ then $\{P, P'\}$ gives a $\mathbb{Z}$-basis for L_α.*

Proof. Consider the sublattice $P\mathbb{Z} + P'\mathbb{Z}$ of L_α generated by these lattice points. By Minkowski's first convex body theorem its index in L_α can only be 1 or 2. In the latter case suppose that $P = aQ + bQ'$ and $P' = cQ + dQ'$ where $L_\alpha = Q\mathbb{Z} + Q'\mathbb{Z}$, so $|ad - bc| = 2$. If a were even and c odd we would have that b is even and so $\frac{1}{2}P = (\frac{a}{2})Q + (\frac{b}{2})Q'$ would be a non-zero point in $\mathcal{B}_t \cap L_\alpha$. A similar argument disallows c being even and a odd. Thus a and c are either both even or both odd. Similarly b and d are either both even or both odd. In any case

$$\tfrac{1}{2}(P + P') = (\tfrac{a+c}{2})Q + (\tfrac{b+d}{2})Q' \quad \text{and} \quad \tfrac{1}{2}(P - P') = (\tfrac{a-c}{2})Q + (\tfrac{b-d}{2})Q'$$

are distinct points of L_α. As $\mathcal{B}_t$ is convex they must lie on the boundary of $\mathcal{B}_t$. It follows that $\mathcal{B}_t$ must be a parallelogram and strong symmetry implies it is a stretched ball for either the 1-norm or the sup norm. As the corners and midpoints of the sides are lattice points we would have to have that L_α contains points of the x-axis, i.e. α would be rational. □

In the next lemma we make the whole process precise. Clearly to represent any L_α we may assume that $\alpha \in (-\frac{1}{2}, \frac{1}{2}]$.

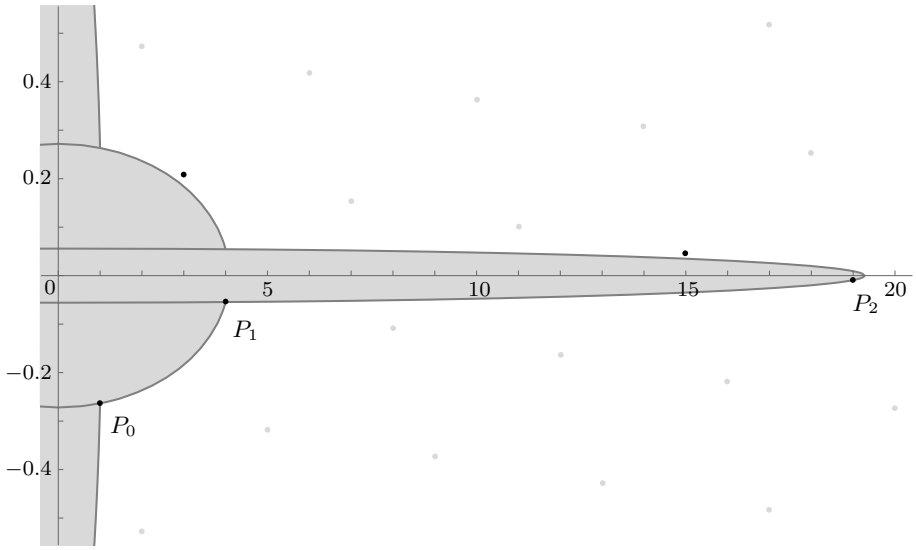

FIGURE 5. The lattice L_α for $\alpha = \frac{1}{1+}\frac{1}{2+}\frac{1}{1+}\frac{1}{3+}\frac{1}{1+}\frac{1}{4+}\cdots$. Dark lattice points correspond to the regular convergents of α. The points $P_0 = (1, -\alpha), P_1 = (4, 1 - 4\alpha)$, and $P_2 = (19, 5 - 19\alpha)$ give best approximations for the 2-norm.

Lemma 5.2. *Fix a strongly symmetric norm F and an irrational $\alpha \in (-\frac{1}{2}, \frac{1}{2})$. There is a sequence $1 = t_{-1} \leq t_0 < t_1 < t_2 < \dots$ tending to ∞ and for each $m = -1, 0, 1, \dots$ there is a $P_m = (x_m, y_m) \in L_\alpha$ with the following properties. For each $m \geq 0$*

(i) $x_m > x_{m-1}$ and $|y_m| < |y_{m-1}|$,
(ii) $\{P_{m-1}, P_m\}$ gives a minimal basis for L_α with respect to F_{t_m},
(iii) for any $t \in (t_{m-1}, t_m)$ there is no $P' = (x', y') \in L_\alpha$ different from P_{m-1} with $x' > 0$ and

$$F_t(P') \leq F_t(P_{m-1}).$$

Proof. Consider for $P = (x, y) \in L_\alpha$ the ball

$$\mathcal{B}(P, t) \stackrel{\text{def}}{=} \{P' \in \mathbb{R}^2; F_t(P') < F_t(P)\}, \tag{5.2}$$

for which $\operatorname{area} \mathcal{B}(P, t) = F_t^2(P) \operatorname{area} \mathcal{B}$. Suppose that L_α is admissible for $\mathcal{B}(P, t)$. Now by Lemma A.2

$$F_t(P) = F(t^{-1}x, ty) \geq F(0, ty) = t|y|.$$

Thus, as long as $y \neq 0$, by Minkowski's first convex body theorem there will be a maximal $t' \geq t$ for which L_α is admissible for $\mathcal{B}(P, t')$. For any of the resulting $P' \neq -P$ with $F_{t'}(P') = F_{t'}(P)$, we have by Lemma 5.1 that $\{P, P'\}$ gives a minimal basis for L_α with respect to $F_{t'}$.

Let $P_{-1} = (0, 1)$. Then L_α is admissible for $\mathcal{B}(P_{-1}, 1)$. Let $t_0 \geq 1$ be maximal for which L_α is admissible for $\mathcal{B}(P_{-1}, t_0)$. Our assumption that $\alpha \in (-\frac{1}{2}, \frac{1}{2})$ implies that we can take $P_0 = (1, -\alpha)$ as a solution to $F_{t_0}(P_0) = F_{t_0}(P_{-1})$.

Now L_α is admissible for $\mathcal{B}(P_0, t_0)$ and so we find $t_1 > t_0$ maximal so that L_α is admissible for $\mathcal{B}(P_0, t_1)$. That $t_1 > t_0$ with strict inequality is assured by our choice of P_0. Among the finitely many $P' = (x', y') \in L_\alpha$ with $F_{t_1}(P') = F_{t_1}(P_0)$ there will be unique one with maximal x' since α is irrational. We let $P_1 = (x_1, y_1)$ be this point. Clearly $x_1 > x_0$ and $|y_1| < |y_0|$.

We continue this process to construct t_m and P_m. That we have $t_m > t_{m-1}$ is guaranteed by choosing among the new points on the boundary the one with maximal x-coordinate. From the form of L_α, where α is irrational, it follows that $x_m > x_{m-1}$ and $|y_{m-1}| > |y_m| > 0$ for each $m \geq 0$ and that this process never terminates.

We will have all the stated properties of t_m and P_m once we show that $t_m \to \infty$. We have by Lemma A.2 and Minkowski's first convex body theorem again that for each $m \geq 0$

$$x_m t_m^{-1} = F(x_m t_m^{-1}, 0) \leq F(x_m t_m^{-1}, y_m t_m) = F_{t_m}(P_m) \leq 2(\operatorname{area} \mathcal{B})^{-\frac{1}{2}}.$$

Thus $t_m \gg x_m \to \infty$ as $x_m > x_{m-1}$ are integers. □

Proof of Theorem 2. Observe that $\mathcal{B}(P_m, t_m)$ as defined by (5.2), with P_m and t_m from Lemma 5.2, contains a parallelogram of area 2 since $\{P_{m-1}, P_m\}$ is a minimal basis for L_α with respect to F_{t_m}. Therefore

$$\Delta F_{t_m}^2(P_m) \geq 2\Delta(\operatorname{area} \mathcal{B})^{-1} \geq \tfrac{1}{2},$$

where to get the second inequality we have applied Minkowski's bound (4.1). Since $t_m \to \infty$ as $n \to \infty$ we have that

$$\delta_F(\alpha) \geq \Delta \limsup_{m \to \infty} F_{t_m}^2(P_m) \geq \tfrac{1}{2}, \tag{5.3}$$

thus proving Theorem 2. □

6. The continued fraction

Next we relate to each other the two definitions of best approximation given in and below Definition 2.

Lemma 6.1. *If the fraction $\frac{p}{q}$ with $q > 0$ is a best approximation of an irrational α with respect to a strongly symmetric norm F then it is a best approximation in the usual sense. Conversely, if $\frac{p}{q}$ with $q > 0$ is a best approximation of an irrational α in the usual sense it is a best approximation with respect to the sup-norm.*

Proof. Suppose that there is a $t > 1$ such that $\frac{r}{s} \neq \frac{p}{q}$ with $s > 0$ implies

$$F_t(q, p - \alpha q) < F_t(s, r - \alpha s).$$

If $s \leq q$ then $|r - \alpha s| > |p - \alpha q|$ by Lemma A.2.

Conversely, suppose that $\frac{r}{s} \neq \frac{p}{q}$ and $0 < s \leq q$ implies that $|p - \alpha q| < |r - \alpha s|$. Choose t such that $t^{-1}q = t|p - \alpha q|$ and note that such a $t > 1$.

