ACTA ARITHMETICA
LXIV.1 (1993)

On numbers with a unique representation by a binary quadratic form

by

Mariusz Skałba (Warszawa)

We present a generalization of Davenport's constant and give some number-theoretic application of this notion.

In Section 1 we define the relative Davenport constant $D_a(A)$ and prove some basic theorems about it. In particular, we calculate the Davenport constant with respect to any element of a cyclic group and of a p-group.

The main result of Section 2 is the following theorem:

Let $F(x,y)$ be a quadratic form with nonsquare discriminant D and conductor f. If a natural number n, relatively prime to f, is uniquely representable by F then

$$n = r(n)s(n)$$

where $r(n)$ is a squarefree divisor of D relatively prime to f, $s(n)$ is relatively prime to D and

$$\Omega(s(n)) \le D_{[F]^2}(C(D)^2)$$

where $C(D)$ is the corresponding form class group and $\Omega(s(n))$ is the number of prime factors of $s(n)$, counted with multiplicities.

We also obtain an asymptotic formula for the number $N_F(x)$ of natural numbers not greater than x, relatively prime to f and uniquely representable by the form F:

$$N_F(x) = (C_F + o(1))\frac{x}{\log x}(\log\log x)^{D_{[F]^2}(C(D)^2)-1}$$

where $C_F > 0$.

I wish to express my sincere gratitude to Prof. J. Browkin for his help in writing this paper and many valuable suggestions. I am also greatly indebted to the referee for his critical remarks and several improvements.

1. We start with the basic definitions. A sequence $a_1, \ldots, a_k$ will be called *irreducible* provided no sum of less than k of its distinct elements vanishes. If in addition $a_1 + \ldots + a_k \neq 0$, then this sequence will be called *primitive*.

For any finite Abelian group A and a in A we define $D_a(A)$, the *relative Davenport constant of A with respect to a*, as the greatest integer k with the property that a can be written as the sum of k elements of A forming an irreducible sequence.

For $a = 0$ we have $D_0(A) = D(A)$ ([3]).

We need the following easy lemmas:

LEMMA 1. *Let A be an Abelian group, and $\mathcal{A} = (a_1, \ldots, a_k)$ a sequence of its elements. The following conditions are equivalent*:

(i) *$\mathcal{A}$ is primitive*,

(ii) *$\mathcal{A}' = (a_1, \ldots, a_k, -\sum_{i=1}^{k} a_i)$ is irreducible.* ■

LEMMA 2. *If $\mathcal{A} = (a_1, \ldots, a_k)$ is a maximal primitive sequence in A, then every element of A is a sum of elements of $\mathcal{A}$.* ■

First of all we get the following general estimate.

THEOREM 1. *If A is a finite Abelian group and $a \in A$, $a \neq 0$, then*

$$\tfrac{1}{2}D(A) \leq D_a(A) < D(A)\,.$$

Proof. Let $\mathcal{A} = (a_1, \ldots, a_k)$ be an irreducible sequence with sum a. Since $a \neq 0$, therefore $\mathcal{A}$ is primitive. From Lemma 1 we see that $\mathcal{A}' = (a_1, \ldots, a_k, -a)$ is also irreducible. Hence $D_a(A) < D(A)$.

To prove the estimate from below fix any primitive sequence $\mathcal{A} = (a_1, \ldots \ldots, a_k)$ with $k = D(A) - 1$. By Lemma 2 we have $a = \sum_{i \in X} a_i$ for some X.

If $|X| \geq (k+1)/2$ we define $\mathcal{A}' = (a_j)_{j \in X}$. Then $\mathcal{A}'$ has sum a, is irreducible (even primitive), and therefore

$$D_a(A) \geq \frac{k+1}{2} = \frac{D(A)}{2}\,.$$

If $|X| < (k+1)/2$ we proceed otherwise. Let

$$Y = \{1, \ldots, k+1\} - X, \qquad a_{k+1} = -\sum_{i=1}^{k} a_i$$

and consider the sequence $\mathcal{A}'' = (-a_j)_{j \in Y}$. It has sum a and is irreducible by Lemma 1, hence

