ON A THEOREM OF DAVENPORT AND SCHMIDT

NICKOLAS ANDERSEN AND WILLIAM DUKE

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 more irrationals can be improved and if so, by how much. In this paper 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 $\alpha$ 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 $\alpha$), 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 $\alpha$. Clearly $\delta(\alpha) = 0$ for rational $\alpha$ so we only consider irrational $\alpha$.

An easy direct argument proves the fact, perhaps surprising at first, that any irrational $\alpha$ for which $\delta(\alpha) < 1$ must be badly approximable. For $\alpha$ 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 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 $\alpha$.

Recall that an irrational $\alpha$ has a unique infinite regular continued fraction expansion

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

**Date:** July 3, 2019.

Supported by NSF grant DMS 1701638.
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

$$u_n = \frac{1}{b_{n+1}} + \frac{1}{b_{n+2}} + \cdots$$

and

$$v_n = \frac{1}{b_n} + \frac{1}{b_{n-1}} + \frac{1}{b_{n-2}} + \cdots \frac{1}{b_1}.$$

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

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

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 $\alpha$ being badly approximable [49, 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

$$e - 1 \frac{e + 1}{2 + 6 + 10 + 14 + \cdots}.$$

By a well-known result of Khintchine [24, 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

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

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 or more numbers simultaneously (see also [11] and [48]). In this paper we will consider a different kind of generalization of Dirichlet’s results, one that was conceived of by Hermite and Minkowski.

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

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

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).

\(^1\)For further results about the set of values of $\delta(\alpha)$ see [22] and the references therein. See also our §12.
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 $B$. Fix $\alpha \in \mathbb{R}$. Then for any real $t \geq 1$ there exist integers $p, q$ with $q > 0$ such that

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

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 $\alpha$ there are infinitely many integers $p, q$ with $q > 0$ such that

$$q|p - \alpha q| < \frac{1}{\sqrt{3}},$$

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 $\alpha$. 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 $\delta$ in the Davenport–Schmidt theorem.

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

$$F(x, y) = F(|x|, |y|).$$

At least for the study of $\delta_F$, we may assume without any further loss of generality that the norm $F$ also satisfies

$$F(0, \pm 1) = F(\pm 1, 0) = 1.$$

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

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

while $F(\infty)(x, y) = \sup\{|x|, |y|\}$. Denote the corresponding $B$ by $B^p$, $\Delta$ by $\Delta_p$ and $\delta$ by $\delta_p$. Other interesting examples are the two unique strongly symmetric norms whose unit balls are regular octagons: $B^{oct_1}$ and $B^{oct_2}$ (see Figure 1).

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 $\alpha$.

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

$$\delta_F(\alpha) \geq \frac{1}{2}.$$  

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 $\alpha$ is $\Delta$.

We will see in the proof of Theorem 3 that $\delta_p(\alpha) = \Delta$ for any $\alpha$ 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 $\alpha$.

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

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

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

The value of $\Delta_p$ is given below in (4.2). See Figure 2 for graphs of $\Delta_p$ and the minimum value of $\delta_p$. 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

$$\delta_2\left( \frac{1}{2}(-1 + \sqrt{3}) \right) = 1.$$  

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 $\delta_p$ among all $p$-norms. It can be shown that the minimum value of
\( \delta_F(\alpha) \) for both of the octagonal norms is \( \frac{1}{8} (3\sqrt{2} + 2) = 0.78033 \ldots \) (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](image-url)

**Figure 2.** The minimum of \( \delta_p \) for \( p \geq 1 \). The dotted line is \( \Delta_p \), the minimum of \( \delta_p(\alpha) \) for well approximable \( \alpha \).

**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 [50, 51]. See also [47] 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

\[
T(u, v) = \left( \frac{1}{u} - \left\lfloor \frac{1}{u} \right\rfloor, \frac{1}{v + \left\lfloor \frac{1}{u} \right\rfloor} \right).
\]

(2.9)

It has an invariant measure \( \omega \) with density function

\[
\frac{1}{\log 2} \frac{1}{(1 + uv)^2}.
\]

(2.10)

The ergodicity of this system, 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 [23], 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 \( \pm 1 \) as partial numerators. The continued fractions we need are examples of \( S \)-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 \( SL(2, \mathbb{Z}) \backslash SL(2, \mathbb{R}) \) 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

\[ a_0 + \frac{\varepsilon_1}{a_1+} \frac{\varepsilon_2}{a_2+} \ldots, \quad \varepsilon_m = \pm 1, \quad a_m \in \mathbb{Z} \]

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 \( \frac{p_m}{q_m} = a_0 + \frac{\varepsilon_1}{a_1+} \frac{\varepsilon_2}{a_2+} \ldots \frac{\varepsilon_m}{a_m} \) uniquely defines relatively prime integers \( p_m, q_m \) with \( q_m > 0 \), where \( p_0 = a_0 \) and \( q_0 = 1 \).

Tietze ([55], see also [43, p. 135]) showed that there is an irrational \( \alpha \) 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 \( \alpha \) 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 \( \alpha \) 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 \( \alpha \) with respect to \( F \).

We will refer to this continued fraction as the \( F \)-continued fraction of \( \alpha \) and, for the \( p \)-norm, as the \( p \)-continued fraction of \( \alpha \). For the sup-norm the \( \infty \)-continued fraction is closely related to, but not always equal to, the regular continued fraction. Suppose that

\[ \alpha = b_0 + \frac{1}{b_1 + \frac{1}{b_2 + \frac{1}{b_3 + \cdots}}} \]

is the regular continued fraction of an irrational \( \alpha \). Recall that Lagrange showed ([25], see also [43, §15]) that every best approximation in the usual sense is a convergent of the regular continued fraction of \( \alpha \) and that every convergent, except possibly \( b_0 \), is a best approximation to \( \alpha \). In view of Theorem 5, (3.2) coincides with the \( \infty \)-continued fraction of \( \alpha \) if and only if \( b_1 > 1 \). If \( b_1 = 1 \) the \( \infty \)-continued fraction of \( \alpha \) is

\[ \alpha = b_0 + 1 + \frac{-1}{b_2 + \frac{1}{b_3 + \cdots}}. \]

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 $\alpha$ in this case. This well-known exceptional case does not occur for the $\infty$-continued fraction. With this one possible exception, however, the convergents of the $\infty$-continued fraction and those of the regular continued fraction coincide.