If $0 < s \leq q$ and $\frac{r}{s} \neq \frac{p}{q}$ then

$$\sup(qt^{-1}, |p - \alpha q|t) = t|p - \alpha q| < t|r - \alpha s| \leq \sup(st^{-1}, |r - \alpha s|t).$$

If $s > q$ then

$$\sup(qt^{-1}, |p - \alpha q|t) = t^{-1}q < t^{-1}s \leq \sup(st^{-1}, |r - \alpha s|t).$$

This finishes the proof. □

Proof of Theorem 5. For any $\alpha \in \mathbb{R}$ write $\alpha = \alpha' + a_0$, where $a_0 \in \mathbb{Z}$ and $\alpha' \in (-\frac{1}{2}, \frac{1}{2}]$. Suppose that α is irrational. In the notation of Lemma 5.2 (taking there $\alpha = \alpha'$) for $m \geq -1$ write $P_m = (x_m, y_m)$. For $m \geq 0$ define

$$g_m = \begin{pmatrix} x_m & y_m \\ x_{m-1} & y_{m-1} \end{pmatrix}$$

and set $g_{-1} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & -\alpha \end{smallmatrix}\right)$. We know by Lemma 5.2 that for each $m \geq 1$ there is a positive integer a_m and $\varepsilon_n = \pm 1$ so that

$$g_m = \begin{pmatrix} a_m & \varepsilon_m \\ 1 & 0 \end{pmatrix} g_{m-1}. \tag{6.1}$$

This also holds for $m = 0$ if we set $\varepsilon_0 = 1$. Clearly for $m \geq 0$

$$\gamma_m \stackrel{\text{def}}{=} \det g_m = (-1)^m \varepsilon_1 \cdots \varepsilon_m. \tag{6.2}$$

The numerator p_m and denominator q_m of the convergents

$$\frac{p_m}{q_m} = a_0 + \frac{\varepsilon_1}{a_1 +} \frac{\varepsilon_2}{a_2 +} \cdots \frac{\varepsilon_m}{a_m}$$

of our continued fraction are determined recursively for $m \geq 0$ through

$$p_m = a_m p_{m-1} + \varepsilon_m p_{m-2}, \qquad p_{-1} = 1, \quad p_{-2} = 0, \tag{6.3}$$

$$q_m = a_m q_{m-1} + \varepsilon_m q_{m-2}, \qquad q_{-1} = 0, \quad q_{-2} = 1. \tag{6.4}$$

It is easy to see that for $m \geq 0$

$$\begin{pmatrix} p_m & q_m \\ p_{m-1} & q_{m-1} \end{pmatrix} = \begin{pmatrix} a_m & \varepsilon_m \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_{m-1} & \varepsilon_{m-1} \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{6.5}$$

By (6.1) and (6.5) for each $m \geq 0$ we get that

$$g_m = \begin{pmatrix} x_m & y_m \\ x_{m-1} & y_{m-1} \end{pmatrix} = \begin{pmatrix} q_m & p_m - \alpha q_m \\ q_{m-1} & p_{m-1} - \alpha q_{m-1} \end{pmatrix}. \tag{6.6}$$

By Lemma 5.2 the basis of rows of g_m is minimal for the norm F_{t_m}. Choose any $t \in (t_m, t_{m+1})$. By (iii) of Lemma 5.2 for any $\frac{r}{s} \neq \frac{p_m}{q_m}$ we have

$$F_t(q_m, p_m - \alpha q_m) < F_t(s, r - \alpha s). \tag{6.7}$$

Conversely, suppose that $\frac{r}{s}$ is a best approximation to α with respect to F, and write $Q = (s, r - \alpha s)$. Then for some $t > 1$, we have $F_t(Q) \leq F_t(P_m)$ for all $m \geq 0$. By Lemma 5.2 it cannot happen that $t \in (1, t_0)$ since in that case we would have to have $Q = (0, 1)$ and so $s = 0$. Also we cannot have that $t = t_m$ for any $m \geq 0$. On the other hand, if m is such that $t \in (t_m, t_{m+1})$, then by Lemma 5.2 we have $Q = P_m$. It follows that the convergents are precisely the best approximations to α with respect to F.

That the continued fraction converges to α now follows from the first statement of Lemma 6.1 and Lagrange's theorem mentioned below Theorem 5, since they imply that each convergent of our continued fraction is a convergent of the regular continued fraction. It remains to show that it is semi-regular. The condition $\varepsilon_{m+1} + a_m \geq 1$ for all $m \geq 1$ will follow once we relate the μ_m, ν_m from (3.5) to the points $P_m = (x_m, y_m)$, which is also needed to prove our generalization of (1.3).

Lemma 6.2. *For x_m, y_m from (6.6) and μ_m, ν_m from (3.5) we have for $m \geq 0$ that*

$$\mu_m = -\frac{y_m}{y_{m-1}} \quad \text{and} \quad \nu_m = \frac{x_{m-1}}{x_m}. \tag{6.8}$$

Proof. The proof is an adaptation to more general continued fractions of standard arguments used for regular continued fractions (see [51]).

To start with, by (6.6)

$$-\frac{y_m}{y_{m-1}} = \frac{-p_m + \alpha q_m}{p_{m-1} - \alpha q_{m-1}}. \tag{6.9}$$

By (6.5) and (6.2) we have

$$q_{m+1} p_m - p_{m+1} q_m = \gamma_{m+1}. \tag{6.10}$$

Together with (6.3) and (6.4), this yields the following formal identity between rational functions with variables $a_1, \ldots, a_{m+1}$:

$$p_m - q_m \frac{p_{m+1}}{q_{m+1}} = \frac{\gamma_{m+1}}{a_{m+1} q_m + \varepsilon_{m+1} q_{m-1}} \quad \text{where} \quad \frac{p_{m+1}}{q_{m+1}} = a_0 + \frac{\varepsilon_1}{a_1 +} \frac{\varepsilon_2}{a_2 +} \cdots \frac{\varepsilon_{m+1}}{a_{m+1}}. \tag{6.11}$$

The m^{th} *complete quotient* α_m of the expansion $\alpha = \frac{\varepsilon_1}{a_1 +} \frac{\varepsilon_2}{a_2 +} \cdots$ is defined recursively by $\alpha_0 = \alpha$ and for $m \geq 0$ through

$$\alpha_{m+1} = \frac{\varepsilon_{m+1}}{\alpha_m - a_m}.$$

It follows that for $m \geq 0$ we have

$$\alpha = a_0 + \frac{\varepsilon_1}{a_1 +} \frac{\varepsilon_2}{a_2 +} \cdots \frac{\varepsilon_{m+1}}{\alpha_{m+1}}. \tag{6.12}$$

By (6.11) upon setting the variable $a_{m+1} = \alpha_{m+1}$ and using (6.12) we derive that

$$p_m - q_m \alpha = \frac{\gamma_{m+1}}{\alpha_{m+1} q_m + \varepsilon_{m+1} q_{m-1}}.$$

Next solve this equation for α_{m+1} and use (6.10) with m in place of $m+1$ to get

$$\alpha_{m+1} = \frac{\varepsilon_{m+1}(-p_{m-1} + q_{m-1}\alpha)}{p_m - q_m\alpha}. \tag{6.13}$$

From (6.12) we have

$$\alpha_{m+1} = a_{m+1} + \frac{\varepsilon_{m+2}}{a_{m+2}+}\,\frac{\varepsilon_{m+3}}{a_{m+3}+}\cdots. \tag{6.14}$$

so by (3.5)

$$\mu_m = \frac{\varepsilon_{m+1}}{\alpha_{m+1}}. \tag{6.15}$$

The first formula of (6.8) now follows from (6.9) and (6.13).

To prove the second formula of (6.8) start with $\frac{q_{m-1}}{q_m} = \frac{x_{m-1}}{x_m}$ from (6.6). By (3.5) we have that $v_0 = 0$ while for $m \geq 0$

$$\nu_{m+1} = \frac{1}{a_{m+1} + \varepsilon_{m+1}\nu_m}.$$

Using (6.4) we see that $\frac{q_{m-1}}{q_m}$ satisfies the same recurrence. □

We now finish the proof of Theorem 5 by showing that our expansion of the irrational α is semi-regular. By (6.15) and (6.8) we have $|\alpha_m| > 1$ for all $m \geq 1$. Thus

$$\alpha_m = a_m + \frac{\varepsilon_{m+1}}{\alpha_{m+1}} \geq 1 + \frac{\varepsilon_{m+1}}{\alpha_{m+1}} > 0.$$

Now suppose that we had $\varepsilon_{m+1} = -1$ and $a_m = 1$ for some $m \geq 1$. We would then have from (6.14) that $\alpha_m < 1$ which is impossible. It follows that $\varepsilon_{m+1} + a_m \geq 1$ for all $m \geq 1$. Now suppose that $\varepsilon_{m+1} + a_m = 1$ for all but finitely many m. Then for all sufficiently large m we have $a_m = 2$ and $\varepsilon_{m+1} = -1$. But $2 + \frac{-1}{2+}\,\frac{-1}{2+}\,\frac{-1}{2+}\cdots = 1$ which contradicts that α is irrational. This completes the proof of Theorem 5. □

7. A formula for $\delta_F(\alpha)$

We will deduce Theorem 6 from a formula for $\delta_F(\alpha)$ for any strongly symmetric norm F given in terms of the quantities μ_m, ν_m. As usual, we may identify the space of all lattices of determinant one with $\Gamma\backslash G$ where $G = \mathrm{SL}(2,\mathbb{R})$ and $\Gamma = \mathrm{SL}(2,\mathbb{Z})$ by means of

$$g = \left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right) \mapsto L(g) \stackrel{\text{def}}{=} (x,y)\mathbb{Z} + (x',y')\mathbb{Z}. \tag{7.1}$$

Let $\mathcal{D}$ be the set of $g = \left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right) \in G$ such that

$$F(x,y) = F(x',y') \text{ and} \tag{7.2}$$

$$0 \leq x' < x \text{ and } |y| < y'. \tag{7.3}$$

For $g \in \mathcal{D}$ let $F(g) = F(x,y)$.

Lemma 7.1. *The map* $\Phi : \mathcal{D} \to (-1,1) \times [0,1)$ *given by*

$$\Phi\left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right) = \left(-\tfrac{y}{y'}, \tfrac{x'}{x}\right)$$

is a continuous bijection.

Proof. The inverse of Φ is given by

$$(u,v) \mapsto \tfrac{1}{\sqrt{1+uv}} \left(\begin{smallmatrix} t^{-1} & -ut \\ t^{-1}v & t \end{smallmatrix}\right). \tag{7.4}$$

By Lemma A.4 we see that $t = t(u,v) > 0$ exists and is uniquely determined by the condition $F_t(1,-u) = F_t(v,1)$. □

The function

$$D_F(u,v) \stackrel{\text{def}}{=} F^2(\Phi^{-1}(u,v)) \tag{7.5}$$

is easily seen to be continuous on $(-1,1) \times [0,1)$.

The following is our generalization of the formula (1.3).

Lemma 7.2. *Fix a strongly symmetric norm F. For any irrational α whose continued fraction associated to the norm is (3.1) we have that*

$$\delta_F(\alpha) = \limsup_{m\to\infty} \Delta\, D_F\big(\mu_m, \nu_m\big),$$

where μ_m, ν_m are given above in (3.5).