$$D_a(A) \geq k + 1 - \frac{k+1}{2} = \frac{D(A)}{2}\,. \ ■$$

LEMMA 3. *Let A be a finite Abelian group, B a subgroup of A and $a \in B$. Then*

$$D_a(A) \leq D_a(B) \cdot D(A/B)\,.$$

Proof. Consider an irreducible sequence $\mathcal{A} = (a_1, \dots, a_n)$ with sum a. We may represent the set $\{1, \dots, n\}$ as the sum of disjoint subsets $A_1, \dots, A_t$ $(t \ge 1)$ such that

$$\forall 1 \le j \le t, \quad \sum_{i \in A_j} a_i \in B \quad \text{and} \quad \forall \emptyset \ne A \subsetneq A_j, \quad \sum_{i \in A} a_i \notin B.$$

Then $|A_j| \le D(A/B)$. If we put

$$b_j = \sum_{i \in A_j} a_i \qquad (j = 1, \dots, t)$$

then the sequence $b_1, \dots, b_t$ has sum a and is irreducible, hence $t \le D_a(B)$ and our assertion follows. ∎

THEOREM 2. *If A is a finite cyclic group and $a \in A$ then*

$$D_a(A) = \begin{cases} |A| & \text{for } a = 0, \\ |A| - |A|/|a| & \text{for } a \ne 0 \end{cases}$$

($|a|$ denotes the order of a).

Proof. Let $A = \mathbb{Z}_n$ $(n > 1)$. The case $a = 0$ is well known. Assume $a \ne 0$. We use Lemma 3 for $B = \langle a \rangle$:

$$D_a(A) \le D_a(\langle a \rangle) \cdot D(A/\langle a \rangle) = D_a(\langle a \rangle) \cdot \frac{n}{|a|}\,.$$

Consider the sequence

$$\mathcal{A} = (-a, -a, \dots, -a)$$

with $|a| - 1$ terms. $\mathcal{A}$ is primitive and has sum a. Hence

$$D_a(\langle a \rangle) \ge |a| - 1\,.$$

On the other hand, any irreducible sequence with sum a is primitive $(a \ne 0)$ and hence its length is less than $D(\langle a \rangle) = |a|$. This gives the equality

$$D_a(\langle a \rangle) = |a| - 1\,.$$

From the above,

$$D_a(A) \le n - \frac{n}{|a|}\,.$$

To get equality it suffices to construct an irreducible sequence $\mathcal{A}$ with sum a and length $n - n/|a|$. Using an automorphism of $A = \mathbb{Z}_n$ if necessary, we may assume that

$$a = \frac{n}{|a|} \bmod n$$

and then the sequence

$$\mathcal{A} = (-1 \bmod n, -1 \bmod n, \dots, -1 \bmod n)$$

meets our demand. ∎

Now we deduce from [5] a formula for $D_a(A)$ in case of p-groups. We need the following technical definition: Let A be a finite Abelian p-group. For any $a \in A$ let

$$\alpha(a) = p^n$$

where n is the greatest nonnegative integer such that

$$a = b^{p^n}$$

for $b \in A$ $(\alpha(1) = \infty)$.

THEOREM 3. *If* $A \cong \prod_{i=1}^{r} C_{p^{e_i}}$ *(where* C_n *denotes the cyclic group of order* n*),* $r \geq 1$, $e_i \geq 1$, *then for every nonzero* a *in* A *we have*

$$D_a(A) = D(A) - \alpha(a)\,.$$

Proof. We write the group A multiplicatively. If $a \neq 1$ and $a = a_1 \ldots a_k$ with $(a_1, \ldots, a_k)$ irreducible then Lemma 1 implies that the sequence
$(a_1, \ldots, a_k, a^{-1})$ is irreducible and $(a_2, \ldots, a_k, a^{-1})$ is primitive. Thus the product $(1-a_2)\ldots(1-a_k)(1-a^{-1})$ in the group ring $\mathbb{Z}_p[A]$ is nonzero and Theorem 2 of [5] implies