For any $1 \leq p < \infty$, 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 $\alpha$ to be a convergent of the $p$-continued fraction is that

$$s_n|r_n - \alpha s_n| < \left(4^{1/p} \epsilon_p\right)^{-1}.$$  \hspace{1cm} (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 $\mu_m, \nu_m$ determined by our continued fraction $\alpha = a_0 + \frac{\varepsilon_1}{a_1 + a_2 +} \frac{\varepsilon_2}{a_3 +} \cdots$. Namely, for a fixed norm we define $\mu_0 = \alpha$ and $\nu_0 = 0$, while for $m \geq 1$ we let

$$\mu_m = \frac{\varepsilon_{m+1}}{a_{m+1}} + \frac{\varepsilon_{m+2}}{a_{m+2}} + \cdots \quad \text{and} \quad \nu_m = \frac{1}{a_m} + \frac{\varepsilon_{m-1}}{a_{m-1} + a_{m-2} +} \cdots \frac{\varepsilon_2}{a_1}.$$  \hspace{1cm} (3.5)

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

$$D_p(u, v) = \frac{1}{1 + uv} \left(\frac{(1 - |u|^p v^p)^2}{(1 - |u|^p)(1 - v^p)}\right)^{\frac{1}{p}},$$  \hspace{1cm} (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 $1 \leq p \leq \infty$. For any irrational $\alpha$ whose $p$-continued fraction is (3.1) we have that

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

where $\mu_m, \nu_m$ are given above in (3.5).

The $F$-continued fraction of an irrational $\alpha$ for any strongly symmetric norm is an example of an $\mathcal{S}$-expansion. Their theory has been developed by Kraaikamp [28] and others (see also [4], [9] and [26]). Recall the definition of $T$ and $\omega$ from (2.9) and (2.10). A Borel set $\mathcal{S} \subset \Omega_0$ is called a singularity area if $\omega(\partial \mathcal{S}) = 0$ and if

(i) $\mathcal{S} \subseteq [\frac{1}{2}, 1] \times [0, 1]$ and

(ii) $TS \cap \mathcal{S} \subseteq \{\beta, \beta\}$, where $\beta = \frac{1}{2}(-1 + \sqrt{5})$.

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

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

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 for $\alpha$ 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 $\infty$-continued fraction discussed above is the $\mathcal{S}$ expansion for $\mathcal{S} = [\frac{1}{2}, 1] \times \{0\}$. As usual, we denote $\mathcal{S}$ by $\mathcal{S}_p$ in the case of the $p$-norm (see Figure 3).
Theorem 7. Fix a strongly symmetric norm $F$. There exists a singularization area $S$ so that the $F$-continued fraction of any irrational $\alpha$ is the $S$-expansion of $\alpha$.

**Figure 3.** The $S_p$ 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 [27] that the 1-continued fraction (known as Minkowski’s diagonal continued fraction) is an $S$-expansion. For related work on the 1-continued fraction see [40]. That the $p$-continued fraction for $p \neq 1, \infty$ is also an $S$-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:

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

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

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

This $B$ is open, bounded, convex and symmetric around 0 and every such body arises as the unit ball of some norm (see e.g. [52]). Denote by area($B$) the Lebesgue measure of $B$ on $\mathbb{R}^2$. It is convenient to define the stretched ball for $t > 0$

$$B_t = \{(x, y) \in \mathbb{R}^2; F_t(x, y) < 1\}.$$
Let $L \subset \mathbb{R}^2$ be a (full) lattice. By the determinant of $L$, denoted $\det L$, we mean $|\det g|$ for any $g \in \text{GL}(2, \mathbb{R})$ whose rows give a $\mathbb{Z}$-basis for $L$. The lattice $L$ is admissible for $B$ if $B$ 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 $B$ then

$$\text{area } B \leq 4 \det L.$$  

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

$$\Delta \geq \frac{1}{4}(\text{area } B).$$  

(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 $\Delta$ exactly. Finding the critical determinant of a given $B$ 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 $B$ the boundary of $B$ 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 $B$ is admissible for $B$ (see [7, Thm XI p. 160]). Therefore, as Minkowski already knew, computing $\Delta$ 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 $B$. This justifies our definition of $\Delta$ in the statement of the Minkowski approximation theorem.

Next we review what is known about the value of $\Delta_p$ for all $p$. Let

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

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

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

(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 [56, 57] and Malyshev [31]. Building on their work, the proof of the full conjecture was finally completed by Glazunov, Golovanov and Malyshev [16].

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

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

(4.3)

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

$$P = (2^{-\frac{1}{p}}(1 - \tau_p)^{-1}, -2^{-\frac{1}{p}}\tau_p(1 - \tau_p)^{-1}) \quad \text{and} \quad P' = (2^{-\frac{1}{p}}, 2^{-\frac{1}{p}})$$

(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 $\infty$ these parallelograms are unique up to obvious symmetries. When $p = 1, 2$ or $\infty$ 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 \( 3\Delta(B) \) is the minimal area of an affinely regular symmetric hexagon inscribed in \( B \). A useful alternative due to Reinhardt [46] is that \( 4\Delta(B) \) 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 \( \Delta(B_{oct1}) = \sqrt{2} - \frac{1}{2} \) for the regular octagon from Figure 2. Thus we also have \( \Delta(B_{oct2}) = \frac{1}{8}(3\sqrt{2} + 2) \), 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

\[
(5.1) \quad L_\alpha = (1, -\alpha)\mathbb{Z} + (0, 1)\mathbb{Z}.
\]

Obviously \( L_\alpha \) 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 \( L_\alpha \) 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 \( L_\alpha \) for any \( t \in (t_{m-1}, t_m) \). We will start with \( P_{-1} = (0, 1) \) and \( P_0 = (1, -\alpha) \). Roughly speaking, given \( P_{m-1} \), to find the new point \( P_m \) and the associated \( t_m \), we simultaneously expand \( B \) 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 \( L_\alpha \) that are linearly independent over \( \mathbb{R} \) and on a ball for which \( L_\alpha \) is admissible. First we need to know that they give a basis for \( L_\alpha \).