Proof. Fix an m and write as before $P_m = (x_m, y_m)$. Let

$$\Phi^{-1}(\mu_m, \nu_m) = \left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right)$$

where $0 \le x' < x$ and $|y| < y'$. Recall that by (i) of Lemma 5.2 we know that

$$0 \le x_{m-1} < x_m \quad \text{and} \quad |y_m| < |y_{m-1}|.$$

Equation (7.4) and Lemmas 5.2 and 6.2 now imply that

$$x = t_m^{-1} x_m, \;\; x' = t_m^{-1} x_{m-1}, \;\; y = \gamma_m t_m y_m, \;\; y' = \gamma_m t_m y_{m-1},$$

where $\gamma_m = \pm 1$ was defined in (6.2). Note that in this case $\gamma_m = \operatorname{sgn} y_{m-1}$. By strong symmetry of the norm we have $F_{t_m}(x,y) = F_{t_m}(x',y') = F_{t_m}(P_m)$. Hence

$$F_{t_m}^2(P_m) = F^2(\Phi^{-1}(\mu_m,\nu_m)) = D_F(\mu_m,\nu_m).$$

Now we need to show that

$$\delta_F(\alpha) = \Delta \limsup_{m\to\infty} F_{t_m}^2(P_m). \tag{7.6}$$

For $t \ge 1$ let $m(t)$ be such that $t_{m(t)} \le t \le t_{m(t)+1}$. By Lemma 5.2 we have that

$$\delta_F(\alpha) \le \Delta \limsup_{t\to\infty} F_{t_{m(t)}}^2(P_m),$$

and by Lemma A.6 we have that

$$F_t(P_m) \le \max\big(F_{t_m}(P_m), F_{t_{m+1}}(P_m)\big) \quad \text{if} \quad t_m \le t \le t_{m+1}.$$

Now apply the first inequality in (5.3) to establish (7.6) and therefore finish the proof of Lemma 7.2. □

Proof of Theorem 6. To conclude formula (3.6) from Lemma 7.2, first observe that for the p-norm with $1 \le p < \infty$ we have from (7.4) that for $(u, v) \in (-1, 1) \times [0, 1)$ the value of t that makes the rows of $\Phi^{-1}(u, v)$ have the same norm F_t is given by

$$t = \Big(\frac{1 - v^p}{1 - |u|^p}\Big)^{\frac{1}{2p}}.$$

The corresponding value of $D_{F^{\langle p\rangle}}(u, v)$ from (7.5) is

$$\begin{aligned} D_{F^{\langle p\rangle}}(u, v) &= (1 + uv)^{-1}\Big(\big(\tfrac{1-v^p}{1-|u|^p}\big)^{-\frac{1}{2}} + |u|^p\big(\tfrac{1-v^p}{1-|u|^p}\big)^{\frac{1}{2}}\Big)^{\frac{2}{p}} \\ &= (1 + uv)^{-1}\Big(\tfrac{(1-|u|^p v^p)^2}{(1-|u|^p)(1-v^p)}\Big)^{\frac{1}{p}} = D_p(u, v), \end{aligned}$$

giving (3.6). The case $p = \infty$ is immediate. This completes the proof of Theorem 6. □

8. $\mathcal{S}$-EXPANSIONS

To prove Theorem 7 we want to characterize in terms of the norm those convergents of the regular continued fraction of an irrational α that are also convergents of the continued fraction of α associated to a strongly symmetric norm F. We will use the notation and results of Lemma 5.2. Write $P_m = (q_m, p_m - \alpha q_m)$ for points coming from this norm with corresponding t_m and let $Q_n = (s_n, r_n - \alpha s_n)$ be the points coming from the convergents of the regular continued fraction of α. Furthermore, the partial quotient b_n is associated to Q_n while a_m is associated to P_m.

Lemma 8.1. *For a fixed $n \ge 0$ there are integers c_ℓ and d_ℓ with $c_\ell > 0$ and $d_\ell \ge 0$ so that for each $\ell \ge 0$*

$$Q_{n+\ell} = c_\ell Q_n + d_\ell Q_{n-1}$$

where $c_\ell \ge d_\ell$ for all $\ell \ge 0$, while for $\ell \ge 2$ we have

$$c_\ell \ge d_\ell + 1.$$

Proof. The integers r_n, s_n are determined recursively for $n \ge 0$ by

$$r_n = b_n r_{n-1} + r_{n-2}, \qquad r_{-2} = 0, \quad r_{-1} = 1, \tag{8.1}$$

$$s_n = b_n s_{n-1} + s_{n-2}, \qquad s_{-2} = 1, \quad s_{-1} = 0. \tag{8.2}$$

It is easy to check using (8.1) and (8.2) that c_ℓ and d_ℓ satisfy for fixed n and $\ell \ge 1$ the recurrence relations

$$c_{\ell+1} = b_{n+\ell+1} c_\ell + c_{\ell-1}, \qquad c_1 = b_{n+1}, \quad c_0 = 1, \tag{8.3}$$

$$d_{\ell+1} = b_{n+\ell+1} d_\ell + d_{\ell-1}, \qquad d_1 = 1, \qquad d_0 = 0. \tag{8.4}$$

The claim of the lemma follows from a straightforward inductive argument. □

The following result will be used to characterize those convergents of the regular continued fraction that occur as convergents in the continued fraction associated to the norm.

Lemma 8.2. *For $m \ge 1$ let n and ℓ be such that $P_{m-1} = Q_{n-1}$ and $P_m = Q_{n+\ell}$. Then*

(i) $\ell \in \{0, 1\}$.

(ii) There is a unique $t \ge 1$ such that $F_t(Q_n) = F_t(Q_{n-1})$, and $\ell = 1$ if and only if

$$F_t(Q_n + Q_{n-1}) \le F_t(Q_n).$$

If this holds we have that $b_{n+1} = 1$.

(iii) $Q_0 = P_0$ if and only if $a_0 = b_0$.

Proof. We have $P_{-1} = Q_{-1}$ and by Lemma 6.1 we know that for each $m \geq 1$ we have $P_{m-1} = Q_{n-1}$ for some n and $P_m = Q_{n+\ell}$ for some $\ell \geq 0$. We can check directly that $Q_0 = P_0$ if and only if $a_0 = b_0$.

By Lemma 8.1 for $\ell \geq 0$ we have

$$c_\ell Q_n = P_m - d_\ell P_{m-1}. \tag{8.5}$$

By Lemma 5.2 we have $F_{t_m}(Q_n) \geq F_{t_m}(P_m) = F_{t_m}(P_{m-1})$ and hence

$$c_\ell F_{t_m}(P_m) \leq F_{t_m}(P_m - d_\ell P_{m-1}) < F_{t_m}(P_m) + d_\ell F_{t_m}(P_{m-1}) \tag{8.6}$$

by Lemma A.5. Thus we have

$$c_\ell < d_\ell + 1. \tag{8.7}$$

so by Lemma 8.1 we have that either $\ell = 0$ or $\ell = 1$.

Now by Lemma A.4 applied to the norm F_{t_m} and using that

$$F_{t_m}(Q_n) \geq F_{t_m}(Q_{n-1}),$$

there is a $t \geq 1$ (indeed $t \geq t_m$) so that

$$F_t(Q_n) = F_t(Q_{n-1}).$$

In case $\ell = 1$ we have $b_{n+1} = 1$ by (8.7) and (8.3)–(8.4). By (8.5) we have that

$$Q_n + Q_{n-1} = P_m = Q_{n+1},$$

so we must have

$$F_t(Q_n + Q_{n-1}) = F_t(Q_{n+1}) = F_t(P_m) \leq F_t(Q_n),$$

where the inequality follows from Lemma 5.2(iii).

If $\ell = 0$ we have $Q_{n-1} = P_{m-1}$ and $Q_n = P_m$ so that $t = t_m$ and

$$F_t(Q_n + Q_{n-1}) = F_{t_m}(P_m + P_{m-1}) > F_{t_m}(P_m) = F_t(Q_n),$$

at least when $m > 0$, since then the x-coordinate of $P_m + P_{m-1}$ is strictly larger than that of P_m and so by Lemma 5.2(iii) strict inequality must hold. □

Proof of Theorem 7. Lemma 8.2 gives instructions for obtaining the sequence of convergents p_m/q_m of α associated to the norm F from the sequence of regular convergents r_n/s_n of α, namely

$$\text{omit the regular convergent } \frac{r_n}{s_n} \ (n \geq 1) \iff F_t(Q_n + Q_{n-1}) \leq F_t(Q_n), \tag{8.8}$$

where $t \geq 1$ is such that $F_t(Q_n) = F_t(Q_{n-1})$, and

$$\text{omit } \frac{r_0}{s_0} \iff \lfloor \alpha \rfloor \text{ is not the nearest integer to } \alpha \iff \alpha \in [\tfrac{1}{2}, 1) + \mathbb{Z}. \tag{8.9}$$

We must define a singularization area that encodes both of these instructions. Let $\mathcal{D}$ and Φ be as in Section 7. For each $g = \left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right) \in \mathcal{D}$ we write

$$P = (x, y) \quad \text{and} \quad P' = (x', y'),$$

and we define

$$\mathcal{S} = \Phi(\{g \in \mathcal{D};\, F(P + P') \leq F(P)\}) \cup \left([\tfrac{1}{2}, 1) \times \{0\}\right). \tag{8.10}$$

Note that $\mathcal{S}$ is a closed set in the induced topology on $[\frac{1}{2}, 1) \times [0, 1]$. The portion of $\mathcal{S}$ that lies on the u-axis encodes the rule (8.9). Suppose that $n \geq 1$ and let $\left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right) = \Phi^{-1}(u_n, v_n)$,

where u_n and v_n are defined in (1.2). Then Lemmas 7.1 and 6.2 imply that $Q_n = (tx, t^{-1}y)$ and $Q_{n-1} = (tx', t^{-1}y')$, with t defined by $F_t(Q_n) = F_t(Q_{n-1})$. So the condition on the right-hand side of (8.8) is equivalent to $F(P + P') \le F(P)$. It follows that (8.8)–(8.9) are encoded by the rule

$$\text{omit the regular convergent } \frac{r_n}{s_n} \iff (u_n, v_n) \in \mathcal{S}. \tag{8.11}$$

It is helpful to have some more concrete information about the set $\mathcal{S}$. For a generic norm it is difficult to describe $\mathcal{S}$ explicitly, so we will relate $\mathcal{S}$ to the set $\mathcal{S}_1$, which is easy to describe. As usual, we denote $\mathcal{S}$ by $\mathcal{S}_p$ when F is the p-norm. By (8.10) we have

$$\mathcal{S}_1 = \Phi\big(\{\left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right) \in \mathcal{D};\ |x + x'| + |y + y'| \le |x| + |y|\}\big) \cup \big([\tfrac{1}{2}, 1) \times \{0\}\big).$$

Using (7.4) we can rewrite this as

$$\mathcal{S}_1 = \big\{(u, v) \in (-1, 1) \times [0, 1);\ t^{-1}(1 + v) + t(1 - u) \le t^{-1} + t|u|\big\} \cup \big([\tfrac{1}{2}, 1) \times \{0\}\big), \tag{8.12}$$

where $t^2 = \frac{1-v}{1-|u|}$. Some algebraic manipulation reduces the inequality in (8.12) to

$$1 + uv \le u + |u|.$$

This, together with $|uv| < 1$, implies that $u \ge \frac{1}{2}$, so we find that

$$\mathcal{S}_1 = \big\{(u, v) \in [\tfrac{1}{2}, 1) \times [0, 1);\ v \le 2 - \tfrac{1}{u}\big\}. \tag{8.13}$$

The interior of the set (8.13) agrees with the S-region given in [27] for Minkowski's diagonal continued fraction.