$$\sum_{i=2}^{k} \alpha(a_i) + \alpha(a^{-1}) < D(A)\,.$$

We have $\alpha(a_i) \geq 1$ for $i = 2, \ldots, k$ $(a_i \neq 1)$ and $\alpha(a^{-1}) = \alpha(a)$, therefore

$$(k-1) + \alpha(a) < D(A)\,,$$

hence

$$D_a(A) \leq D(A) - \alpha(a)\,.$$

To finish the proof it suffices to construct a primitive sequence $\mathcal{A}$ with product a and length $D(A) - \alpha(a)$. Let $b \in A$ be such that

$$b^{\alpha(a)} = a\,.$$

The element b generates a maximal cyclic subgroup of A, therefore using possibly an automorphism of A we can write

$$\exists 1 \leq i \leq r, \quad b = (1, \ldots, x_i, \ldots, 1) \quad \text{where } x_i \text{ generates } C_{p^{e_i}}\,.$$

Define

$$\mathcal{A} = \Big(\varepsilon_1, \ldots, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_2, \ldots, \varepsilon_i^{-1}, \ldots, \varepsilon_i^{-1}, \ldots, \varepsilon_r, \ldots, \varepsilon_r, \Big(\prod_{j=1}^{r} \varepsilon_j\Big)\varepsilon_i^{-2}\Big)$$

where for $j \neq i$, $\varepsilon_j := (1, \ldots, x_j, \ldots, 1)$ appears $p^{e_j} - 1$ times and ε_i^{-1} $(= b^{-1})$ appears $p^{e_i} - \alpha(a) - 1$ times. ∎

COROLLARY. *If* $A \cong \prod_{i=1}^r C_{p^{e_i}}$, $r \ge 1$, $e_i \ge 1$, *then*

$$D_a(A) = \begin{cases} \sum_{i=1}^{r} (p^{e_i} - 1) + 1 & \text{for } a = 0, \\ \sum_{i=1}^{r} (p^{e_i} - 1) + 1 - \alpha(a) & \text{for } a \neq 0. \end{cases}$$

Proof. By Theorem 1 of [5]. ∎

2. Let $F(x,y)$ be a binary quadratic form, positive if definite, corresponding to a class X of invertible ideals in an order $\mathcal{O}_f$ in a suitable quadratic field K. The classical theory of quadratic forms ([1], [2]) shows that if we choose an arbitrary invertible ideal I in X under the unique restriction that in case $X^2 = E$, the unit class, the ideal I should be ambiguous, i.e. $\bar{I} = I$, then one can choose a $\mathbb{Z}$-basis a, b of I such that

$$F(x,y) = N(ax - by)/N(I)\,.$$

Thus we have

$$F(x,y) = n$$

with $(x,y) = 1$ if and only if there is a principal ideal A with

$$N(A) = nN(I), \qquad A \subseteq I\,, \tag{1}$$

which has no rational divisor > 1. (Actually $A = (ax - by)\mathcal{O}_f$.)

We shall say that n *is uniquely representable by the form* F provided the ideal A in (1) is unique in the case $X^2 \neq E$, and unique up to conjugacy in the case $X^2 = E$.

LEMMA 4. *Let X be the class of the ideal I, assume $(n,f) = 1$ and let A, B be distinct, principal and moreover, in the case $X^2 = E$, nonconjugate ideals satisfying* (1). *Write*

$$A = I \cdot D_1 \cdot P_1 \cdot \ldots \cdot P_s, \qquad B = I \cdot D_2 \cdot Q_1 \cdot \ldots \cdot Q_t$$

where D_j are ideals without unramified prime ideal divisors, and P_i, Q_j are unramified prime ideals in $\mathcal{O}_f$; finally, let a_i be the class of P_i and b_j be the class of Q_j. Then with suitable i_j and $r < s$ we have

$$(a_{i_1} \cdot \ldots \cdot a_{i_r})^2 = E\,. \tag{2}$$

The converse is also true.

Proof. Obviously we have $s = t$ and after a suitable regrouping we can assume that Q_j either equals P_j or is conjugate to it. Assume that the first possibility happens for $j = 1, \ldots, w$. Then

$$b_j = a_j \qquad (j = 1, \ldots, w)$$

and
$$b_j = a_j^{-1} \qquad (j = w+1, \ldots, s)\,.$$
Since $a_1 \cdot \ldots \cdot a_s = b_1 \cdot \ldots \cdot b_s$ we get (2) with $r = s - w$, and it remains to show that w is positive.