**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 \( B_t \) for some \( t > 0 \) and are linearly independent over \( \mathbb{R} \). If \( L_\alpha \) is admissible for \( B_t \) then \( \{P, P'\} \) gives a \( \mathbb{Z} \)-basis for \( L_\alpha \).
Proof. Consider the sublattice \( P\mathbb{Z} + P'\mathbb{Z} \) of \( L_\alpha \) generated by these lattice points. By Minkowski’s first convex body theorem its index in \( L_\alpha \) 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 \( \mathcal{B}_t \cap \mathcal{L}_\alpha \). 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

\[
\frac{1}{2}(P + P') = \left( \frac{a+c}{2} \right)Q + \left( \frac{b+d}{2} \right)Q' \quad \text{and} \quad \frac{1}{2}(P - P') = \left( \frac{a-c}{2} \right)Q + \left( \frac{b-d}{2} \right)Q'
\]

are distinct points of \( L_\alpha \). 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 \( L_\alpha \) contains points of the \( x \)-axis, i.e. \( \alpha \) would be rational. \( \square \)

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

**Lemma 5.2.** Fix a strongly symmetric norm \( F \) and an irrational \( \alpha \in \left( \frac{-1}{2}, \frac{1}{2} \right] \). There is a sequence \( 1 = \tau_{-1} \leq \tau_0 < \tau_1 < \tau_2 < \cdots \) tending to \( \infty \) and for each \( m = -1, 0, 1, \ldots \) 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 \( L_\alpha \) with respect to \( F_{\tau_m} \),
(iii) for any \( t \in (\tau_{m-1}, \tau_m) \) 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

\[
(5.2) \quad \mathcal{B}(P, \tau) \overset{\text{def}}{=} \{P' \in \mathbb{R}^2; F_t(P') < F_t(P)\},
\]
from the form of \( L \) and \( B \) for \( t \) is a best approximation with respect to the sup-norm. Thus, as long as \( y \neq 0 \), by Minkowski’s first convex body theorem there will be a maximal \( t' \geq t \) for which \( L_\alpha \) is admissible for \( B(P,t') \). 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 \( L_\alpha \) with respect to \( F_{t'} \).

Let \( P_{-1} = (0,1) \). Then \( L_\alpha \) is admissible for \( B(P_{-1},1) \). Let \( t_0 \geq 1 \) be maximal for which \( L_\alpha \) is admissible for \( B(P_{-1},t_0) \). 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 \( L_\alpha \) is admissible for \( B(P_0,t_0) \) and so we find \( t_1 > t_0 \) maximal so that \( L_\alpha \) is admissible for \( B(P_0,t_1) \). 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 \( \alpha \) 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 \( L_\alpha \), where \( \alpha \) 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 \( A.1 \) and Minkowski’s first convex body theorem again that for each \( m \geq 0 \)
\[
x_m t_m^{-1} = F(x_m t_m^{-1},0) \leq F(x_m t_m^{-1},y_m t_m) = F_{t_m}(P_m) \leq 2(\text{area } B)^{-\frac{1}{2}}.
\]
Thus \( t_m \gg x_m \to \infty \) as \( x_m > x_{m-1} \) are integers. \( \square \)

**Proof of Theorem 2.** Observe that \( B(P_m, t_m) \) 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 \( L_\alpha \) with respect to \( F_{t_m} \). Therefore
\[
\Delta F_{t_m}^2(P_m) \geq 2\Delta(\text{area } B)^{-1} \geq \frac{1}{2},
\]
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
\[
\delta_F(\alpha) \geq \Delta \limsup_{m \to \infty} F_{t_m}^2(P_m) \geq \frac{1}{2},
\]
thus proving Theorem 2. \( \square \)

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 \( \alpha \) 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 \( \alpha \) 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.1.

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. \( \square \)

**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 \( \alpha \) 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 \( g_{-1} = (\begin{smallmatrix} 0 & 1 \\ 1 & -\alpha \end{smallmatrix}) \). We know by Lemma 5.2 that for each \( m \geq 1 \) there is a positive integer \( a_m \) and \( \varepsilon_m = \pm 1 \) so that
\[
(6.1) \quad g_m = \begin{pmatrix} a_m & \varepsilon_m \\ 1 & 0 \end{pmatrix} g_{m-1}.
\]
This also holds for \( m = 0 \) if we set \( \varepsilon_0 = 1 \). Clearly for \( m \geq 0 \)
\[
(6.2) \quad \gamma_m \overset{\text{def}}{=} \det g_m = (-1)^m \varepsilon_1 \cdots \varepsilon_m.
\]

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
\[
(6.3) \quad p_m = a_mp_{m-1} + \varepsilon_mp_{m-2}, \quad p_{-1} = 1, \quad p_{-2} = 0,
\]
\[
(6.4) \quad q_m = a_q q_{m-1} + \varepsilon_q q_{m-2}, \quad q_{-1} = 0, \quad q_{-2} = 1.
\]

It is easy to see that for \( m \geq 0 \)
\[
(6.5) \quad \begin{pmatrix} p_m & q_m \\ p_{m-1} & q_{m-1} \end{pmatrix} = \begin{pmatrix} a_m & \varepsilon_m \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_{m-1} & \varepsilon_{m-1} \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix}.
\]

By (6.1) and (6.5) for each \( m \geq 0 \) we get that
\[
(6.6) \quad g_m = \begin{pmatrix} x_m & y_m \\ x_{m-1} & y_{m-1} \end{pmatrix} = \begin{pmatrix} q_m & p_m - \alpha q_m \\ q_{m-1} & p_{m-1} - \alpha q_{m-1} \end{pmatrix}.
\]

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 \( \frac{r}{s} \neq \frac{p}{q} \) we have
\[
(6.7) \quad F_t(q_m, p_m - \alpha q_m) < F_t(s, r - \alpha s).
\]

Conversely, suppose that \( \frac{r}{s} \) is a best approximation to \( \alpha \) 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 \( \alpha \) with respect to \( F \).