Lemma 8.3. *For any strongly symmetric norm F we have $\mathcal{S} \subseteq \mathcal{S}_1$.*

Proof. Since $\mathcal{S}$ and $\mathcal{S}_1$ are closed sets in the induced topology on $[\frac{1}{2}, 1) \times [0, 1]$, it suffices to show that a dense subset of $\mathcal{S}$ is contained in $\mathcal{S}_1$. Suppose that $(u, v) \in \mathcal{S}$ with $u \notin \mathbb{Q}$ and $v \in \mathbb{Q}$, and write

$$u = \frac{1}{b_{n+1}+}\ \frac{1}{b_{n+2}+} \cdots \quad \text{and} \quad v = \frac{1}{b_n+}\ \frac{1}{b_{n-1}+}\ \frac{1}{b_{n-2}+} \cdots \frac{1}{b_1} \tag{8.14}$$

for the regular continued fractions of u and v. If we define

$$\alpha = \frac{1}{b_1+}\ \frac{1}{b_2+}\ \frac{1}{b_3+} \cdots$$

then $(u, v) = (u_n, v_n)$ for α. Since $(u, v) \in \mathcal{S}$ we have $Q_{n+1} = Q_{n-1} + Q_n$ in the notation of Section 8, and for some m we have $P_{m-1} = Q_{n-1}$ and $P_m = Q_{n+1}$. Thus

$$F_{t_m}(Q_{n-1}) = F_{t_m}(Q_{n+1}) \le F_{t_m}(Q_n).$$

Since the regular convergents of α alternate between lying to the left of α and lying to the right of α, the points Q_{n-1} and Q_{n+1} are in the same quadrant. Let t be such that $F_t^{\langle 1\rangle}(P_{m-1}) = F_t^{\langle 1\rangle}(P_m)$, where $F^{\langle 1\rangle}$ denotes the 1-norm. By convexity of F, the closed stretched ball $\overline{\mathcal{B}(P_m, t_m)}$ contains the line segment connecting the points P_{m-1} and P_m. This line segment comprises all of the points P in the same quadrant as Q_{n-1}, Q_{n+1} with x-coordinate between x_{m-1} and x_m, and with $F_t^{\langle 1\rangle}(P) = F_t^{\langle 1\rangle}(P_m)$. Since the x-coordinate of Q_n is between x_{m-1} and x_m and Q_n is outside the ball $\mathcal{B}(P_m, t_m)$, we have

$$F_t^{\langle 1\rangle}(Q_n) \ge F_t^{\langle 1\rangle}(P_m) = F_t^{\langle 1\rangle}(Q_n + Q_{n-1}).$$

By (8.8) and (8.11) it follows that $(u, v) \in \mathcal{S}_1$. □

Lemma 8.3, together with the explicit description (8.13), shows that $\mathcal{S} \subseteq [\frac{1}{2}, 1) \times [0, 1]$. We also have that $T\mathcal{S} \cap \mathcal{S} = \emptyset$, where T is given in (2.9), since $\mathcal{S} \subseteq \mathcal{S}_1$ and

$$T\mathcal{S}_1 = \big\{(u, v) \in [0, 1) \times [\tfrac{1}{2}, 1); u \leq 2 - \tfrac{1}{v}\big\}. \tag{8.15}$$

It follows that $\mathcal{S}$ is a singularization area as defined above Theorem 7. This fact and (8.11) together prove Theorem 7. □

We also immediately obtain the following lemma, which we will use several times in the coming sections.

Lemma 8.4. *For every strongly symmetric norm F, there is a neighborhood $\mathcal{U}$ of the line segment $u = v$ with $u, v \in (0, 1)$ such that $\mathcal{U} \cap ((-1, 1) \times [0, 1))$ does not intersect $\mathcal{S}$.*

We finish this section with a quick proof of our claim (2.8) that

$$\delta_2\Big(\tfrac{1}{2}(-1 + \sqrt{3})\Big) = 1.$$

The regular continued fraction expansion of $\alpha = \frac{1}{2}(-1 + \sqrt{3})$ is

$$\alpha = \frac{1}{2+}\,\frac{1}{1+}\,\frac{1}{2+}\,\frac{1}{1+}\cdots, \tag{8.16}$$

from which it follows that

$$u_n = \begin{cases} \alpha & \text{if } n \text{ is even,} \\ 2\alpha & \text{if } n \text{ is odd,} \end{cases}$$

while $v_{2n} \to 2\alpha$ from below and $v_{2n+1} \to \alpha$ from above. By an argument similar to the one used to establish (8.13), we find that the region $\mathcal{S}_2$ comprises those points (u, v) for which $u(2 + v) > 1 + 2v$. The points (u_n, v_n) are all outside $\mathcal{S}_2$, so the 2-continued fraction expansion of α is the same as the regular continued fraction and thus $(\mu_n, \nu_n) = (u_n, v_n)$. Since $D_2(u, v) = D_2(v, u)$, we have

$$\delta_2(\alpha) = \Delta_2 \max_{k\in\{0,1\}} \lim_{n\to\infty} D_2(\mu_{2n+k}, \nu_{2n+k}) = \Delta_2 D_2(\alpha, 2\alpha) = 1,$$

where the last equality uses (3.6) and the fact that $\Delta_2 = \frac{\sqrt{3}}{2}$.

9. Values of $\delta_F(\alpha)$ for well approximable numbers

We now prove Theorem 3, which gives the smallest value of $\delta_F(\alpha)$ for F any strongly symmetric norm and α well approximable.

Lemma 9.1. *Suppose that α is well approximable. Then $\delta_F(\alpha) \geq \Delta$.*

Proof. By definition, for any $\varepsilon > 0$ there are arbitrarily large $q > 0$ so that for some $p \in \mathbb{Z}$

$$|\tfrac{p}{q} - \alpha| < \tfrac{\varepsilon}{q^2}.$$

For such a q let $t = q$ and note that for any $r, s \in \mathbb{Z}$ with $s > 0$

$$\begin{aligned} F_t(s, r - \alpha s) &= F(t^{-1}s, t(s\alpha - r)) = F(\tfrac{s}{q}, q(s\alpha - r)) \\ &= F(\tfrac{s}{q}, q(s\tfrac{p}{q} - r + \tfrac{\sigma s}{q^2})) = F(\tfrac{s}{q}, sp - rq + \tfrac{\sigma s}{q}) \end{aligned}$$

for some σ with $|\sigma| \leq \varepsilon$. By Lemma A.2 if $s \geq q$ we have

$$F(\tfrac{s}{q}, sp - rq + \tfrac{\sigma s}{q}) \geq F(1, 0) = 1,$$

while for $0 < s < q$ we have $F(\frac{s}{q}, sp - rq + \frac{\sigma s}{q}) \geq F(0, 1-\varepsilon)$, since $q \nmid s$. By the continuity of F, for any $\varepsilon' > 0$ there is an $\varepsilon > 0$ so that $F(0, 1-\varepsilon) \geq 1 - \varepsilon'$. It follows that $\limsup_{t\to\infty} F_t(s, r - \alpha s) \geq 1$ and hence that $\delta_F(\alpha) \geq \Delta$. □

To finish the proof of Theorem 3, we need to find well approximable α for which $\delta_F(\alpha) = \Delta$.

Lemma 9.2. *Suppose that the partial quotients b_n of the regular continued fraction expansion of α are eventually strictly increasing with n. Then*

$$\delta_F(\alpha) = \Delta.$$

Proof. If the regular partial quotients b_n of α are eventually strictly increasing, then for any $\varepsilon > 0$ the points (u_n, v_n) all eventually lie within ε of the point $(0, 0)$. So by Lemma 8.4, the points (u_n, v_n) are outside $\mathcal{S}$ for sufficiently large n. Thus

$$\lim_{n\to\infty} (\mu_n, \nu_n) = \lim_{n\to\infty} (u_n, v_n) = (0, 0).$$

Finally, $D_F(0, 0) = F^2(\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)) = 1$, therefore $\delta_F(\alpha) = \Delta$. □

10. Values of $\delta_p(\alpha)$ for any irrational α

Proof of Theorem 4. Fix $p \in [1, \infty]$. Throughout the proof let

$$\alpha = \frac{\varepsilon_1}{a_1+} \frac{\varepsilon_2}{a_2+} \frac{\varepsilon_3}{a_3+} \dots \tag{10.1}$$

denote the p-continued fraction expansion of $\alpha \in (0, 1)$ and define μ_m and ν_m as in (3.5). By Theorem 6 it suffices to show that for every α we have

$$\limsup_{m\to\infty} D_p(\mu_m, \nu_m) \geq \begin{cases} 1 & \text{if } 1 \leq p \leq 2, \\ \frac{1}{10}\big(\sqrt{5}+5\big)\Big(\big(\frac{1}{2}(\sqrt{5}-1)\big)^p + 1\Big)^{2/p} & \text{if } p > 2, \end{cases} \tag{10.2}$$

and that there is at least one α for which equality holds. In both cases the number on the right-hand side of (10.2) is ≤ 1. Since $(1 - |u|^p v^p)^2 \geq (1 - |u|^p)(1 - v^p)$, we have

$$D_p(u, v) \geq \frac{1}{1 + uv}. \tag{10.3}$$

It follows that $D_p(u, v) \geq 1$ for nonpositive u, so if $\mu_m \leq 0$ for infinitely many m, the inequality (10.2) holds trivially. Thus we may assume that the p-continued fraction expansion of α has $\mu_m \geq 0$ for all sufficiently large m.

Lemma 10.1. *If $0 \leq u, v < 1$ then*

$$D_p(u, v) \geq D_p\Big(\frac{u+v}{2}, \frac{u+v}{2}\Big),$$

with equality only when $u = v$.

Proof. The inequalities

$$1 - (uv)^p \geq 1 - \left(\frac{u+v}{2}\right)^{2p} \quad \text{and} \quad 1 + uv \leq 1 + \left(\frac{u+v}{2}\right)^2 \tag{10.4}$$

both reduce to $(u - v)^2 \geq 0$. It remains to show that

$$(1 - u^p)(1 - v^p) \leq \left(1 - \left(\frac{u+v}{2}\right)^p\right)^2.$$

This inequality is implied by the first inequality of (10.4) and

$$u^p + v^p \geq 2\Big(\frac{u+v}{2}\Big)^p,$$

which follows immediately from Hölder's inequality. □

It is convenient to define

$$d_p(x) = D_p(x,x) = \frac{(1+x^p)^{2/p}}{1+x^2}.$$

Then

$$\frac{\partial}{\partial x}[d_p(x)]^p = \frac{2p(1+x^p)(x^p-x^2)}{x(x^2+1)^{p+1}}.$$

If $1 \leq p < 2$ then $d_p(x)$ is strictly increasing, so by Lemma 10.1 we have

$$\min_{u,v\in[0,1]} D_p(u,v) = \min_{x\in[0,1]} d_p(x) = d_p(0) = 1. \tag{10.5}$$

If $p = 2$ then $d_p(x) = 1$ for all x. In either case, we can use Lemma 9.2 to find examples of α for which $\delta_p(\alpha) = \Delta_p$.

Suppose that $p > 2$, and let $\beta = \frac{1}{2}(\sqrt{5}-1)$. The sequence (u_n, v_n) associated to the regular continued fraction

$$\beta = \frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\cdots$$

approaches (β,β) as $n \to \infty$. By Lemma 8.4 it follows that the sequence (μ_m, ν_m) associated to the p-continued fraction of β also converges to (β,β). Thus, for $p > 2$ we have $\delta_p(\beta) = \Delta_p D_p(\beta,\beta)$, which is the number in (2.7).

