ONE CUBIC DIOPHANTINE INEQUALITY

D. ERIC FREEMAN

1. *Introduction*

Suppose that $G(\mathbf{x})$ is a form, or homogeneous polynomial, of odd degree d in s variables, with real coefficients. Schmidt [**15**] has shown that there exists a positive integer $s_0(d)$, which depends only on the degree d, so that if $s \geqslant s_0(d)$, then there is an $\mathbf{x} \in \mathbb{Z}^s \setminus \{\mathbf{0}\}$ satisfying the inequality

$$|G(\mathbf{x})| < 1. \tag{1}$$

In other words, if there are enough variables, in terms of the degree only, then there is a nontrivial solution to (1). Let $s_0(d)$ be the minimum integer with the above property. In the course of proving this important result, Schmidt did not explicitly give upper bounds for $s_0(d)$. His methods do indicate how to do so, although not very efficiently. However, in fact much earlier, Pitman [**13**] provided explicit bounds in the case when G is a cubic. We consider a general cubic form $F(\mathbf{x})$ with real coefficients, in s variables, and look at the inequality

$$|F(\mathbf{x})| < 1. \tag{2}$$

Specifically, Pitman showed that if

$$s \geqslant (1314)^{256} - 1, \tag{3}$$

then inequality (2) is non-trivially soluble in integers. We present the following improvement of this bound.

THEOREM 1. *Suppose that s is an integer with*

$$s \geqslant 359\,551\,882, \tag{4}$$

and that $F(\mathbf{x})$ is a cubic form with real coefficients in s variables. Then there exists $\mathbf{x} \in \mathbb{Z}^s \setminus \{\mathbf{0}\}$ with

$$|F(\mathbf{x})| < 1.$$

We observe that for any $\epsilon > 0$, we can also solve the inequality

$$|F(\mathbf{x})| < \epsilon$$

non-trivially, by the standard technique of applying Theorem 1 to the form $\epsilon^{-1}F(\mathbf{x})$.

It is of interest to compare this theorem with related results. For example, much work has been done in the special case in which all the coefficients of F are integral. The inequality (2) then reduces to the equation

$$F(\mathbf{x}) = 0. \tag{5}$$

Received 15 May 1998.

2000 *Mathematics Subject Classification* 11D75 (primary), 11J25, 11D72, 11D25 (secondary).

The author was supported in part by NSF grant DMS-9622773.

In this case, Davenport [**6**] has shown that if $s \geqslant 16$, then there is a nontrivial integral solution $\mathbf{x}$ to equation (5). With the additional requirement that F is nonsingular, Heath-Brown [**7**] has shown that the condition $s \geqslant 10$ suffices to ensure nontrivial solubility of (5). Hooley [**8**] has improved this result, demonstrating that nine variables are sufficient, provided that F is nonsingular and has a nontrivial zero in every field $\mathbb{Q}_p$. Hooley [**9**, **10**] has further improved this result in additional work on cubic equations in nine variables.

On the other hand, one can consider the situation in which the coefficients of F are real, but not all in rational ratio. In the case of quadratic forms, we have the remarkable result of Margulis [**12**], which states that the values taken at integral points by any indefinite real quadratic form $Q(\mathbf{x})$ in at least three variables, and whose coefficients are not all in rational ratio, are dense on the real line. This of course implies that any such quadratic form takes arbitrarily small values at integer points, which is an analogue of inequality (2) for quadratic forms. One might expect a similar result if the coefficients of a cubic form F are not all in rational ratio. However, in fact, Wooley (in an unpublished result), by making use of properties of norm forms, has provided an example which shows that at least four variables are necessary to non-trivially solve inequality (2). Even in the special case where F is additive and the coefficients are not all in rational ratio, the best known result is the recent breakthrough by Baker, Brüdern and Wooley [**1**]. They have proved that seven variables are sufficient to guarantee a nontrivial solution to (2). It is difficult to say what one might expect to be the smallest number of variables required to non-trivially solve (2) when the coefficients are not all in rational ratio for a general homogeneous F, but it seems fairly likely that the bound (4) is quite far from the best possible.

Our methods are based on those of Pitman. The general idea, which can be traced to work of Brauer [**3**], is to find t linearly independent integral vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t \in \mathbb{Z}^s$ which make the form F 'almost diagonal'. By this we mean that for any choices of $u_1, \ldots, u_t \in \mathbb{R}$, we have

$$F(u_1\mathbf{x}_1 + \ldots + u_t\mathbf{x}_t) = F(\mathbf{x}_1)u_1^3 + \ldots + F(\mathbf{x}_t)u_t^3 + E, \tag{6}$$

where E is an error term bounded in terms of, among other quantities, the size of the variables u_i. We then apply a result due to Brüdern [**4**], which improves on a theorem of Pitman and Ridout [**14**, Theorem 2] and yields a small nontrivial integral solution of the diagonal form when $t = 9$. As this result allows us to choose the components of $(u_1, \ldots, u_9)$ to be small, we are able to make E small enough so that the right-hand side of (6) is small. Thus we have a nontrivial solution of the original inequality (2).