Note that D_1, D_2 are both products of distinct ramified prime ideals, since otherwise A resp. B would have a nontrivial rational factor. In view of $N(D_1) = N(D_2)$ this implies $D_1 = D_2$ and so $D_1 D_2$ must be principal.

If $w = 0$, then
$$AB = I^2 D_1 D_2 J \qquad \text{where } J \text{ is principal,}$$
showing that I^2 is principal. But in this case our assumptions give $I = \bar{I}$ and this immediately implies that A and B are conjugate.

To prove the converse we proceed very similarly. After a suitable regrouping we can assume that
$$(a_1 \cdot \ldots \cdot a_r)^2 = E\,.$$
Now if we take
$$D_2 = D_1, \quad Q_i = \overline{P}_i \quad \text{for } i = 1, \ldots, r \quad \text{and} \quad Q_i = P_i \quad \text{otherwise}\,,$$
then $B \neq A$, since equality would imply $P_1 \cdot \ldots \cdot P_s = \overline{P}_1 \cdot \ldots \cdot \overline{P}_s$ and hence A would have a rational divisor > 1. Moreover, $B \neq \overline{A}$ in the case $X^2 = E$, since equality would imply $P_{r+1} \cdot \ldots \cdot P_s = \overline{P}_{r+1} \cdot \ldots \cdot \overline{P}_s$ which also contradicts the assumptions. ∎

The following theorem is an easy consequence of the above lemma and the definition of the relative Davenport constant:

THEOREM 4. *Let $F(x, y)$ be a form with nonsquare discriminant D and conductor f. If a natural number n, relatively prime to f, is uniquely representable by F then*
$$n = r(n)s(n)$$
where $r(n)$ is a squarefree divisor of D relatively prime to f, $s(n)$ is relatively prime to D and
$$\Omega(s(n)) \leq D_{[F]^2}(C(D)^2)$$
where $[F]$ denotes the class of the form F in the form class group $C(D)$. ∎

COROLLARY. *Let d be a natural number, $d \geq 4$. Moreover, let f be the conductor of the form $F(x, y) = x^2 + dy^2$. If a natural number $x \in [1, \sqrt{3d})$ is such that $(x^2 + d, f) = 1$ then either*
$$x^2 + d = t^2 \qquad \textit{for some } t \in \mathbb{N}$$
or
$$x^2 + d = rs$$

where r is a squarefree divisor of $4d$, $(s,4d)=1$ *and*

$$\Omega(s)\le D(C(-4d)^2)\,.$$

P r o o f. Let $n=x^2+d$ for some $x\in[1,\sqrt{3d})$ and assume that $(n,f)=1$. We have

$$n<3d+d=4d$$

and

$$F(x,y)\ge 4d \qquad \text{for } |y|\ge 2,$$

therefore if $n\ne t^2$ then n is uniquely representable by F. Now the assertion results from Theorem 4. ∎

Example. Let $d=5005=5\cdot 7\cdot 11\cdot 13$. Since d is squarefree and $d\equiv 1 \pmod 4$, therefore the conductor f of the form $F(x,y)=x^2+5005y^2$ is 1. Hence for each $x\in[1,122]$,

$$x^2+5005=t^2 \quad\text{or}\quad x^2+5005=rs$$

where $r\mid 10010$, $(s,10010)=1$ and $\Omega(s)\le 2$.