That the continued fraction converges to \( \alpha \) 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. Since \( \alpha \) is irrational we need only show that \( \varepsilon_{m+1} + a_m \geq 1 \) for all \( m \geq 1 \). This will follow once we relate the \( \mu_m, \nu_m \) 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 \( \mu_m, \nu_m \) from (3.5) we have for \( m \geq 0 \) that

\[
\mu_m = -\frac{y_m}{y_{m-1}} \quad \text{and} \quad \nu_m = \frac{x_{m-1}}{x_m}.
\]

**Proof.** The proof is an adaptation to more general continued fractions of standard arguments used for regular continued fractions (see [49]).

To start with, by (6.6)

\[
-\frac{y_m}{y_{m-1}} = \frac{-p_m + \alpha q_m}{p_{m-1} - \alpha q_{m-1}}.
\]

By (6.5) and (6.2) we have

\[
q_{m+1}p_m - p_{m+1}q_m = \gamma_{m+1}.
\]

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

\[
\frac{p_m - q_m}{q_{m+1}} = \frac{\gamma_{m+1}}{a_{m+1}q_m + \varepsilon_{m+1}q_{m-1}} \quad \text{where} \quad \frac{p_{m+1}}{q_{m+1}} = a_0 + \frac{\varepsilon_1}{a_1 + \varepsilon_2} + \frac{\varepsilon_3}{a_2 + \varepsilon_3} + \cdots + \frac{\varepsilon_{m+1}}{a_m + \varepsilon_{m+1}}.
\]

The \( m^{th} \) complete quotient \( \alpha_m \) of the expansion \( \alpha = \frac{\varepsilon_1}{a_1 + \varepsilon_2} + \frac{\varepsilon_3}{a_2 + \varepsilon_3} + \cdots \) is defined recursively by \( \alpha_0 = \alpha \) and for \( m \geq 0 \) through

\[\alpha_{m+1} = \frac{\varepsilon_{m+1}}{a_m - a_m}.
\]

It follows that for \( m \geq 0 \) we have

\[
\alpha = a_0 + \frac{\varepsilon_1}{a_1 + \varepsilon_2} + \frac{\varepsilon_3}{a_2 + \varepsilon_3} + \cdots + \frac{\varepsilon_{m+1}}{a_m + \varepsilon_{m+1}}.
\]

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 \( \alpha_{m+1} \) and use (6.10) with \( m \) in place of \( m + 1 \) to get

\[
\alpha_{m+1} = \frac{\varepsilon_{m+1}(-p_{m-1} + q_m - \alpha)}{p_m - q_m \alpha}.
\]

From (6.12) we have

\[
\alpha_{m+1} = a_{m+1} + \frac{\varepsilon_{m+2}}{a_{m+2} + \varepsilon_{m+3}} + \cdots.
\]
so by (3.5)

\[ \mu_m = \frac{\varepsilon_{m+1}}{\alpha_{m+1}}. \]  

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. \( \Box \)

We now finish the proof of Theorem 5 by showing that our expansion is semi-regular. 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 \( \alpha_m < 1 \) and so from (6.15) that

\[ |\mu_{m-1}| = |\frac{y_{m-1}}{y_{m-2}}| > 1, \]

which is impossible. This completes the proof of Theorem 5. \( \Box \)

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 \( \mu_m, \nu_m \). As usual, we may identify the space of all lattices of determinant one with \( \Gamma \setminus G \) where \( G = SL(2, \mathbb{R}) \) and \( \Gamma = SL(2, \mathbb{Z}) \) by means of

\[ g = (x \ y \ x' \ y') \mapsto L(g) \overset{\text{def}}{=} (x, y) \mathbb{Z} + (x', y') \mathbb{Z}. \]

Let \( D \) be the set of \( g = (x \ y \ x' \ y') \in G \) such that

\[ F(x, y) = F(x', y') \]  

and

\[ 0 \leq x' < x \quad \text{and} \quad |y| < y'. \]

For \( g \in D \) let \( F(g) = F(x, y) \).

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

\[ \Phi \left( \frac{x}{x'}, \frac{y}{y'} \right) = \left( \frac{-y}{y'}, \frac{x}{x'} \right) \]

is a continuous bijection.

**Proof.** The inverse of \( \Phi \) is given by

\[ (u, v) \mapsto \frac{1}{\sqrt{1+uv}} \left( \begin{array}{c} t^{-1} \ v \\ -ut \ t \end{array} \right). \]

By Lemma A.3 we see that \( t = t(u, v) > 0 \) exists and is uniquely determined by the condition \( F_t(1, -u) = F_t(v, 1) \). \( \Box \)

The function

\[ D_F(u, v) \overset{\text{def}}{=} F^2(\Phi^{-1}(u, v)) \]

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 $F$ strongly symmetric norm. For any irrational $\alpha$ whose continued fraction associated to the norm is (3.1) we have that

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

where $\mu_m, \nu_m$ are given above in (3.5).

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

$$\Phi^{-1}(\mu_m, \nu_m) = \left( \frac{x'}{x}, \frac{y'}{y} \right)$$

where $0 \leq x' < x$ and $|y| < y'$. Recall that by (i) of Lemma 5.2 we know that $0 \leq x_{m-1} < x_m$ and $|y_m| < |y_{m-1}|$.

