

Davenport's constant for indefinite binary quadratic forms

by

JANE PITMAN (Cambridge)

1. Introduction. Let f(x, y) be an indefinite binary quadratic form with real coefficients and discriminant D = D(f), and write $\Delta = \Delta(f) = \sqrt{(D(f))}$. If P = (x', y') is any real point, let $M(f; P) = M(f; x', y') = \inf[|f(x+x', y+y')|; x, y \text{ integral}]$; the inhomogeneous minimum, M(f), of f is now defined by

$$M(f) = \sup_{P} M(f; P),$$

where the supremum is taken over all real points P. Davenport [4] showed that there exists a constant k such that, for all f,

$$M(f) > k\Delta(f)$$

(and that the statement is true for k=1/128 [5]). We may therefore define an absolute constant K by

$$K = \sup[k; M(f) > k\Delta(f)],$$

where the supremum is taken over all forms f.

Ennola [6] improved on previous lower bounds for K by showing that

$$K \geqslant 1/(16+6\sqrt{6}) = 1/30 \cdot 69...$$

In the opposite direction, I had shown [7] that

$$K \leqslant 1/12$$
,

and it had already been conjectured that the value of K might be 1/12. However, as a result of computations on EDSAC 2 at the Cambridge Mathematical Laboratory, I have found several forms f for which $M(f)/\Delta(f)$ is less than 1/12, and one form F for which

$$M(F)/\Delta(f) = \frac{1}{4} \times 309564... = .077391... = 1/12.921...$$

39

Hence we now have

$$K \leq .077391... = 1/12.921...$$

The results which form the theoretical basis for the computation of M(f) are outlined in § 2; some further ideas which underlie the method of computation are given in § 3; and the organization of the computations for EDSAC 2 is described in § 4. Finally, in § 5, I discuss some of the forms with low minima (in particular, the form F), and the way in which these forms were found.

Mr H. P. F. Swinnerton-Dyer wrote the programs for the computations on EDSAC 2 and helped with the description of them in this paper; I wish to thank him very much for this and for all his help. I am also grateful to Professor H. Davenport and Dr. J. W. S. Cassels for reading the manuscript. I am indebted to the Director of the Cambridge Mathematical Laboratory for the opportunity of using EDSAC 2 for these computations. I am grateful for an 1851 Overseas Scholarship which enabled me to carry out this work in Cambridge.

2. Theoretical basis. The computations were based on the "divided cell method" for obtaining the inhomogeneous minimum of a given form, which was devised by Barnes and Swinnerton-Dyer [2] and extended by Barnes [3]; a brief account of the method is given in [7], § 2, and further results related to it are given in [7], § 3. Here, I shall briefly outline the results needed in the computations.

Let f = (a, b, e) denote an indefinite binary quadratic form which has real coefficients and does not represent zero:

$$f(x, y) = ax^2 + bxy + cy^2,$$

and let

$$\Delta = \Delta(t) = +\sqrt{(b^2-4ac)}$$

Let f be written in the shape

$$f(x,y) = \pm \frac{\Delta}{|\theta \Phi - 1|} (\theta x + y)(x + \Phi y);$$

then f is called inhomogeneously reduced or I-reduced if

$$| heta|>1, \quad | heta|>1,$$

and Gauss-reduced if

$$\theta < -1, \quad \Phi > 1.$$

_

Let

$$f_0(x, y) = \pm \frac{\Delta}{|\theta_0 \Phi_0 - 1|} (\theta_0 x + y)(x + \Phi_0 y)$$

be any I-reduced form equivalent to a given form f. Let $\{a_n\}$ $(-\infty < n < \infty)$ be a chain of integers such that $|a_n| \ge 2$ for each n and

(1)
$$\theta_0 = a_0 - \frac{1}{a_{-1}} - \frac{1}{a_{-2}} \dots, \quad \Phi_0 = a_1 - \frac{1}{a_2 - a_3} - \dots;$$

then $\{a_n\}$ is called an *a-chain* of the form f from the form f_0 . (As continued fraction expansions of the above type are not unique, there will be many such *a*-chains from one form f_0 .) A unique chain of equivalent I-reduced forms $\{f_n\}$ corresponding to $\{a_n\}$ is then defined by

$$f_n(x, y) = \pm \frac{\Delta}{|\theta_n \Phi_n - 1|} (\theta_n x + y)(x + \Phi_n y),$$

where

$$heta_n = a_n - rac{1}{a_{n-1} -} rac{1}{a_{n-2} -} \dots, \quad heta_n = a_{n+1} - rac{1}{a_{n+2} -} rac{1}{a_{n+3} -} \dots$$