Brüdern's result provides some of our improvement over Pitman's bound (3), but much of the savings come from being able to find large sets of linearly independent integral points where certain linear forms are small. The kernel of this idea stems from Lewis and Schulze-Pillot [**11**]. We use the above sets to inductively find nine vectors which make F almost diagonal: at each step, we use the inductive hypothesis to choose t such vectors, and then find one additional vector. When undertaking an 'almost diagonalization' procedure, one considers the form $F(\mathbf{x}+\mathbf{y})$. Then one uses the binomial theorem to write $F(\mathbf{x}+\mathbf{y})$ as the sum of four forms F_0, F_1, F_2, and F_3, where F_i has degree i in $\mathbf{x}$ and $3-i$ in $\mathbf{y}$. We then have $F_0(\mathbf{x}, \mathbf{y}) = F(\mathbf{y})$ and $F_3(\mathbf{x}, \mathbf{y}) = F(\mathbf{x})$. We apply the inductive hypothesis to the form F_3 to obtain t vectors, and then choose $\mathbf{y}$ to be the extra desired vector. Here our method diverges from Pitman's. Pitman's technique involves choosing the vector $\mathbf{y}$ so that $F_2(\mathbf{z}, \mathbf{y})$ is small for suitably bounded $\mathbf{z}$ in a subspace of large dimension, say r. Then one must require that r be large enough

to find t suitable vectors within this subspace. This method causes rapid growth of the number of variables required because the form $F_2(\mathbf{z}, \mathbf{y})$ is a sum of $\binom{r+1}{2}$ linear forms in $\mathbf{y}$, and one chooses $\mathbf{y}$ by means of a lemma due to Birch and Davenport [**2**]. Roughly, this forces the number of variables needed to be squared in each step. Our technique, which is similar to a method of Wooley [**16**] stemming from work of Lewis and Schulze-Pillot [**11**], is to find a large set S of linearly independent integral points $\mathbf{v}_i$ which make the forms $F_1(\mathbf{v}_i, \mathbf{z})$ small for $\mathbf{z}$ of suitable size, with $\mathbf{z}$ in a subspace U. Within the lattice generated by the set S, we will use the inductive hypothesis to find t of our desired vectors. Call them $\mathbf{x}_1, \ldots, \mathbf{x}_t$, say. We proceed to choose $\mathbf{y}$ in the subspace U, suitably bounded, so that the form $F_2(v_1\mathbf{x}_1 + \ldots + v_t\mathbf{x}_t, \mathbf{y})$ is small for all suitably bounded choices of $v_1, \ldots, v_t$. This only yields $\binom{t+1}{2}$ forms, which is not very large in comparison with the number of variables needed to choose t 'diagonalizing' vectors, that is, the number of variables needed for the previous step. What makes this method more efficient is the fact that although the size of the set S is relatively large, the lemma which generalizes that of Birch and Davenport requires that we only add this number of variables in each step, that is, the cardinality of the set S, thus reducing the growth dramatically.

It should be noted that further small improvements are possible. For example, the choice of parameters at the end of the work could be made slightly more judiciously, but would seem to provide very little substantive improvement. A trick used by Pitman [**13**] would allow one to replace the right-hand side of (4) with 359 547 172, but at the price of a more complicated argument.

2. *Finding sets of linearly independent points*

As a preliminary to proving Theorem 1, we require a lemma. First, we must establish some notation.

Suppose that $\mathbf{x} \in \mathbb{R}^s$. Then we define

$$\|\mathbf{x}\| = \max_{j=1,\ldots,s} \{|x_j|\}.$$

Suppose that $F(\mathbf{x})$ is a form with real coefficients. Define $|F|$ to be the maximum of the absolute values of the coefficients of F.

We may now state a generalization of a well-known result originally proved by Birch and Davenport [**2**].

LEMMA 1. *Suppose that $L_1, \ldots, L_h$ are h real linear forms in s variables. Suppose also that N is a real number satisfying $N \geqslant C(h, s)$ for some constant $C(h, s)$, which depends only on h and s. Then, if $m \leqslant s$, there exist m linearly independent vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{Z}^s$ such that $\|\mathbf{x}_j\| \leqslant N$ for all j with $1 \leqslant j \leqslant m$, and*

$$|L_i(\mathbf{x}_j)| \leqslant N^{1-s/h+m/h}|L_i| \quad \text{for all } i \text{ with } 1 \leqslant i \leqslant h.$$

In the course of the proof, we will require an elementary lemma due to Davenport [**5**].