Lemmas 7.1 and 6.2 now imply that

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

where $\gamma_m = \pm 1$ was defined in (6.2). Note that in this case $\gamma_m = \text{sgn} y_{m-1}$. 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

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

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

$$\delta_F(\alpha) \leq \limsup_{t \to \infty} \|F_{m(t)}\|_t^2,$$

and by Lemma A.5 we have that

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

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 $1 \leq p < \infty$ 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

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

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

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

giving (3.6). The case $p = \infty$ is immediate. This completes the proof of Theorem 6. \qed
8. 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 \( \alpha \) that are also convergents of the continued fraction of \( \alpha \) 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 \( \alpha \). 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 \( n \geq 0 \) there are integers \( c_\ell \) and \( d_\ell \) with \( c_\ell > 0 \) and \( d_\ell \geq 0 \) so that for each \( \ell \geq 0 \)

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

where \( c_\ell \geq d_\ell \) for all \( \ell \geq 0 \), while for \( \ell \geq 2 \) we have

\[
c_\ell \geq d_\ell + 1.
\]

**Proof.** The integers \( r_n, s_n \) are determined recursively for \( n \geq 0 \) by

\[
\begin{align*}
r_n &= b_n r_{n-1} + r_{n-2}, & r_2 &= 0, & r_1 &= 1, \\
s_n &= b_n s_{n-1} + s_{n-2}, & s_2 &= 1, & s_1 &= 0.
\end{align*}
\]

(8.1) (8.2)

It is easy to check using (8.1) and (8.2) that \( c_\ell \) and \( d_\ell \) satisfy for fixed \( n \) and \( \ell \geq 1 \) the recurrence relations

\[
\begin{align*}
c_{\ell+1} &= b_n + c_{\ell-1}, & c_1 &= b_n + 1, & c_0 &= 1, \\
d_{\ell+1} &= b_n + d_{\ell-1}, & d_1 &= 1, & d_0 &= 0.
\end{align*}
\]

(8.3) (8.4)

The claim of the lemma follows from a straightforward inductive argument. \( \square \)

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 \( m \geq 1 \) let \( n \) and \( \ell \) 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 \( t \geq 1 \) such that \( F_t(Q_n) = F_t(Q_{n-1}) \), and \( \ell = 1 \) if and only if

\[
F_t(Q_n + Q_{n-1}) \leq F_t(Q_n).
\]

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 know that \( P_{-1} = Q_{-1} \) and 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

\[
c_\ell Q_n = P_m - d_\ell P_{m-1}.
\]

(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

\[
c_\ell F_{t_m}(P_m) \leq F_{t_m}(P_m - d_\ell P_{m-1}) < F_{t_m}(P_m) + d_\ell F_{t_m}(P_{m-1}).
\]

(8.6)

by Lemma A.4. Thus we have

\[
c_\ell < d_\ell + 1.
\]

(8.7)

so by Lemma 8.1 we have that either \( \ell = 0 \) or \( \ell = 1 \).
Now by Lemma A.3 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 by Lemma 5.2 we must have
\[ F_t(Q_n + Q_{n-1}) = F_t(Q_{n+1}) = F_t(P_m) \leq F_t(Q_n). \]

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 $m \geq 0$, 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 strict inequality must hold.

\[ \square \]

**Proof of Theorem 7.** Lemma 8.2 gives instructions for obtaining the sequence of convergents $p_m/q_m$ of $\alpha$ associated to the norm $F$ from the sequence of regular convergents $r_n/s_n$ of $\alpha$, namely
\[ (8.8) \quad \text{omit the regular convergent } \frac{r_n}{s_n} \text{ } (n \geq 1) \iff F_t(Q_n + Q_{n-1}) \leq F_t(Q_n), \]
where $t \geq 1$ is such that $F_t(Q_n) = F_t(Q_{n-1})$, and
\[ (8.9) \quad \text{omit } \frac{r_0}{s_0} \iff \lfloor \alpha \rfloor \text{ is not the nearest integer to } \alpha \iff \alpha \in \left[ \frac{1}{2}, 1 \right] + \mathbb{Z}. \]

We must define a singularity area that encodes both of these instructions. Let $\mathcal{D}$ and $\Phi$ be as in Section 7. For each $g = \left( \begin{array}{c} x \\ y \end{array} \right) \in \mathcal{D}$ we write
\[ P = (x, y) \quad \text{and} \quad P' = (x', y'), \]
and we define
\[ (8.10) \quad \mathcal{S} = \Phi\{ \{ g \in \mathcal{D} : F(P + P') \leq F(P) \} \} \cup \left( \left[ \frac{1}{2}, 1 \right] \times \{0\} \right). \]

The portion of $\mathcal{S}$ that lies on the $u$-axis encodes the rule (8.9). Suppose that $n \geq 1$ and let $\left( \begin{array}{c} x \\ y \end{array} \right) = \Phi^{-1}(u, v_n)$. 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 $F(P + P') \leq F(P)$. It follows that (8.8)–(8.9) are encoded by the rule
\[ (8.11) \quad \text{omit the regular convergent } \frac{r_n}{s_n} \iff (u, v_n) \in \mathcal{S}. \]

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. We have
\[ (8.12) \quad \mathcal{S}_1 = \{ (u, v) \in \left[ \frac{1}{2}, 1 \right] \times [0, 1] ; v \leq 2 - \frac{1}{u} \}, \]
as one can see by reducing the system of equations and inequalities
\[ |x + x'| + |y + y'| \leq |x| + |y| = |x'| + |y'|, \quad x > x' \geq 0, \quad y' > |y|, \quad xy' - yx' = 1 \]
defining $\mathcal{S}_1$. The interior of the set (8.12) agrees with the $S$-region given in [28] for Minkowski’s diagonal continued fraction.
Lemma 8.3. For any strongly symmetric norm \( F \) we have \( S \subseteq S_1 \).