Any chain of integers $\{\varepsilon_n\}$ $(-\infty < n < \infty)$ such that, for each n, ε_n has the same parity as a_{n+1} , and

$$|\varepsilon_n| \leqslant |a_{n+1}| - 2$$
,

is called an ε -chain corresponding to $\{a_n\}$. The pair of chains $\{a_n\}$, $\{\varepsilon_n\}$ is then called a *chain-pair* (of the form f from the form f_0).

For a given chain-pair, let

$$\sigma_n = arepsilon_{n-1} + \sum_{r=1}^{\infty} \left(-1
ight)^r rac{arepsilon_{n-r-1}}{ heta_{n-1} heta_{n-2}\dots heta_{n-r}},$$

$$\tau_n = \varepsilon_n + \sum_{r=1}^{\infty} (-1)^r \frac{\varepsilon_{n+r}}{\Phi_{n+1} \Phi_{n+2} \dots \Phi_{n+r}};$$

and, for each n, let

(2)
$$\pi_n =$$

$$= \frac{\Delta}{|\theta_n \Phi_n - 1|} \min \left[\frac{|(1 + \theta_n + \sigma_n)(1 + \Phi_n + \tau_n)|, \quad |(-1 + \theta_n + \sigma_n)(1 - \Phi_n + \tau_n)|,}{|(-1 - \theta_n + \sigma_n)(-1 - \Phi_n + \tau_n)|, \quad |(1 - \theta_n + \sigma_n)(-1 + \Phi_n + \tau_n)|} \right]$$

Then we define

(2a)
$$M(\lbrace a_n \rbrace, \lbrace \varepsilon_n \rbrace) = \inf_n \frac{1}{4} \pi_n.$$

It can be shown that $M(\{a_n\}, \{\epsilon_n\})$ has the same value if we transform the chains $\{a_{n+1}\}, \{\epsilon_n\}$ by reversing them both about the same point, or by changing the signs of all the a_n and of alternate ϵ_n , or by changing the signs of all the ϵ_n . We shall therefore regard two chain-pairs as completely distinct only when one cannot be obtained from the other by one of these transformations.

The following two theorems give the results needed for computing M(f).

THEOREM 1 ([2] and [7], § 2). We have

$$M(f) = \sup M(\{a_n\}, \{\varepsilon_n\}),$$

where the supremum is taken over all completely distinct chain-pairs $\{a_n\}$, $\{\varepsilon_n\}$ of f for which $\{\varepsilon_n\}$ satisfies certain conditions (E) (which are given in [7], § 2, and are not needed explicitly here).

THEOREM 2 ([7], \S 3). If g=(A,B,C) is a Gauss-reduced form equivalent to f, then all the completely distinct a-chains of f are included in the sets of a-chains from the triad of forms:

$$g = (A, B, C),$$

 $g_1 = (A, 2A + B, A + B + C),$
 $g_2 = (A - B + C, B - 2C, C).$

If, further, f has integral coefficients, then the chain of forms $\{f_n\}$ corresponding to any a-chain of f contains infinitely many "copies" of forms of the triad, and the distance between successive "copies" (which may or may not be of the same form) is bounded.

(The last statement is not given explicitly in [7], but it follows immediately from § 3 of [7] and the existence of the fundamental automorph of f and the fact that the number of I-reduced forms equivalent to f is finite.)

For any form f = (a, b, c) we define

(3)
$$\lambda(t) = \min |a \pm b + c|;$$

it can be shown that, if $\{f_n\}$ is the chain of forms corresponding to $\{a_n\}$, then, for each n,

$$\pi_n \leqslant \lambda(f_n),$$

and that the following theorem holds.