It remains to show that for every α with $\mu_m \geq 0$ for sufficiently large m, we have $D_p(\mu_m,\nu_m) \geq D_p(\beta,\beta)$ for infinitely many m. Since $p > 2$, the function $d_p(x)$ is strictly decreasing, so by Lemma 10.1 it suffices to show that

$$\mu_m + \nu_m \leq \sqrt{5} - 1 = 1.23606\ldots$$

for infinitely many m.

If there are infinitely many m such that $a_{m+1} \geq 5$ then for such m we have $\mu_m \leq \frac{1}{5}$ and therefore $\mu_m + \nu_m \leq 1.2$. So we may suppose that $a_m \leq 4$ for all sufficiently large m. The following lemma covers the remaining cases.

Lemma 10.2. *Let $\ell \in \{2,3,4\}$ and suppose that $\varepsilon_m = 1$ and $a_m \leq \ell$ for sufficiently large m. If $a_m = \ell$ for infinitely many m, then*

$$\mu_m + \nu_m < 1.18 \tag{10.6}$$

for infinitely many m.

Proof. Suppose that $\varepsilon_m = 1$ and $a_m \leq \ell$ for $m \geq M$. Recall that increasing (resp. decreasing) the partial quotients of a continued fraction in even (resp. odd) indices increases the resulting number. Thus for any $m \geq M+3$ with $a_{m+1} = \ell$ we have

$$\mu_m \leq \frac{1}{\ell+}\,\frac{1}{\ell+}\,\frac{1}{1+}\,\frac{1}{\ell+}\,\frac{1}{1+}\,\frac{1}{\ell+}\cdots,$$
$$\nu_m \leq \frac{1}{1+}\,\frac{1}{\ell+1}.$$

The lemma now follows from an easy computation. □

This completes the proof of Theorem 4. □

We remark that it is sometimes possible to compute the minimum value of $\delta_F(\alpha)$ for other norms as well. The composition of strongly symmetric norms is, up to scaling, also strongly symmetric (see Lemma A.3). For instance, the norms F^{oct_1} and F^{oct_2} with regular octagonal unit balls mentioned in §2 can be given in terms of compositions of the 1-norm and the sup-norm. Explicitly,

$$F^{\mathrm{oct}_1}(P) = F^{\langle\infty\rangle}(Q) \quad \text{and} \quad F^{\mathrm{oct}_2}(P) = (2-\sqrt{2})F^{\langle 1\rangle}(Q) \tag{10.7}$$

where $Q = (\frac{1}{\sqrt{2}}F^{\langle 1\rangle}(P), F^{\langle\infty\rangle}(P))$. Some effort involving Mahler's computation of the critical determinant of the regular octagon recalled at the end of §4 and Lemma 8.4, leads to the following results. The minimum of $\delta_F(\alpha)$ for $F = F^{\mathrm{oct}_1}$ is $\frac{1}{8}\big(3\sqrt{2}+2\big)$, which is attained when

$$\alpha = \sqrt{2}-1 = \frac{1}{2+}\,\frac{1}{2+}\,\frac{1}{2+}\cdots.$$

For $F = F^{\mathrm{oct}_2}$ the minimal value is also $\frac{1}{8}\big(3\sqrt{2}+2\big)$, but now this is the value of Δ and is attained when $\alpha = \frac{e-1}{e+1}$, for instance.

11. The dynamical system

The goal of this section is to prove Theorem 1. We employ the notation of Section 8. Say that $g = \left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right) \in G$ is *reduced with respect to the norm* F if $g \in \mathcal{D}$ and

$$\big(F(P)\overline{\mathcal{B}}\big) \cap L(g) = \{0, \pm P, \pm P'\}, \tag{11.1}$$

where $\mathcal{B}$ is the open unit ball for F, the overline denotes the closure, and $L(g)$ was defined in (7.1). Let $\mathcal{R}$ be the set of all g that are reduced with respect to F and define $\Omega \subset (-1,1)\times[0,1]$ as

$$\Omega \stackrel{\mathrm{def}}{=} \Phi(\mathcal{R}) \cup \mathcal{A}, \tag{11.2}$$

where

$$\mathcal{A} = \overline{\Phi(\mathcal{R})} \cap \big((-\tfrac{1}{2},\tfrac{1}{2})\times\{0\}\big).$$

We will show that $(\mu_n,\nu_n) \in \Omega$ for all $n \geq 0$.

We want to apply the ergodic theory of $\mathcal{S}$-expansions as developed in [26]. For that we need to show that Ω defined by (11.2) coincides with the set

$$\Omega_{\mathcal{S}} = ([0,1)\times[0,1] \setminus (\mathcal{S} \cup T\mathcal{S})) \cup (M\circ T)\mathcal{S}$$

defined in Section 5 of [26], where

$$M(u,v) = \Big(\frac{-u}{1+u}, 1-v\Big), \quad (u,v) \in T\mathcal{S}.$$

The following equivalent description of reduced matrices is helpful.

Lemma 11.1. *A matrix $g \in \mathcal{D}$ is reduced if and only if*

$$\min\big(F(P+P'), F(P-P')\big) > F(P). \tag{11.3}$$

Proof. Clearly a matrix g satisfying (11.1) also satisfies (11.3). Suppose $g \in \mathcal{D}$ satisfies (11.3). We will show that $F(aP + bP') > F(P)$ for all $(a, b) \in \mathbb{Z}^2 \setminus \{(0, 0), (0, \pm 1), (\pm 1, 0)\}$. If $|a| = |b|$ then

$$F(aP + bP') = |a|F(P \pm P') > F(P).$$

Otherwise, if $|a| > |b|$, say, then by the reverse triangle inequality

$$F(aP + bP') \geq |a|F(P) - |b|F(P') = (|a| - |b|)F(P).$$

This is strictly greater than $F(P)$ if $|a| - |b| \geq 2$. If $|a| = |b| + 1$ then

$$F(aP + bP') = F(a(P \pm P') \pm P') \geq |a|F(P + P') - F(P') > (|a| - 1)F(P).$$

This completes the proof since $|b| \geq 1$ so $|a| \geq 2$. □

It follows that the set $(-1, 1) \times [0, 1]$ decomposes as $\Omega \sqcup \mathcal{S} \sqcup \mathcal{S}' \sqcup \mathcal{S}''$, where

$$\mathcal{S}' = \Phi(\{g \in \mathcal{D} : F(P - P') \leq F(P) \text{ and } y < 0\}),$$
$$\mathcal{S}'' = \Phi(\{g \in \mathcal{D} : F(P - P') \leq F(P) \text{ and } y \geq 0\}) \cup \big(((-1, \tfrac{1}{2}) \times \{0\}) \setminus \mathcal{A}\big).$$

See Figure 6 for the case $p = 2$.

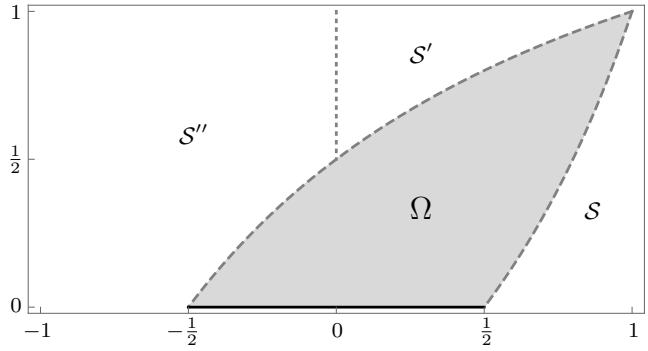

FIGURE 6. The sets Ω, $\mathcal{S}$, $\mathcal{S}'$, and $\mathcal{S}''$ for $p = 2$.

Since the critical lattices for F (see §4) are among those for which two basis vectors and their sum all have equal norm, we will refer to the set

$$\big\{g \in \mathcal{D}; \min\big(F(P + P'), F(P - P')\big) = F(P)\big\} \tag{11.4}$$

as the potentially critical matrices. The next lemma describes the boundary of Ω in terms of the distinguished subset

$$\mathcal{P} = \{g \in \mathcal{D} : F(P) = F(P + P')\}$$

of the potentially critical matrices. The following shows that $\mathcal{P}$ is a subset of the potentially critical matrices.

Lemma 11.2. *If $F(P) = F(P') = F(P + P')$ then $F(P - P') \geq F(P)$.*

Proof. Without loss of generality assume that P, P', and $P + P'$ all lie on the boundary of $\mathcal{B}$. Let L, L' denote the parallel lines

$$L = \gamma P, \qquad L' = \gamma P - P',$$

where $\gamma \in \mathbb{R}$. Note that $-P, 0, P$ are points in $\overline{\mathcal{B}}$ on L, while $-P' - P$ and $-P'$ are points on the boundary of $\mathcal{B}$ on L'. We argue by contradiction. Suppose that $F(P - P') < F(P)$

so that $Q = P - P'$ is in the interior of $\mathcal{B}$. Then we can move Q slightly so that it remains inside $\mathcal{B}$ but the midpoint of the line segment between $-P - P'$ and Q is outside $\overline{\mathcal{B}}$. This contradicts the convexity of $\mathcal{B}$. □

Lemma 11.3. *The part of the boundary of* Ω *that lies in* $(-1,1)\times[0,1]$ *is* $\partial \cup \partial' \cup \partial'' \cup \mathcal{A}$, *where*

$$\begin{aligned}\partial &= \Phi(\mathcal{P}),\\ \partial' &= \Big\{\Phi\Big(\begin{pmatrix} x+x' & y+y' \\ x & y\end{pmatrix}\Big);\ g\in\mathcal{P}\Big\},\\ \partial'' &= \Big\{\Phi\Big(\begin{pmatrix} x+x' & y+y' \\ x' & y'\end{pmatrix}\Big);\ g\in\mathcal{P}\Big\}.\end{aligned}$$

Proof. The boundary of Ω is $\mathcal{A}\cup\mathcal{C}$, where

$$\mathcal{C} = \Phi\big(\{g\in\mathcal{D};\ x'>0 \text{ and } \min\big(F(P+P'),F(P-P')\big) = F(P)\}\big).$$

Lemma 11.2 implies that ∂ is the part of $\mathcal{C}$ adjacent to $\mathcal{S}$. The remaining set, $\mathcal{C}\setminus\partial$, is the image of the set of $g'\in\mathcal{D}$ satisfying

$$F(Q) = F(Q-Q') < F(Q+Q'),\ \text{where } g' = \begin{pmatrix} Q \\ Q'\end{pmatrix}.$$