Proof. Since \( S \) and \( S_1 \) are closed sets in the induced topology on \([\frac{1}{2}, 1] \times [0, 1]\), it suffices to show that a dense subset of \( S \) is contained in \( S_1 \). Suppose that \((u, v) \in S\) with \( u \notin \mathbb{Q} \) and \( v \in \mathbb{Q} \), and write

\[
(8.13) \quad u = \frac{1}{b_{n+1}} + \frac{1}{b_{n+2}} + \cdots \quad \text{and} \quad v = \frac{1}{b_n} + \frac{1}{b_{n-1}} + \frac{1}{b_{n-2}} + \cdots + \frac{1}{b_1}
\]

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 \( \alpha \). Since \((u, v) \in S\) 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

\[
F_{tm}(Q_{n-1}) = F_{tm}(Q_{n+1}) \leq F_{tm}(Q_n).
\]

Let \( t \) be such that \( F_t(P_{m-1}) = F_t(P_m) \), where \( F_t \) denotes the 1-norm. By convexity of \( F \), the closed stretched ball \( B(P_m, t_m) \) 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 \( F_t(P) = F_t(P_m) \). Since the \( x \)-coordinate of \( Q_n \) is between \( x_{m-1} \) and \( x_m \) and \( Q_n \) is outside the ball \( B(P_m, t_m) \), we have

\[
F_t(P_m) \geq F_t(Q_m) = F_t(Q_n + Q_{n-1}).
\]

By (8.8) and (8.11) it follows that \((u, v) \in S_1\). \( \square \)

Lemma 8.3, together with the explicit description (8.12), shows that the set \( S \) is a singularization area as defined above Theorem 7. This fact and (8.11) together prove Theorem 7. \( \square \)

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 of the line segment \( u = v \) with \( u, v \in (0, 1) \) that does not intersect \( S \).

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

\[
\delta_2 \left( \frac{1}{2}(-1 + \sqrt{3}) \right) = 1.
\]

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

\[
(8.14) \quad \alpha = \frac{1}{2+} + \frac{1}{1+} + \frac{1}{2+} + \frac{1}{1+} \cdots,
\]

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. The region \( S_2 \) comprises those points \((u, v)\) for which \( u(2 + v) > 1 + 2v \). The points \((u_n, v_n)\) are all outside \( S_2 \), so the 2-continued fraction expansion of \( \alpha \) is the same as the regular continued fraction and thus \((\mu_n, v_n) = (u_n, v_n)\). Since \( D_2(u, v) = D_2(v, u) \), we have

\[
\delta_2(\alpha) = \Delta_2 \max_{k \in \{0, 1\}} \lim_{n \to \infty} D(\mu_{2n+k}, \nu_{2n+k}) = \Delta_2 D(\alpha, 2\alpha) = 1.
\]
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 $\alpha$ well approximable.

Lemma 9.1. Suppose that $\alpha$ 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 $p \in \mathbb{Z}$
\[
|\frac{p}{q} - \alpha| < \frac{\varepsilon}{q^2}.
\]
For such a $q$ let $t = q$ and note that for any $r, s \in \mathbb{Z}$ with $s > 0$
\[
F_t(s, s - \alpha s) = F(t^{-1}s, t(s\alpha - r)) = F(\frac{s}{q}, q(s\alpha - r)) = F(\frac{s}{q}, q(s\alpha - r) + \frac{as}{q})
\]
for some $\sigma$ with $|\sigma| \leq \varepsilon$. By Lemma A.1 if $s \geq q$ we have
\[
F(\frac{s}{q}, sp - rq + \frac{as}{q}) \geq F(1, 0) = 1,
\]
while for $0 < s < q$ we have $F(\frac{s}{q}, sp - rq + \frac{as}{q}) \geq F(0, 1 - \varepsilon)$, 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 $F_t(s, s - \alpha s) \geq 1$ and hence that $\delta_F(\alpha) \geq \Delta$. \qed

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

Lemma 9.2. Suppose that the partial quotients $b_n$ of the regular continued fraction expansion of $\alpha$ are eventually strictly increasing with $n$. Then
\[
\delta_F(\alpha) = \Delta.
\]

Proof. If the regular partial quotients $b_n$ of $\alpha$ are eventually strictly increasing, then for any $\varepsilon > 0$ the points $(u_n, v_n)$ all eventually lie within $\varepsilon$ of the point $(0, 0)$. So by Lemma 8.4, the points $(u_n, v_n)$ are outside $S$ for sufficiently large $n$. Thus
\[
\lim_{n \to \infty} (\mu_n, \nu_n) = \lim_{n \to \infty} (u_n, v_n) = (0, 0).
\]
Finally, $D_F(0, 0) = F^2((\frac{1}{0}, \frac{1}{1})) = 1$, therefore $\delta_F(\alpha) = \Delta$. \qed

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

Proof of Theorem 4. Fix $p \in [1, \infty]$. Throughout the proof let
\[
\alpha = \frac{\varepsilon_1}{a_1} + \frac{\varepsilon_2}{a_2} + \frac{\varepsilon_3}{a_3} + \ldots
\]
denote the $p$-continued fraction expansion of $\alpha \in (0, 1)$ and define $\mu_m$ and $\nu_m$ as in (3.5). By Theorem 6 it suffices to show that for every $\alpha$ we have
\[
\limsup_{m \to \infty} D_p(\mu_m, \nu_m) \geq \begin{cases} 
1 & \text{if } 1 \leq p \leq 2, \\
\frac{1}{16}((\sqrt{5} + 5)(\frac{1}{2}(\sqrt{5} - 1))^p + 1)^{2/p} & \text{if } p > 2,
\end{cases}
\]
and that there is at least one $\alpha$ for which equality holds. In both cases the number on the right-hand side of (10.2) is $\leq 1$. Since $(1 - |u|^p|v|^p)^2 \geq (1 - |u|^p)(1 - |v|^p)$, we have
\[
D_p(u, v) \geq \frac{1}{1 + uv}.
\]
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 $\alpha$ has $\mu_m \geq 0$ for all sufficiently large $m$.

**Lemma 10.1.** If $0 \leq u, v < 1$ then $$D_p(u, v) \geq D_p\left(\frac{u + v}{2}, \frac{u + v}{2}\right),$$ with equality only when $u = v$.