THEOREM 3 ([3] and [7], § 2). If the a_n are all even and not constantly equal to 2 or to -2, and if the ε_n are all zero, then the conditions (E) are satisfied, and, for each n.

$$\pi_n = \lambda(f_n)$$
.

As the even continued fraction expansion of the type (1) of any number with modulus greater than 1 is unique, it follows from Theorem 2 that there are at most three completely distinct even a-chains of the form f, given by the unique even a-chains from the triad g, g_1 , g_2 ; the minima obtained by taking zero ε -chains with these three even chains are, respectively,

$$M(g; \frac{1}{2}, \frac{1}{2}), M(g; 0, \frac{1}{2}), M(g; \frac{1}{2}, 0).$$

(We note this because the inhomogeneous minima of some of the forms which we shall consider in § 5 correspond to chain-pairs of this type.)

The following results ([3] and [7], § 2) are needed for the computation of σ_n and τ_n or of bounds for them, and hold for all n:

(5)
$$\theta_n = a_n - \frac{1}{\theta_{n-1}}, \quad \Phi_n = a_{n+1} - \frac{1}{\Phi_{n+1}};$$

(6)
$$\sigma_n = \varepsilon_{n-1} - \frac{\sigma_{n-1}}{\theta_{n-1}}, \quad \tau_n = \varepsilon_n - \frac{\tau_{n+1}}{\Phi_{n+1}};$$

(7)
$$|\sigma_n| \leqslant |\theta_n| - 1, \quad |\tau_n| \leqslant |\Phi_n| - 1.$$

Also, as Mr Swinnerton-Dyer pointed out, π_n is less than or equal to the smaller of

$$(8) \quad \frac{\varDelta}{(\theta_n \varPhi_n - 1)} \min \left[\frac{(\theta_n + 1) \left((\varPhi_n + 1)^2 - \tau_n^2 \right)}{\varPhi_n + 1}, \frac{(\theta_n - 1) \left((\varPhi_n - 1)^2 - \tau_n^2 \right)}{\varPhi_n - 1} \right],$$

$$(9) \quad \frac{\varDelta}{(\theta_n\varPhi_n-1)} \ \min\bigg[\frac{(\varPhi_n+1)\big((\theta_n+1)^2-\sigma_n^2\big)}{\theta_n+1}, \ \frac{(\varPhi_n-1)\big((\theta_n-1)^2-\sigma_n^2\big)}{\theta_n-1}\bigg].$$

This follows from (2) by using (7) and the fact that, if x, y, z, t are real and y, t have the same sign, then

$$\min[xy, zt] \leqslant \frac{(x+z)yt}{y+t}.$$

3. Further ideas underlying the method of computation. The aim was to use the divided cell method to obtain the inhomogeneous minimum of a form whose coefficients are in rationed ratios; therefore we assume in the remainder of this discussion that the coefficients of the given form f are integral. It is conjectured (see [1]) that in this case the inhomogeneous minimum M(f) is the value of M(f; P) corresponding

to a rational point P and so is the minimum corresponding to a periodic chain-pair $\{a_{n+1}\}, \{\epsilon_n\}$. By Theorem 2, all the periodic a-chains are included in the periodic a-chains from a triad g, g_1, g_2 , and it seems possible that the minimum M(f) corresponds to one of these periodic a-chains whose period is not too long. This idea underlies the whole approach to the problem. At least the largest value of $M(\{a_{n+1}\}, \{\epsilon_n\})$ corresponding to periodic chain-pairs of fairly short period from g, g_1, g_2 should give a good lower bound for M(f). (A program was written to obtain this lower bound but it turned out to be too slow because of the large number of ε -chains corresponding to a given a-chain.)