If the y-coordinate of Q is negative, then $(P,P') = (Q', Q-Q')$ gives an element of $\mathcal{P}$, and

$$\Phi(g') = \Phi\Big(\begin{pmatrix} x+x' & y+y' \\ x & y\end{pmatrix}\Big).$$

Otherwise $(P,P') = (Q-Q', Q')$ yields an element of $\mathcal{P}$; in this case

$$\Phi(g') = \Phi\Big(\begin{pmatrix} x+x' & y+y' \\ x' & y'\end{pmatrix}\Big).$$

This completes the proof. □

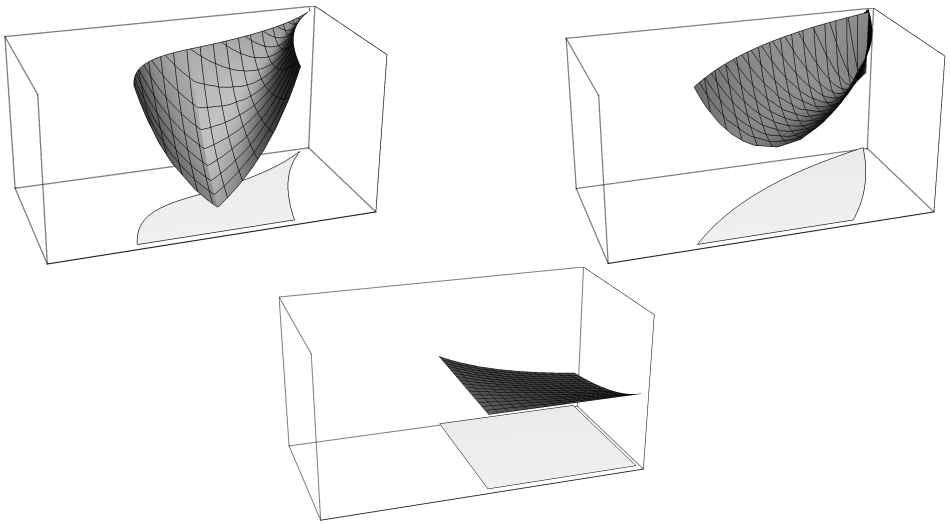

FIGURE 7. The functions $D_p(u,v)$ over Ω_p for $p = 1, 2, \infty$.

Lemma 11.4. *The function* $D(u,v)$ *is continuous on* $\overline{\Omega}\setminus\{(1,1)\}$ *and assumes its maximum value* $1/\Delta$ *on that set.*

Proof. The continuity statement is clear from the definition of $D(u,v)$.

For $(u,v) \in \overline{\Omega}\backslash\{(1,1)\}$, let the points $P=(x,y)$ and $P'=(x',y')$ be such that $\Phi^{-1}(u,v) = \left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right)$. Then

$$D(u,v) = F\big(\Phi^{-1}(u,v)\big)^2 = F(P)^2 = (\det L)^{-1} \leq \tfrac{1}{\Delta},$$

where L is the lattice generated by the unit vectors $\frac{1}{F(P)}P$ and $\frac{1}{F(P')}P'$. Say that a point (u,v) in $\overline{\Omega}\setminus\{(1,1)\}$ is a critical point if $D(u,v) = 1/\Delta$ so that (u,v) maximizes $D(u,v)$. Then (u,v) is a critical point if and only if L is a critical lattice that does not correspond to $(1,1)$, so by Lemma 11.3 all critical points lie on the boundary of Ω. If critical points in $\overline{\Omega}\setminus\{(1,1)\}$ exist, then we are done.

Suppose that a critical lattice L corresponds to the point $(1,1)$ in the uv-plane. Then there exists a t such that

$$F_t(P) = F_t(P') = F_t(P+P'), \quad \text{where } P = \tfrac{1}{\sqrt{2}}(1,-1), P' = \tfrac{1}{\sqrt{2}}(1,1).$$

Then the matrix $g = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 2t^{-1} & 0 \\ t^{-1} & -t \end{smallmatrix}\right) \in \overline{\mathcal{R}}$ is critical and satisfies $\Phi(g) = (0, \frac{1}{2})$. So if $(1,1)$ corresponds to a critical lattice L, then the point $(0,\frac{1}{2}) \in \overline{\Omega}$ is a critical point. □

That $\Omega = \Omega_{\mathcal{S}}$ follows from the next lemma.

Lemma 11.5. *We have*

$$\mathcal{S}' = T\mathcal{S},$$
$$\mathcal{S}'' = ((-1,0]\times[0,1]) \setminus M\mathcal{S}'.$$

Proof. We begin by showing that $\mathcal{S}' = T\mathcal{S}$. For $(u,v) \in \mathcal{S}$ we have $\frac{1}{2} < u < 1$ or $(u,v) = (\frac{1}{2}, 0)$, so (2.9) simplifies to $T(u,v) = \left(\frac{1-u}{u}, \frac{1}{v+1}\right)$ and $T(\frac{1}{2},0) = (0,\frac{1}{2})$. We show that the boundary of $\mathcal{S}$ maps to the boundary of $\mathcal{S}'$ under T; the lemma then follows by the continuity of T on $\mathcal{S}\setminus\{(\frac{1}{2},0)\}$. Since $T([\frac{1}{2},1]\times\{0\}) = ([0,1]\times\{1\}) \cup \{(0,\frac{1}{2})\}$ and $T(\{1\}\times[0,1)) = \{0\}\times(\frac{1}{2},1]$, it suffices to show that $T(\partial) = \partial'$. Suppose that $(u,v) \in \partial$ and that $\Phi(g) = (u,v)$. Then

$$T\circ\Phi\left(\left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right)\right) = \left(\tfrac{1-u}{u}, \tfrac{1}{v+1}\right) = \left(\tfrac{-(y+y')}{y}, \tfrac{x}{x'+x}\right) = \Phi\left(\left(\begin{smallmatrix} x+x' & y+y' \\ x & y \end{smallmatrix}\right)\right).$$

Since this is clearly invertible, we conclude that $T(\partial) = \partial'$.

We prove $\mathcal{S}'' = ((-1,0]\times[0,1]) \setminus M\mathcal{S}'$ similarly. We have $M([0,1]\times\{1\}) = [-\frac{1}{2},0]\times\{0\}$ and $M(\{0\}\times[\frac{1}{2},1]) = \{0\}\times[0,\frac{1}{2}]$ for the straight line segments, so it suffices to show that $M(\partial') = \partial''$. We will show that $(M\circ T)\partial = \partial''$, using that

$$(M\circ T)(u,v) = \left(u-1, \tfrac{v}{v+1}\right).$$

Suppose that $(u,v) \in \partial$ and that $\Phi(g) = (u,v)$. Then

$$M\circ T\circ\Phi\left(\left(\begin{smallmatrix} x & y \\ x' & y' \end{smallmatrix}\right)\right) = \left(u-1, \tfrac{v}{v+1}\right) = \left(-\tfrac{(y+y')}{y'}, \tfrac{x'}{x+x'}\right) = \Phi\left(\left(\begin{smallmatrix} x+x' & y+y' \\ x' & y' \end{smallmatrix}\right)\right),$$

which completes the proof. □

Let $\omega_{\mathcal{S}} = (1-\omega(\mathcal{S}))^{-1}\omega$, where ω was defined in (2.10), and the scaling makes $\omega_{\mathcal{S}}$ a probability measure on Ω.

Lemma 11.6. *Define $\mathcal{S}$ and Ω by* (8.10) *and* (11.2)*. Then for almost all irrational α the sequence (μ_m, ν_m) is uniformly distributed over Ω with respect to the measure $\omega_{\mathcal{S}}$.*

Proof. Since $\Omega = \Omega_{\mathcal{S}}$, Lemma 11.6 follows from Theorem 5.4.23 of [9] (see also [26]) and Theorem 7. □

Proof of Theorem 1. By Lemma 11.4, the function $\Delta D(u, v)$ assumes the value 1 at some point in $\overline{\Omega} \setminus \{(1, 1)\}$. By Lemma 7.2 it follows that $\delta_F(\alpha) = 1$ if and only if the sequence (μ_n, ν_n) is infinitely often arbitrarily close to such a critical point. It follows from Lemma 11.6 that $\delta_F(\alpha) = 1$ for almost all α.

To finish the proof it suffices to show that there are uncountably many α for which Minkowski's Approximation Theorem can be improved. Since we already know Theorem 1 is true in the case of the sup-norm, suppose that F is not the sup-norm. Then the lattice generated by $(1, 0)$ and $(0, 1)$ is not potentially critical, so we have $\Delta D(0, 0) < 1$. Thus any α for which (μ_n, ν_n) converges to $(0, 0)$ has $\delta_F(\alpha) < 1$, and there are uncountably many such α (for example, the set of α with strictly increasing partial quotients). □

12. Concluding remarks

In addition to the proof of Theorem 1, there are other applications of the metric theory of $\mathcal{S}$-expansions and ergodic theory to quantities related to $\delta_p(\alpha)$. For instance we may treat the distribution of the values of

$$\delta_p(\alpha; m) \stackrel{\text{def}}{=} \Delta_p \, D_p(\mu_m, \nu_m)$$

from Theorem 6. For almost all α the distribution function

$$\lim_{M \to \infty} \tfrac{1}{M} \#\{1 \le m \le M;\ \delta_p(\alpha; m) \le z\}$$

exists for all $z \in [0, 1]$. For $p = 1, 2, \infty$ it can be evaluated explicitly, as was done for $p = \infty$ in Theorem 4 of [5]. In particular, for almost all α

$$\lim_{M \to \infty} \frac{1}{M} \sum_{1 \le m \le M} \delta_p(\alpha; m) = c_p$$

exists where

$$c_1 = \tfrac{1}{2}(3 - \log 4) = 0.806853\ldots, \quad c_2 = \tfrac{1}{\log 3} = 0.910239\ldots, \quad c_\infty = \tfrac{1+\log 4}{\log 16} = 0.860674\ldots.$$

It is well known that a close connection exists between dynamical systems associated to various kinds of continued fractions and the geodesic flow on $\mathrm{SL}(2, \mathbb{Z})\backslash\mathrm{SL}(2, \mathbb{R})$. See [14] and a discussion in [2] for more on this connection and for references to the literature. Roughly speaking, the natural extension of a continued fraction transformation can be identified with a cross section for the geodesic flow. For example, the transformation T from (2.9) of the regular continued fraction's natural extension gives a planar representation of the first return map and ω corresponds to the Liouville measure. Geodesics can be identified with (proper classes of) indefinite binary quadratic forms and a cross section with a reduction domain. The trajectories we study in this paper correspond to cuspidal geodesics or, equivalently, forms with one rational root.