**Proof.** The inequalities

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

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

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

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

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

which follows immediately from Hölder’s inequality. $\square$

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

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

If $p = 2$ then $d_p(x) = 1$ for all $x$. In either case, we can use Lemma 9.2 to find examples of $\alpha$ 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 $(\beta, \beta)$ as $n \to \infty$. By Lemma 8.4 it follows that the sequence $(\mu_m, \nu_m)$ associated to the $p$-continued fraction of $\beta$ also converges to $(\beta, \beta)$. Thus, for $p > 2$ we have $\delta_p(\beta) = D_p(\beta, \beta)$, which is the number in (2.7).

It remains to show that for every $\alpha$ 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

\begin{equation}
\mu_m + \nu_m < 1.18
\end{equation}

for infinitely many $m$.

Proof. Suppose that $\varepsilon_m = 1$ and $a_m \leq \ell$ for $m \geq M$. Then for any $m \geq M + 3$ with $a_{m+1} = \ell$ we have

\begin{align*}
\mu_m &\leq \frac{1}{\ell + 1} + \frac{1}{\ell + 1} + \frac{1}{\ell + 1} + \cdots, \\
\nu_m &\leq \frac{1}{1 + \ell + 1}.
\end{align*}

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.2). For instance, the norms $F_{\text{oct}}^1$ and $F_{\text{oct}}^2$ with regular octagonal unit balls mentioned in §1 can be given in terms of compositions of the 1-norm and the sup-norm. Explicitly,

\begin{equation}
F_{\text{oct}}^1(P) = F^{(\infty)}(Q) \quad \text{and} \quad F_{\text{oct}}^2(P) = (2 - \sqrt{2})F^{(1)}(Q)
\end{equation}

where $Q = (\frac{1}{\sqrt{2}}F^{(1)}(P), F^{(\infty)}(P))$. These formulas, together with Mahler’s computation of the critical determinant of the regular octagon recalled at the end of §4 and Lemma 8.4, lead to the result referred to at the end of §1. The minimum of $\delta_F(\alpha)$ for $F = F_{\text{oct}}^1$ is $\frac{1}{8}(3\sqrt{2} + 2)$, which is attained when $\alpha = \frac{\sqrt{2} - 1}{2 + \frac{1}{2 + \frac{1}{2 + \cdots}}}$. For $F = F_{\text{oct}}^2$ the minimal value is also $\frac{1}{8}(3\sqrt{2} + 2)$, but now this is the value of $\Delta$ 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 $g \in D$ and

\begin{equation}
(F(P)\overline{\mathcal{B}}) \cap L(g) = \{0, \pm P, \pm P'\},
\end{equation}

where 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

\begin{equation}
\Omega \overset{\text{def}}{=} \Phi(\mathcal{R}) \cup \mathcal{A},
\end{equation}

where

\begin{equation*}
\mathcal{A} = \overline{\Phi(\mathcal{R})} \cap \left((-\frac{1}{2}, \frac{1}{2}) \times \{0\}\right).
\end{equation*}

We will show that $(\mu_n, \nu_n) \in \Omega$ for all $n \geq 0$. 
We want to apply the ergodic theory of $S$-expansions as developed in [27]. For that we need to show that $\Omega$ defined by (11.2) coincides with the set

$$\Omega_S = ([0, 1] \times [0, 1] \setminus (S \cup TS)) \cup (M \circ T)S$$

defined in Section 5 of [27], where

$$M(u, v) = \left( \frac{-u}{1+u}, 1-v \right), \quad (u, v) \in TS.$$

The following equivalent description of reduced matrices is helpful.

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

$$\min \left( F(P + P'), F(P - P') \right) > F(P). \quad (11.3)$$

**Proof.** Clearly a matrix $g$ satisfying (11.1) also satisfies (11.3). Suppose $g \in D$ 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 $|a| \geq 2$. \qed

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

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

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

![Figure 6](image_url)

**Figure 6.** The sets $\Omega$, $S$, $S'$, and $S''$ for $p = 2$.

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

$$\left\{ g \in D; \min \left( F(P + P'), F(P - P') \right) = F(P) \right\} \quad (11.4)$$
as the potentially critical matrices. The next lemma describes the boundary of $\Omega$ in terms of the distinguished subset

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

of the potentially critical matrices.

**Lemma 11.2.** The part of the boundary of $\Omega$ that lies in $(-1, 1) \times [0, 1]$ is $\partial \cup \partial' \cup \partial'' \cup A$, where

$$\partial = \Phi(\mathcal{P}),$$

$$\partial' = \left\{ \Phi\left( \left( \begin{array}{c} x + x' y + y' \\ x \\ y \end{array} \right) \right) ; g \in \mathcal{P} \right\},$$

$$\partial'' = \left\{ \Phi\left( \left( \begin{array}{c} x + x' y + y' \\ x' \\ y' \end{array} \right) \right) ; g \in \mathcal{P} \right\}.$$ 

**Proof.** The boundary of $\Omega$ is $A \cup C$, where

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

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

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

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

$$\Phi(g') = \Phi\left( \left( \begin{array}{c} x + x' y + y' \\ x \\ y \end{array} \right) \right).$$

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

$$\Phi(g') = \Phi\left( \left( \begin{array}{c} x + x' y + y' \\ x' \\ y' \end{array} \right) \right).$$

This completes the proof. □

![Figure 7](image-url)
Lemma 11.3. 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 $D(u,v)$. Say that a point $(u,v)$ in $\overline{\Omega} \setminus \{(1,1)\}$ is a critical point if $D(u,v) = 1/\Delta$. For $(u,v) \in \overline{\Omega} \setminus \{(1,1)\}$, let the points $P = (x,y)$ and $P' = (x',y')$ be such that $\Phi^{-1}(u,v) = (x,y)$. Then

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

where $L$ is the lattice generated by the unit vectors $\frac{1}{F(\Phi(P))}P$ and $\frac{1}{F(P')}P'$. Thus $(u,v)$ is a critical point if and only if $L$ is a critical lattice, so by Lemma 11.2 all critical points lie on the boundary of $\Omega$. 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

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

Then the matrix $g = \frac{1}{\sqrt{2}} \begin{pmatrix} 2^{-1} & 0 \\ 0 & 1 \end{pmatrix}$ is in $\overline{\mathcal{R}}$ 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. \hfill \Box

That $\Omega = \Omega_S$ follows from the next lemma.