Consider a given "block" of length n forwards from the form f_0 (equivalent to f) to the form f_n :

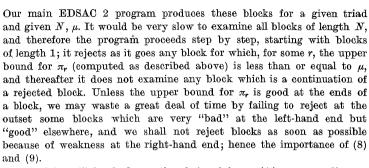
$$a_1, a_2, \ldots, a_n;$$

 $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{n-1}.$

We have θ_0 , Φ_0 , θ_n , Φ_n determined by f_0 , f_n , and we can compute θ_r , Φ_r for $r=1,\ldots,n-1$ by using (5). We can use (6) and (7) to obtain upper and lower bounds for σ_r , τ_r for $r=0,1,\ldots,n$ over the set of all chain-pairs of f which contain the given block; we can then use (2), (8), and (9) to obtain upper and lower bounds for π_r for $r=0,1,\ldots,n$ for all such chain-pairs. Clearly the bounds for σ_r will be crude at the left-hand end, while those for τ_r will be crude at the right-hand end; however, as (8) is independent of σ_n , and (9) of τ_n , we can get fairly good estimates for the upper bound of π_r , even at the ends. If we know that some chain-pair has minimum $\frac{1}{4}\mu$, and we can show that the upper bound for π_r is less than μ for some r in a given block, then we can reject all chain-pairs containing this block, because, for such chain-pairs, $M(\{a_n\}, \{e_n\})$ must be less than $\frac{1}{4}\mu$ (by (2a)), and so less than M(f).

Care is needed in the computations leading to the upper and lower bounds for π_r for a given block, because we are working with inequalities, so that we must carry upper and lower bounds for most quantities, and must, at each stage, choose the correct bound and round off multiplications and divisions in the correct direction, instead of rounding off in the ordinary way. (Because of this, computations on EDSAC 2 had to be carried out in the "fixed point" scale rather than in "floating point".)

4. The EDSAC 2 programs used for computing M(f). The basic idea in finding M(f) for given f by means of computation on EDSAC 2 is as follows. We guess a constant $\frac{1}{4}\mu$ slightly less than M(f) and pick a triad of forms given by Theorem 2. Then, for some given N, we find those blocks of length N forwards from forms of the triad for which the inequalities for π_r ($r=0,1,\ldots,N$) do not contradict $\pi_r > \mu$ (i. e. we reject those blocks for which the upper bound for π_r is less than or equal to μ for some r).



If N is sufficiently large, the chain of forms $\{f_n\}$ corresponding to a block of length N from a form of the triad must (by Theorem 2) contain another copy of a form of the triad. Thus, from the set of blocks leading forwards from forms of the triad, we get a set of (shorter) blocks leading forward from one form of this triad to itself or to another form of the triad, and a set of possible successors to each block. With luck, it is possible, by fitting these blocks together in every way which avoids rejected blocks, to obtain a finite set of periodic chain-pairs. By Theorem 2, any chain-pair for which $M(\{a_n\}, \{\varepsilon_n\}) > \frac{1}{4}\mu$ must be one of these, and so by examining $M(\{a_n\}, \{\varepsilon_n\})$ for these chain-pairs we should be able to find M(f).

In practice there are two ways in which things may go wrong. If we have picked μ too large, we may get no blocks at all, or all the chain-pairs built of the blocks produced may give $M(\{a_n\}, \{\epsilon_n\}) \leq \frac{1}{4}\mu$; in either case we at least deduce that $\frac{1}{4}\mu$ is an upper bound for M(f). If we have picked μ too small, we get too many blocks and an infinity of possible chains. However then we can often pick out a few plausible chain-pairs, some of which usually give values of $M(\{a_n\}, \{\epsilon_n\})$ which are greater than $\frac{1}{4}\mu$ and can be used as the new value for $\frac{1}{4}\mu$; if this does not happen, we must have chosen N too small, so that our results have been vitiated by uncertainty at the ends of the blocks. It is desirable to take N at least twice the longest distance between successive forms of the triad in chains from forms of the triad; by working out all the chains from forms of the triad, this distance can be computed, but in practice it was more efficient to guess N and increase it if necessary.

It follows from (4) that we cannot have $M(\{a_n\}, \{\varepsilon_n\}) > \frac{1}{4}\mu$ for any chain-pair from a form for which $\lambda \leqslant \mu$. Thus if we pick a triad of forms given by Theorem 2 for which one or more of the forms has $\lambda \leqslant \mu$, we can reject these forms immediately, and need only apply our main program to the remaining forms of the triad. In order to take advantage of this, and to gain some idea of the possible chains, we used:

45

(1) Program 1, which, for a given form f and given μ , examines the possible triads of Theorem 2, picks a triad in which as many forms as possible have $\lambda \leqslant \mu$, and gives either all possible a-chains from the forms of the triad which do not lead to forms with $\lambda \leqslant \mu$ or all possible even a-chains with this property. In fact the first alternative was very slow and was therefore not used, while the second was fast and usually gave sufficient indication of the type of chain to be expected.