Of course there is great interest in similar Diophantine problems about general indefinite forms and hence general geodesic trajectories. A prime example is the Markov problem [32] about the minima of such forms and their possible values; these values determine the

Markov spectrum (see [3] and its references). The Lagrange spectrum is similarly defined using cuspidal trajectories; it is determined by the values of

$$\lambda(\alpha) = \liminf_{t\geq 1} \rho_t(\alpha), \quad \text{where } \rho_t(\alpha) = t \min_{\substack{p\in\mathbb{Z}\\ 1\leq q\leq t}} |p-\alpha q| \text{ for } t \geq 1.$$

The Dirichlet spectrum is determined by the values of $\delta(\alpha) = \limsup_{t\geq 1} \rho_t(\alpha)$; in [23] it is defined to be the set of values of $\frac{\delta(\alpha)}{1-\delta(\alpha)}$. There is a spectrum that is related to the Dirichlet spectrum in the same way that the Markov spectrum is related to the Lagrange spectrum. Like the Markov problem, its study involves general geodesic trajectories and their associated continued fractions. Again speaking roughly, we replace lim sup over cuspidal geodesics in the definition of $\delta(\alpha)$ by the supremum over all geodesics. Mordell [38] introduced this problem (actually an n-dimensional version), which he posed as a kind of converse to Minkowski's linear forms theorem. The case of two dimensions was treated in more detail by Szekeres [56], Oppenheim [43] and Burger [6]. This problem in higher dimensions has also attracted a lot of attention (see e.g. [46, 47, 55]).

It should be apparent that a general spectrum of this type can be defined for any strongly symmetric norm F, not just the sup-norm, and that an associated reduction theory for indefinite binary quadratic forms can be developed that uses F-continued fractions. For the 2-norm the problem was introduced by Oppenheim [44] and the relevant reduction theory was already found by Hermite. Minkowski developed the reduction theory for the 1-norm with Hermite's theory in mind and certainly knew that a version could be based on the p-norm for a general p [36, footnote on p. 166]. However, outside of the sup-norm, only isolated aspects of the spectrum and reduction theory have been considered and only for the p-norm for $p = 1, 2$.

Appendix A. Lemmas about norms

Here we state and prove a number of technical lemmas that are referred to in the body of the paper. We begin with a lemma about decreasing concave down functions that we will use later in the appendix.

Lemma A.1. *Fix $s > 0$ and $t > 1$ and let $S \subseteq \mathbb{R}$. Suppose that $f : S \to \mathbb{R}$ is nonnegative, concave down, and strictly decreasing. Then the function $g(x) = stf(tx/s) - f(x)$ has at most one zero.*

Proof. If $t \leq s$ then since f is decreasing we have $g(x) \geq (t^2-1)f(x) \geq 0$, with equality if and only if $t = s$ and $x = \sup S \in S$. So we may assume that $t > s$.

Let $a, b \in S$ and pick any $\theta \in [0, 1]$ such that $\theta a + (1-\theta)b \in S$. Since f is concave down we have

$$f(\theta a + (1-\theta)b) \geq \theta f(a) + (1-\theta)f(b).$$

Suppose that $c, d \in S$ and that $a < b < c < d$. Taking $\theta = \frac{c-b}{c-a}$ we find that

$$f(b) \geq \left[\frac{c-b}{c-a}\right] f(a) + \left[\frac{b-a}{c-a}\right] f(c),$$

from which it follows that

$$\frac{f(b)-f(a)}{b-a} \geq \frac{f(c)-f(a)}{c-a}.$$

Similarly, we have

$$\frac{f(c)-f(a)}{c-a} \geq \frac{f(d)-f(c)}{d-c}. \tag{A.1}$$

The last two inequalities together imply that

$$\frac{f(b)-f(a)}{b-a} \geq \frac{f(d)-f(c)}{d-c}. \tag{A.2}$$

By a similar argument we also have

$$\frac{f(c)-f(a)}{c-a} \geq \frac{f(d)-f(b)}{d-b}. \tag{A.3}$$

Suppose that there exist $x_1 < x_2$ such that $g(x_1) = g(x_2)$, i.e.

$$st\big[f(tx_2/s) - f(tx_1/s)\big] = f(x_2) - f(x_1). \tag{A.4}$$

There are three cases, depending on the ordering of x_2 and tx_1/s. In each case we will prove the inequality

$$st\left[\frac{f(tx_2/s)-f(tx_1/s)}{x_2-x_1}\right] = \frac{f(x_2)-f(x_1)}{x_2-x_1} \geq \frac{f(tx_2/s)-f(tx_1/s)}{tx_2/s - tx_1/s}, \tag{A.5}$$

which implies that $t^2 \leq 1$ because f is strictly decreasing. This is a contradiction.

Case 1: If $x_1 < x_2 < tx_1/s < tx_2/s$ then (A.5) follows from (A.4) and (A.2).

Case 2: If $x_1 < tx_1/s < x_2 < tx_2/s$, we use (A.3) instead of (A.2).

Case 3: Lastly, if $x_1 < x_2 = tx_1/s < tx_2/s$ then we use (A.1). □

In the remainder of the appendix F is a norm on $\mathbb{R}^2$ with unit ball $\mathcal{B}$ and $P = (x, y), P' = (x', y') \in \mathbb{R}^2$. For $t > 0$ we define as above $F_t(x, y) = F(t^{-1}x, ty)$. The following lemmas give various properties of norms that satisfy the first condition (2.4) of strong symmetry. Note that if F satisfies (2.4) then so does F_t for any $t > 0$. The first result is crucial and is used repeatedly in this paper.

Lemma A.2. *Suppose that F satisfies (2.4). If $|x'| \leq |x|$ and $|y'| \leq |y|$ then we have that*

$$F(P') \leq F(P).$$

Proof. Observe that if $F(P) = s$ then $F(\pm x, \pm y) = s$ hence $F(x', y') \leq s$ by convexity. □

Lemma A.3. *If F, G, H satisfy (2.4) then so does K defined by*

$$K(P) = H(F(P), G(P)).$$

Proof. This follows easily using Lemma A.2. □

Lemma A.4. *Suppose that F satisfies (2.4). The following properties hold.*

(i) If $F(P') \geq F(P)$ with $|x| \neq |x'|$ and $|y'| < |y|$ then for some unique $t \geq 1$ we have

$$F_t(P') = F_t(P).$$

(ii) If $F(P') \geq F(P)$ with $|y| \neq |y'|$ and $|x'| < |x|$ then for some unique $t \leq 1$ we have

$$F_t(P') = F_t(P).$$

Proof. We only prove (i) as (ii) is a consequence of (i) applied to the norm $G(x,y) = F(y,x)$. *Existence:* If $F(P') = F(P)$ take $t = 1$. Otherwise for any $P \in \mathbb{R}^2$ define the continuous function $f_P : [1,\infty) \to \mathbb{R}^+$ by $f_P(t) = t^{-1}F_t(P)$. Now by Lemma A.2

$$f_P(t) = F(t^{-2}x, y) \geq F(0,y) = |y|F(0,1).$$

On the other hand, $f_{P'}(t) = F(t^{-2}x', y') \to F(0,y') = |y'|F(0,1) < |y|F(0,1)$ as $t \to \infty$. Because $f_P(1) < f_{P'}(1)$ the existence of desired t follows by the intermediate value theorem.

Uniqueness: Suppose that for $t_2 > t_1 \geq 1$ we have

$$F_{t_1}(P) = F_{t_1}(P') = s_1 \quad \text{and} \quad F_{t_2}(P) = F_{t_2}(P') = s_2.$$

By (2.4) we may assume that P and P' both lie in the first quadrant. Letting $Q = (t_2^{-1}x/s_2, t_2y/s_2)$ and $Q' = (t_2^{-1}x'/s_2, t_2y'/s_2)$ we find that

$$F_1(Q) = F_1(Q') = 1 \ \text{ and } \ F_t(sQ) = F_t(sQ') = 1, \quad \text{where } t = t_1/t_2 < 1 \text{ and } s = s_2/s_1.$$

We first assume that the boundary of $\mathcal{B}$ has no horizontal or vertical segments. Then the portion of the graph of the boundary of $\mathcal{B}$ in the first quadrant defines a nonnegative function $f(X)$ which is strictly decreasing and concave down. Furthermore, the points Q and Q' both lie on the curves $Y = f(X)$ and $tY/s = f(t^{-1}X/s)$. It follows that the function $g(X) = st^{-1}f(t^{-1}X/s) - f(X)$ has at least two zeros which contradicts Lemma A.1.

If the boundary of $\mathcal{B}$ contains a horizontal segment then the points Q, Q' cannot both lie on that segment because $y \neq y'$. Similarly, the points Q, Q' cannot both lie on a vertical segment since $x \neq x'$. Any horizontal segment must be of the form $[0,a] \times \{d\}$ and any vertical segment must be of the form $\{b\} \times [0,c]$. The portion of the graph of the boundary of $\mathcal{B} \cap ([a,b] \times [c,d])$ defines a nonnegative function $f_1(X)$ which is decreasing and concave down. There are several cases to consider. If neither Q nor Q' lies on a horizontal or vertical segment, we can apply the argument in the previous paragraph to the function $f_1(X)$ on the interval $[a,b]$. If, say, $Q = (u,v)$ lies on the horizontal segment and Q' is not on the vertical segment, instead use the function

$$f_2(X) = \begin{cases} v & \text{if } X = u \leq a, \\ f_1(X) & \text{if } X \in (a,b], \end{cases}$$

which is also strictly decreasing and concave down (but not necessarily continuous). If $Q' = (u',v')$ lies on the vertical segment and Q is not on the horizontal segment, use the function

$$f_3(X) = \begin{cases} f_1(X) & \text{if } X \in [a,b), \\ v' & \text{if } X = u' = b. \end{cases}$$

If Q lies on the horizontal segment and Q' lies on the vertical segment, a suitable function f_4 can be defined similarly. □

The following result is trivial in case the norm is strictly convex.

Lemma A.5. *Suppose that that F satisfies (2.4), that we have $F(P) = F(P')$ and that $0 < x' < x$ and $0 < |y| < y'$. Then for any $d \geq 1$*

$$F(P - dP') < F(P) + dF(P'). \tag{A.6}$$

Proof. To see this note first that in order for equality to hold in (A.6) we must have that

$$F(P) + dF(P') = F(P - P' - (d-1)P') \leq F(P - P') + (d-1)F(P'),$$

which implies that

$$F(P - P') \geq F(P) + F(P') \text{ so that } F(P - P') = F(P) + F(P')$$

hence

$$F(\tfrac{1}{2}(P - P')) = \tfrac{1}{2}(F(P) + F(P')) = F(P) = F(-P').$$

That this is impossible follows by a simple convexity argument using the locations of

$$P = (x, y) \quad \text{and} \quad -P' = (-x', -y'),$$

together with (2.4). □

Lemma A.6. *Suppose that F satisfies (2.4). For $\sigma, \sigma' \in [0,1]$ with $\sigma + \sigma' = 1$ and $1 \leq t_1 \leq t_2$ we have*

$$F_{\sigma t_1 + \sigma' t_2}(P) \leq \sigma F_{t_1}(P) + \sigma' F_{t_2}(P).$$

Proof. Using the fact that the function $t \mapsto t^{-1}$ is concave up and applying Lemma A.2 we get that

$$F_{\sigma t_1 + \sigma' t_2}(x, y) \leq F\big(x(\tfrac{\sigma}{t_1} + \tfrac{\sigma'}{t_2}), y(\sigma t_1 + \sigma' t_2)\big).$$

By the defining properties of a norm we finish the proof. □

Acknowledgement

The authors thank the anonymous referee for their very careful reading and their numerous excellent suggestions to improve this manuscript.