Lemma 11.4. We have

$$S' = TS,$$

$$S'' = ((-1,0] \times [0,1]) \setminus MS'.$$

Proof. We begin by showing that $S' = TS$. For $(u,v) \in S$ we have $\frac{1}{2} \leq u < 1$, so (2.9) simplifies to $T(u,v) = (\frac{1-u}{u}, \frac{1}{v+1})$. We show that the boundary of $S$ maps to the boundary of $S'$ under $T$; the lemma then follows by the continuity of $T$ on $S$. Since $T([\frac{1}{2},1] \times \{0\}) = [0,1] \times \{1\}$ 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

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

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

We prove $S'' = ((-1,0] \times [0,1]) \setminus MS'$ 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

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

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

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

which completes the proof. \hfill \Box

Let $\omega_S = (1 - \omega(S))^{-1} \omega$, where $\omega$ was defined in (2.10).

Lemma 11.5. Define $S$ and $\Omega$ by (8.10) and (11.2). Then for almost all irrational $\alpha$ the sequence $(\mu_m, \nu_m)$ is uniformly distributed over $\Omega$ with respect to the measure $\omega_S$.

Proof. Since $\Omega = \Omega_S$, Lemma 11.5 follows from Theorem 5.4.23 of [9] (see also [27]) and Theorem 7. \hfill \Box
Proof of Theorem 1. By Lemma 11.3, the function $\Delta D(u,v)$ assumes the value 1 at some point in $\Omega \setminus \{(1,1)\}$. By Lemma 7.2 it follows that $\delta_F(\alpha) = 1$ if and only if the sequence $(\mu_n, \nu_n)$ is infinitely often arbitrarily close to such a critical point. It follows from Lemma 11.5 that $\delta_F(\alpha) = 1$ for almost all $\alpha$.

To finish the proof it suffices to show that there are uncountably many such $\alpha$. 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 $\alpha$ for which $(\mu_n, \nu_n)$ converges to $(0,0)$ has $\delta_F(\alpha) < 1$, and there are uncountably many such $\alpha$ (for example, the set of $\alpha$ 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 $S$-expansions and ergodic theory to quantities related to $\delta_p(\alpha)$. For instance we may treat the distribution of the values of

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

from Theorem 6. For almost all $\alpha$ the distribution function

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

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 $\alpha$

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

exists where

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

It is well known that a close connection exists between dynamical systems associated to various kinds of continued fractions and the geodesic flow on $\text{SL}(2, \mathbb{Z}) \backslash \text{SL}(2, \mathbb{R})$. 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 $\omega$ 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

$$\lambda(\alpha) = \liminf_{t \geq 1} \rho_t(\alpha), \quad \text{where} \quad \rho_t(\alpha) = t \min_{p \in \mathbb{Z}} \frac{|p - \alpha q|}{1 \leq q \leq t} \text{ for } t \geq 1.$$
The Dirichlet spectrum is determined by the values of \( \delta(\alpha) = \limsup_{t \geq 1} \rho_t(\alpha) \); in [22] 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 [54], Oppenheim [41] and Burger [6]. This problem in higher dimensions has also attracted a lot of attention (see e.g. [44, 45, 53]).

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 [42] 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 simple technical lemmas that are referred to in the body of the paper. Here \( F \) is a norm on \( \mathbb{R}^2 \) with unit ball \( B \) 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 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.1.** Suppose that \( F \) satisfies (2.4). If \( |x'| \leq |x| \) and \( |y'| \leq |y| \) then we have that \( F(P') \leq F(P) \).

**Proof.** To see this observe that if \( F(P) = s \) then \( F(\pm x, \pm y) = s \) hence \( F(x', y') \leq s \) by convexity. \( \square \)

**Lemma A.2.** 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.1. \( \square \)

**Lemma A.3.** Suppose that \( F \) satisfies (2.4). The following properties hold.

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

\[
F_t(P') = F_t(P).
\]

(ii) If \( F(P') \geq F(P) \) 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 $f_P : [1, \infty) \rightarrow \mathbb{R}^+$ by $f_P(t) = t^{-1}F_t(P)$. Now by Lemma A.1

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

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

Uniqueness: Suppose that for $t_1 \neq t_2$ with $t_1, t_2 \geq 1$ we have

$$F_{t_1}(P) = F_{t_1}(P') = F_{t_2}(P) = F_{t_2}(P').$$

This implies that $|x| = |x'|$ and that $|y| = |y'|$, which is not true. \hfill \Box

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

**Lemma A.4.** 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 $d \geq 1$

$$F(P - dP') < F(P) + dF(P').$$

**Proof.** To see this note first that in order for equality to hold in (A.1) we must have that

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

which implies that

$$F(P - P') \geq F(P) + F(P')$$

so that $F(P - P') = F(P) + F(P')$

hence

$$F\left(\frac{1}{2}(P - P')\right) = \frac{1}{2}(F(P) + F(P')) = F(P) = F(-P').$$

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). \hfill \Box

**Lemma A.5.** Suppose that 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.1 we get that

$$F_{\sigma t_1 + \sigma' t_2}(x, y) \leq F\left(x(\frac{\sigma}{t_1} + \frac{\sigma'}{t_2}), y(\sigma t_1 + \sigma' t_2)\right).$$

By the defining properties of a norm we finish the proof. \hfill \Box

**References**


[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.


[22] Lagrange, J.L., Additions aux éléments d’algebra d’Euler, Oeuvres VII .


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