As well as our main program and Program 1, it was convenient to have:

- (2) Program 2, which evaluates $M(\{a_n\}, \{\varepsilon_n\})$ precisely (recording the numbers as p/q) for a given periodic chain-pair. Both this and the computations of Program 1 could, of course, be done by hand.
- (3) Programs to find, for a given form f, chain-pairs with large values of $M(\{a_n\}, \{e_n\})$ which could be used as initial values of $\frac{1}{4}\mu$ in the main program. As it was so slow to consider all the periodic chain-pairs with fairly short periods, we had to use a program which considers only some of the possible e-chains and which sometimes gave a very poor "guess"; however there were many forms for which even this guess was so large that the forms could immediately be rejected as uninteresting. If the guess was very small we were often able to obtain a better one by using (1) and (2) to obtain the minima for the even a-chains with zero e-chains. Often it was best to work by repeating the main program with several different values of μ without first checking that they were valid guesses.

A proof of the value of M(f) (once known) takes 2-5 minutes of EDSAC 2 time, but of course finding M(f) may take much longer if one is unlucky. It would be possible to fit all the programs together into one program which would obtain M(f), but it would be hard to avoid an immense amount of unnecessary calculation in such a program, and so human intervention at some points is more efficient.

5. Forms with low minima. Let u_r , r = 0, 1, ... denote the Fibonacci numbers $(u_0 = 0, u_1 = 1, u_{r+1} = u_r + u_{r-1}, r \ge 1)$, and v_r , r = 0, 1, ... denote the Lucas numbers $(v_0 = 2, v_1 = 1, v_{r+1} = v_r + v_{r-1}, r \ge 1)$, and let

$$G_n(x,y) = u_n x^2 + v_n xy - u_n y^2;$$

it is easily shown that

(10)
$$\Delta(G_n) = \sqrt{(v_n^2 + 4u_n^2)} = \sqrt{(9u_n^2 + (-1)^n 4)}.$$

The forms G_n with n odd are (in a slightly different notation) the symmetric Markov forms which I investigated previously [7]. I showed that, for the

first few (odd) values of n and for all large (odd) n,

(11)
$$M(G_n) = \frac{1}{4}u_n$$
, if $n \not\equiv 0 \pmod{3}$.

It follows from (10) and (11) that, for odd n not divisible by 3, $M(G_n)/\Delta(G_n)$ tends down to 1/12 as n tends to infinity, and this implies that K is not greater than 1/12.

For even n, however, (10) implies that $\frac{1}{4}u_n/\Delta(G_n)$ tends up to 1/12 as n tends to infinity. Thus, if a result corresponding to (11) held for any even n, it would mean that K must be less than 1/12, and the smaller the value of n, the stronger the result would be. Therefore it seemed natural to investigate the forms G_n with n even.

In fact, I found that, for the form $G_{10} = (55, 123, -55)$,

$$M(G_{10}) = 55/4 = \frac{1}{4}u_{10}, \quad 4M(G_{10})/\Delta(G_{10}) = .333309...$$

By using the main program with $\mu=55$ and then with $\mu=54.9$, it was found that there are no chain-pairs for which $M(\{a_n\}, \{\varepsilon_n\})$ is greater than 55/4, and that the only chain-pair for which $M(\{a_n\}, \{\varepsilon_n\})$ might be greater than 54.9/4 consists of the even a-chain from G_{10} with the zero ε -chain; for this chain-pair, the minimum is 55/4 and thus we have the result.