References

[1] Andersen, N. & Duke, W., Markov spectra for modular billiards. *Math. Ann.* 373 (2019), no. 3–4, 1151–1175.

[2] Arnoux, P. & Schmidt, T.A., Cross sections for geodesic flows and α-continued fractions. *Nonlinearity* 26 (2013), no. 3, 711–726.

[3] Bombieri, E., Continued fractions and the Markoff tree. *Expo. Math.* 25 (2007), no. 3, 187–213.

[4] Bosma, W., Optimal continued fractions. Nederl. Akad. Wetensch. Indag. Math. 49 (1987), no. 4, 353–379.

[5] Bosma, W. & Jager, H. & Wiedijk, F., Some metrical observations on the approximation by continued fractions. *Nederl. Akad. Wetensch. Indag. Math.* 45 (1983), no. 3, 281–299.

[6] Burger, E. B., On a question of Mordell and a spectrum of linear forms. *J. London Math. Soc.* (2) 62 (2000), no. 3, 701–715.

[7] Cassels, J. W. S., An introduction to the geometry of numbers. Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete, Bd. 99 Springer-Verlag, Berlin-Göttingen-Heidelberg 1959 viii+344 pp.

[8] Cohn, H. Minkowski's conjecture on critical lattices in the metric $(|\xi|^p + |\eta|^p)^{\frac{1}{p}}$, *Ann. of Math.* 51 (1950) 734–738.

[9] Dajani, K. & Kraaikamp, C., Ergodic theory of numbers. Carus Mathematical Monographs, 29. Mathematical Association of America, Washington, DC, 2002. x+190 pp.

[10] Davenport, H. & Schmidt, W. M., Dirichlet's theorem on Diophantine approximation. 1970 Symposia Mathematica, Vol. IV (INDAM, Rome, 1968/69) pp. 113–132 Academic Press, London.

[11] Davenport, H. & Schmidt, W. M., Dirichlet's theorem on Diophantine approximation. II. *Acta Arith.* 16 1969/1970 413–424.

[12] Davis, C. S., Note on a conjecture by Minkowski. *J. London Math. Soc.* 23, (1948). 172–175.

[13] Dirichlet, L.G.P., Verallgemeinerung eines Satzes aus der Lehre von den Kettenbrüchen nebst einigen Anwendungen auf die Theorie der Zahlen, 1842, Werke I, 633–638.
[14] Einsiedler, M. & Ward, T., Ergodic theory with a view towards number theory. Graduate Texts in Mathematics, 259. Springer-Verlag London, Ltd., London, 2011. xviii+481 pp.
[15] Euler, L., De fractionibus continuis dissertatio, Opera Omnia: Series 1, Volume 14, pp. 187 - 216 (1744)
[16] Glazunov, N. M. & Golovanov, A. S. & Malyshev, A. V., Proof of the Minkowski conjecture on the critical determinant of the region $|x|^p + |y|^p < 1$. (Russian) Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) 151 (1986), Issled. Teor. Chisel. 9, 40–53, 195; translation in J. Soviet Math. 43 (1988), no. 5, 2645–2653.
[17] Gruber, P. M. & Lekkerkerker, C. G., Geometry of numbers. Second edition. North-Holland Mathematical Library, 37. North-Holland Publishing Co., Amsterdam, 1987. xvi+732 pp.
[18] Hermite C., Extraits de lettres de Mr. Ch. Hermite à M. Jacobi sur différents objets de la théorie des nombres. J. Reine Angew. Math. 40, 261–278 in Oeuvres I.
[19] Hermite, C. Sur l'introduction des variables continues dans la theorie des nombres. *J reine Angew Math.* 41: (1851) 191–216.
[20] Humbert, G., Sur la méthode d'approximation d'Hermite. *J Math Pures Appl.* (7th Ser) 2: (1916) 70–103.
[21] Humbert, G., Sur les fractiones continues ordinaires et les formes quadatique binaires indéfinies. *J Math Pures Appl.* (7th Ser) 2: (1916) 104–154.
[22] Iosifescu, M. & Kraaikamp, C., Metrical theory of continued fractions. Mathematics and its Applications, 547 Kluwer Academic Publishers, Dordrecht, (2002) xx+383 pp.
[23] Ivanov, V. A., A theorem of Dirichlet in the theory of Diophantine approximations. (Russian) *Mat. Zametki* 24 (1978), no. 4, 459–474, 589. English translation: Math. Notes 24 (1978), no. 3–4, 747–755 (1979).
[24] Jager, H., Continued fractions and ergodic theory, transcendental numbers and related topics, RIMS Kokyuroko **599** (1986) no.1. 55–59.
[25] Khintchine, A.Ya., Continued fractions, English transl. by P. Wynn, Noordhoff, Groningen, (1963).
[26] Kraaikamp, C., Statistic and ergodic properties of Minkowski's diagonal continued fraction. *Theoret. Comput. Sci.* 65 (1989), no. 2, 197–212.
[27] Kraaikamp, C., A new class of continued fraction expansions. *Acta Arith.* 57 (1991), no. 1, 1–39.
[28] Lagrange, J.L., Additions aux élements d'algebra d'Euler, Oeuvres VII.
[29] Mahler, K. Lattice points in two-dimensional star domains. I. *Proc. London Math. Soc.* (2) 49, (1946) 128–157.
[30] Mahler, K. On the minimum determinant and the circumscribed hexagons of a convex domain. *Nederl. Akad. Wetensch.,* Proc. 50, (1947) 692–703=*Indagationes Math.* 9, (1947) 326–337.
[31] Malyšev, A. V., The application of an electronic computer to the proof of a certain conjecture of Minkowski from the geometry of numbers. (Russian) Modules and representations. Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) 71 (1977), 163–180, 286.
[32] Markoff, A., Sur les formes quadratiques binaires indéfinies. *Math. Ann.* 15 (1879) 381–409, 17 (1880) 379–399.
[33] Minkowski, H., Zur Theorie der Kettenbrüche, *Ann. de l'Ëcole Normale sup.*, ser 3. XIII, 41–60 (1896), in Gesammelte Abhandlungen I . 278–292.
[34] Minkowski, H., Über die Annäherung an eine reele Größe durch rational Zahlen. *Math. Ann.* 54, 91–124 (1901), in Gesammelte Abhandlungen I, 320–352.
[35] Minkowski, H., Dichteste gitterförmige Lagerung kongruenter Körper, Nachr. K. Ges. Wiss. Göttingen, (1904) 311-355, in Gesammelte Abhandlungen. Vol. II, Teubner, Berlin, (1911) pp. 3–42.
[36] Minkowski, H., Geometrie der Zahlen, Teubner (1910)
[37] Minkowski, H., Diophantische Approximationen, 2d ed., Teubner, Leipzig, 1927, pp. 51–58.
[38] Mordell, L. J., Note on an arithmetical problem on linear forms. *London Mathematical Society* 12 (1937): 34–6.
[39] Mordell, L. J., Lattice points in the region $|Ax^4 + By^4| \leq 1$. *J. London Math. Soc.* 16, (1941) 152–156.
[40] Moshchevitin, N., On Minkowski diagonal continued fraction. Analytic and probabilistic methods in number theory, 197–206, TEV, Vilnius, (2012).

[41] Nakada, H., Metrical theory for a class of continued fraction transformations and their natural extensions. *Tokyo J. Math.* 4, (1981), 399–426.
[42] Nakada, H.; Ito, S.; Tanaka, S., On the invariant measure for the transformations associated with some real continued-fractions. *Keio Engrg. Rep.* 30 (1977), no. 13, 159–175.
[43] Oppenheim, A., The continued fractions associated with chains of quadratic forms. *Proc. London Math. Soc.* (2) 44 (1938), no. 5, 323–335.
[44] Oppenheim, A., Two lattice-point problems. *Quart. J. Math.*, Oxford Ser. 18, (1947) 17–24.
[45] Perron, O., Die Lehre von den Kettenbrüchen. Bd I. Elementare Kettenbrüche. (German) 3te Aufl. B. G. Teubner Verlagsgesellschaft, Stuttgart, (1954) vi+194 pp.
[46] Ramharter, G., Über ein Problem von Mordell in der Geometrie der Zahlen. *Monatshefte fr Mathematik* 92 (1981): 143–60.
[47] Ramharter, G., On Mordell's inverse problem in dimension three. *J. Number Th.* 58, (1996) 388–415.
[48] Reinhardt, K., Über die dichteste gitterförmige lagerung kongruenter bereiche in der ebene und eine besondere art konvexer kurven. (German) *Abh. Math. Sem. Univ. Hamburg* 10 (1934), no. 1, 216–230.
[49] Roy, D., On Schmidt and Summerer parametric geometry of numbers. *Ann. of Math.* (2) 182 (2015), no. 2, 739–786.
[50] Schmidt, W. M., Diophantine approximation and certain sequences of lattices. *Acta Arith.* 18 1971 195–178.
[51] Schmidt, W. M., Diophantine approximation. *Lecture Notes in Mathematics*, 785. Springer, Berlin, 1980. x+299 pp.
[52] Schmidt, W. M. & Summerer, L., Parametric geometry of numbers and applications. *Acta Arith.* 140 (2009), no. 1, 67–91.
[53] Schmidt, W. M. & Summerer, L., Diophantine approximation and parametric geometry of numbers. *Monatsh. Math.* 169 (2013), no. 1, 51–104.
[54] Siegel, C.L., Lectures on the geometry of numbers. Notes by B. Friedman. Rewritten by Komaravolu Chandrasekharan with the assistance of Rudolf Suter. With a preface by Chandrasekharan. Springer-Verlag, Berlin, 1989. x+160 pp.
[55] Shapira, U. & Weiss, B., On the Mordell-Gruber spectrum. *Int. Math. Res. Not.* IMRN 2015, no. 14, 5518–5559.
[56] Szekeres, G., On a problem of the lattice plane. *J. London Math. Soc.* 12, (1936) 88–93.
[57] Tietze, H., Über die raschesten Kettenbruchentwicklungen reeller Zahlen, *Monatsh. Math. Phys.*, 24, (1913) 209–241.
[58] Watson, G. L., Minkowski's conjectures on the critical lattices of the region $|x|^p + |y|^p \leq 1$. I. *J. London Math. Soc.* 28, (1953). 305–309.
[59] Watson, G. L., Minkowski's conjectures on the critical lattices of the region $|x|^p + |y|^p \leq 1$. II. *J. London Math. Soc.* 28, (1953). 402–410.

UCLA Mathematics Department, Box 951555, Los Angeles, CA 90095-1555
Email address: nandersen@math.ucla.edu

UCLA Mathematics Department, Box 951555, Los Angeles, CA 90095-1555
Email address: wdduke@ucla.edu