Theorem 5. *Let $F(x,y)$ be a form with discriminant $D<0$ and conductor f. For $x\ge 1$, let $N_F(x)$ denote the number of natural numbers n, not greater than x, relatively prime to f and uniquely representable by F. Then there exists a positive constant C_F such that the following asymptotic equality holds*:

$$N_F(x)=(C_F+o(1))\frac{x}{\log x}(\log\log x)^{D_{[F]^2}(C(D)^2)-1}\,.$$

Moreover, *let*

$$\overline{N}_F(x)=|\{n\in\mathbb{N}:n\le x,(n,f)=1, n \textit{ is uniquely representable by } F \\ \textit{and } \Omega(s(n))=D_{[F]^2}(C(D)^2)\}|\,.$$

Then

$$\lim_{x\to\infty}\frac{\overline{N}_F(x)}{N_F(x)}=1\,.$$

P r o o f. First let us recall some useful definitions. Let X be a set of ideals of the ring $\mathcal{O}_F$, and for each ideal $I<\mathcal{O}_F$ let $\Omega_X(I)$ be the number of prime ideals from X appearing in the decomposition of I into prime factors (counted with multiplicities). If A is a set of prime ideals and

$$\sum_{\mathfrak{p}\in A}N(\mathfrak{p})^{-s}=a\log\frac{1}{s-1}+g(s) \qquad \text{for } \operatorname{Re}s>1$$

where $g(s)$ is regular in the halfplane $\operatorname{Re}s\ge 1$ then A is called a *regular set of prime ideals*; the number a is called the *Dirichlet density* of A.

Lemma 5. *Let $\mathcal{O}_f$ be an order of an imaginary quadratic field K. Let X be a given class of invertible ideals in $\mathcal{O}_f$, and A_X the set of prime ideals in X relatively prime to f. Then the set*

$$\mathcal{A}_X := \{\mathfrak{p} \cdot \mathcal{O}_K : \mathfrak{p} \in A_X\}$$

is regular.

Proof. The assertion follows from the proof of Theorem 9.12 of [2], pp. 188–189. ■

Let A denote the set of all irreducible sequences of the group $C(\mathcal{O}_f)^2$ with product $[I]^{-2}$, where two sequences differing only in the order of terms are considered identical. Let R be the product of all primes dividing D and relatively prime to f, and r a fixed divisor of R. Moreover, let $\mathcal{R}$ be the product of prime ideals of $\mathcal{O}_f$, dividing r.

For each $\mathcal{A} = (\alpha_1, \ldots, \alpha_k) \in A$ we define

$$\mathcal{A}(r) = \Big\{\mathcal{B} = (\beta_1, \ldots, \beta_k) : \beta_i \in C(\mathcal{O}_f),\ \beta_i^2 = \alpha_i \quad \text{for } i = 1, \ldots, k \text{ and } \prod_{i=1}^{k} \beta_i = [I]^{-1}[\mathcal{R}]^{-1}\Big\}.$$

First we prove that for any $\mathcal{A} \in A$,

$$\mathcal{A}(r) \neq \emptyset. \tag{$*$}$$

Let $\mathcal{B}' = (\beta_1', \ldots, \beta_k')$ be an arbitrary sequence of elements of $C(\mathcal{O}_f)$ such that $\beta_i'^2 = \alpha_i$ for $i = 1, \ldots, k$. Since

$$\prod_{i=1}^{k} \beta_i'^2 = \prod_{i=1}^{k} \alpha_i = ([I]^{-1}[\mathcal{R}]^{-1})^2 \qquad ([\mathcal{R}]^2 = 1),$$

there exists $\beta' \in C(\mathcal{O}_f)$ such that $\beta'^2 = 1$ and

$$\beta' \cdot \prod_{i=1}^{k} \beta_i' = [I]^{-1}[\mathcal{R}]^{-1}.$$

Hence

$$\mathcal{B} := (\beta'\beta_1', \beta_2', \ldots, \beta_k') \in \mathcal{A}(r),$$

which ends the proof of $(*)$.