The corresponding result holds for $G_{14} = (337, 843, -337)$, but this is, of course, weaker (as would be corresponding results for higher n), because $u_n/\Delta(G_n)$ increases for even n. Therefore, in order to obtain forms f for which $M(f)/\Delta(f)$ might be smaller, it seemed sensible to examine forms obtained by slightly varying the continued fraction chain of G_{10} . A complete period of the simplest a-chain of G_{10} is

$$(12) 3, 3, 3, 3, 3, -3, -3, -3, -3;$$

and a complete period of the (one and only) even a-chain of G_{10} , which gave the value of $M(G_{10})$, is:

$$2, -2, -2, 2, 2, -2, -2, -4, -2, 2, 2, -2, -4, -2, -2, 2, (13)$$

$$2, -2, -2, 2, -2, 2, 2, -2, -2, 2, 2, 4, 2, -2, -2, 2, 4, 2, (2, -2, -2, 2, 2, 2, 2, 2).$$

By slightly varying (12) and segments of (13), I found a number of forms with $M(f)/\Delta(f) < 1/12$. The best of these was

$$F = (547, 1097, -877),$$

for which

$$M(F) = 547/4$$

and

$$4M(F)/\Delta(F) = \cdot 309564...$$

Jane Pitman

46

In the same way as for G_{10} , it was found that the only chain-pair for which $M(\{a_n\}, \{\varepsilon_n\})$ might be greater than 546/4 consists of the even α -chain from F with the zero ε -chain, for which, in fact, $M(\{a_n\}, \{\varepsilon_n\})$ is 547/4. A complete period of the (sole) even α -chain of F is:

Another form with a very low minimum is

$$t = (151, 739, 193),$$

for which

$$.31701 < 4M(f)/\Delta(f) < .31775.$$

It should perhaps be noted that for this form, as well as for several others with slightly higher minima which were computed precisely before F was found, M(t) does not correspond to an even a-chain.

References

- [1] E. S. Barnes and H. P. F. Swinnerton-Dyer, The inhomogeneous minima of binary quadratic forms II, Acta Math. 88 (1952), p. 279-316.
- [2] The inhomogeneous minima of binary quadratic forms III, Acta Math. 92 (1954), p. 199-234.
- [3] E. S. Barnes, The inhomogeneous minima of binary quadratic forms IV, Acta Math. 92 (1954), p. 235-264.
- [4] H. Davenport, Indefinite binary quadratic forms, Quart. J. Math., Oxford Ser. (2) 1 (1950), p. 54-62.
- [5] Indefinite binary quadratic forms, and Euclid's algorithm in real quadratic fields. Proc. London Math. Soc. (2) 53 (1951), p. 65-82.
- [6] V. Ennola, On the first inhomogeneous minimum of indefinite binary quadratic forms and Euclid's algorithm in real quadratic fields, Ann. Univ. Turkuensis Ser. AI, 28 (1958), p. 9-58.
- [7] Jane Pitman, The inhomogeneous minima of a sequence of symmetric Markov forms, Acta Arithm. 5 (1958), p. 81-116.

GIRTON COLLEGE, CAMBRIDGE

Reçu par la Rédaction le 1. 10. 1959



ACTA ARITHMETICA VI (1960)

On a diophantine equation

bу

J. W. S. Cassels (Cambridge)

The following theorem answers a problem put to me orally by Professor Mordell. He tells me that he had known about the problem for some time and that it had several times been proposed to him (1).

THEOREM I. The system of equations

$$(1) r+s+t=rst=1$$

is insoluble in rationals r. s. t.

As Professor Mordell pointed out, this is equivalent to the following Theorem II. The only rational solutions of

$$(2) (r+s+t)^3 = rst$$

have

$$rst = 0$$
.

The equation of Theorem II represents a curve of genus 1 in homogeneous coordinates. It is, in fact, a particular case of an equation considered by Mordell [4]. He shows that it can be transformed into an apparently quite different shape. Since (2) is homogeneous, we may suppose without loss of generality that r, s, t are integers without common factor. It follows from (2) that r, s, t are coprime in pairs, and so, by (2) again,

$$r=\varrho^3, \quad s=\sigma^3, \quad t=\tau^3$$

where ρ , σ , τ are integers and

$$\rho^3 + \sigma^3 + \tau^3 = \rho \sigma \tau.$$

This is a special case, as Mordell remarks, of the equation

$$(3') a \rho^3 + b \sigma^3 + c \tau^3 + d \rho \sigma \tau = 0$$

considered by Sylvester [10] and Hurwitz [5] (cf. [9], p. 80-81).

⁽¹⁾ I am grateful to Professor Mordell for his comments on my MS.