Define

$$\mathcal{U} = \{n \in \mathbb{N} : (n, f) = 1,\ n \text{ is uniquely representable by } F\}$$

and for each $r \mid R$ let

$$\mathcal{U}(r) = \{n \in \mathcal{U} : r(n) = r\}$$

(with $r(n)$ from Theorem 4). Clearly

$$\mathcal{U} = \bigcup_{r \mid R} \mathcal{U}(r)\,.$$

Hence

$$(**) \qquad N_F(x) = \sum_{r \mid R} N_F^{(r)}(x)$$

where

$$N_F^{(r)}(x) := |\{n \in \mathcal{U}(r) : n \le x\}|\,.$$

We first obtain an asymptotics for $N_F^{(r)}(x)$ at a fixed $r \mid R$ and then use $(**)$. Let

$$h = |C(\mathcal{O}_f)|, \qquad C(\mathcal{O}_f) = \{\gamma_1, \ldots, \gamma_h\},$$
$$\Pi_i = \{\mathfrak{p} \cdot \mathcal{O}_F : \mathfrak{p} \text{ a prime ideal of } \mathcal{O}_f, (N(\mathfrak{p}), f) = 1 \text{ and } [\mathfrak{p}] = \gamma_i\}\,.$$

For each sequence $\mathcal{B} = (\beta_1, \ldots, \beta_n)$ of elements of $C(\mathcal{O}_f)$ let

$$\Omega_{\Pi_i}(\mathcal{B}) := |\{j \in 1, \ldots, n : \beta_j = \gamma_i\}|\,.$$

We define

$$\mathcal{J}(r) = \bigcup_{\mathcal{A} \in A} \bigcup_{\mathcal{B} \in \mathcal{A}(r)} \{J \cdot \mathcal{O}_K : J < \mathcal{O}_f, (N(J), f) = 1 \text{ and}$$
$$\Omega_{\Pi_i}(J \cdot \mathcal{O}_K) = \Omega_{\Pi_i}(\mathcal{B}) \text{ for } i = 1, \ldots, h\}\,.$$

From the above definitions and Lemma 4 it follows that the map $\mathcal{N} : \mathcal{J}(r) \to \mathbb{N}$ given by the formula

$$\mathcal{N}(J) := N(J) \cdot r$$

maps $\mathcal{J}(r)$ onto $\mathcal{U}(r)$ and moreover, for all but finitely many $n \in \mathcal{U}(r)$,

$$(***) \qquad |\mathcal{N}^{-1}(n)| = \begin{cases} 1 & \text{if } X^2 \ne E, \\ 2 & \text{if } X^2 = E. \end{cases}$$

By Lemma 5 and Proposition 9.6 in Ch. 9 of [4],

$$|\{J \in \mathcal{J}(r) : \mathcal{N}(J) \le x\}| = \sum_{\mathcal{A} \in A} \sum_{\mathcal{B} \in \mathcal{A}(r)} (C_{\mathcal{B}} + o(1)) \frac{\frac{x}{r}}{\log \frac{x}{r}} \left(\log\log \frac{x}{r} \right)^{l(\mathcal{B})-1},$$

hence by $(*)$ and the definition of the relative Davenport constant,

$$|\{J \in \mathcal{J}(r) : \mathcal{N}(J) \le x\}| = (C_r' + o(1)) \frac{x}{\log x} (\log\log x)^{D_{[I]^2}(C(\mathcal{O}_f)^2)-1}\,.$$

From $(***)$ and the above formula,

$$N_F^{(r)}(x) = (C_r + o(1)) \frac{x}{\log x} (\log\log x)^{D_{[I]^2}(C(\mathcal{O}_f)^2)-1}\,.$$

To obtain the first part of the assertion of Theorem 5 it suffices to use $(**)$. The second part, concerning the function $\overline{N}_F(x)$, is now obvious. ■

References

[1] Z. I. Borevich and I. R. Shafarevich, *Number Theory*, Nauka, Moscow 1985 (in Russian).
[2] D. A. Cox, *Primes of the Form* $x^2 + ny^2$, Wiley-Interscience, New York 1989.
[3] H. Davenport, in: Proceedings of the Midwestern Conference on Group Theory, Ohio State University, April 1966.
[4] W. Narkiewicz, *Elementary and Analytic Theory of Algebraic Numbers*, PWN, Warszawa 1990.
[5] J. E. Olson, *A combinatorial problem on finite Abelian groups*, *I*, *II*, J. Number Theory 1 (1969), 8–10, 195–199.

INSTITUTE OF MATHEMATICS
WARSAW UNIVERSITY
BANACHA 2
02-097 WARSZAWA, POLAND

Received on 10.3.1992
and in revised form on 5.6.1992 (2242)