LEMMA 2. *Let A be any set of at least T distinct integer points $\mathbf{x}$ in the s-dimensional cube $\{\mathbf{x} \in \mathbb{R}^s : \|\mathbf{x}\| < N\}$, where $N > 1$. Suppose that there are at most r linearly independent points in the set A. Then*

$$T \ll N^r.$$

Here the constant in Vinogradov's notation depends only on s.

Now we may prove Lemma 1.

Proof of Lemma 1. Fix $N \geqslant C(h,s)$. First of all, we make some simplifying assumptions. We can assume that for all i with $1 \leqslant i \leqslant h$, the form L_i is not uniformly zero, for if some form L_i were uniformly zero, we could disregard it, and actually obtain a superior bound. Thus we may assume that $|L_i| \neq 0$ for all i. Consequently, by considering the forms $L_i/|L_i|$, we may assume without loss of generality that for all i with $1 \leqslant i \leqslant h$, the form L_i satisfies $|L_i| = 1$. Now consider, for all $\mathbf{x} \in \mathbb{Z}^s$ with $\|\mathbf{x}\| \leqslant N/2$, the vectors

$$\mathbf{L}(\mathbf{x}) = (L_1(\mathbf{x}), \ldots, L_h(\mathbf{x})). \tag{7}$$

If we insist that $C(h,s) \geqslant 2$, then there are $\geqslant \lfloor N \rfloor^s \geqslant N^s/2^s$ such vectors, where we count such vectors 'with multiplicity', that is to say, we are actually considering the cardinality of the corresponding set $\{\mathbf{x} \in \mathbb{Z}^s : \|\mathbf{x}\| \leqslant N/2\}$. As $|L_i| = 1$ for $1 \leqslant i \leqslant h$, and since $\|\mathbf{x}\| \leqslant N/2$, we know that $|L_i(\mathbf{x})| \leqslant s \cdot (N/2)$ for all i, so we know that all of these vectors are contained in the cube centered at the origin with side length sN. Split this cube into boxes, each having sides parallel to the coordinate axes, with side length at most $N^{(m+h-s)/h}$. We can divide each side of the cube into at most

$$\left\lceil \frac{sN}{N^{(m+h-s)/h}} \right\rceil \leqslant 1 + \frac{sN}{N^{(m+h-s)/h}} \leqslant \frac{2sN}{N^{(m+h-s)/h}}$$

intervals; here we have used the fact that $m+h-s \leqslant h$, whence

$$\frac{sN}{N^{(m+h-s)/h}} \geqslant 1.$$

Thus we may construct these boxes so that there are at most $(2sN)^h/N^{m+h-s}$ of them. By the Pigeonhole Principle, there exists one such box containing at least

$$v = \left\lceil \frac{N^s}{2^s} \cdot \frac{N^{m+h-s}}{(2sN)^h} \right\rceil = \left\lceil \frac{N^m}{2^{s+h} s^h} \right\rceil$$

of the vectors (7). Call these vectors $\mathbf{L}(\mathbf{y}_1), \ldots, \mathbf{L}(\mathbf{y}_v)$.

Now we fix l with $1 \leqslant l \leqslant m$. We note that we always have $v \geqslant 1$. In particular, there always exists a vector $\mathbf{y}_1$. Now, for $2 \leqslant i \leqslant v$, let $\mathbf{z}_i = \mathbf{y}_i - \mathbf{y}_1$. We use induction on l to show that there exist l linearly independent vectors $\mathbf{w}_1, \ldots, \mathbf{w}_l$ among the vectors $\mathbf{z}_2, \ldots, \mathbf{z}_v$.

We first demonstrate that the hypothesis holds in the case $l = 1$. By choosing $C(h,s)$ sufficiently large, and by noting that $m \geqslant 1$, we may assume that $v \geqslant 2$. As therefore $\mathbf{y}_2 \neq \mathbf{y}_1$, the vector $\mathbf{z}_2$ is nonzero. Let $\mathbf{w}_1 = \mathbf{z}_2$. This establishes the hypothesis in the case $l = 1$.

Now take $1 < l \leqslant m$, and assume that we have established the inductive hypothesis for the case $l-1$. Thus we must have at least $l-1$ linearly independent vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{l-1}$ among the vectors $\mathbf{z}_2, \ldots, \mathbf{z}_v$. Suppose that $\{\mathbf{w}_1, \ldots, \mathbf{w}_{l-1}\}$ is in fact a maximal linearly independent subset of $A = \{\mathbf{z}_2, \ldots, \mathbf{z}_v\}$. As we have required $C(h,s) > 1$, we may apply Lemma 2 to A with $r = l-1$. As $\|\mathbf{z}_i\| \leqslant 2N$ for $2 \leqslant i \leqslant v$, we conclude that

$$v - 1 = \operatorname{card}(A) \leqslant C_s (2N)^{l-1}$$

for some constant C_s. By choosing $C(h,s)$ sufficiently large, and noting that $m > l-1$,

we can ensure that $v-1$ is greater than $C_s(2N)^{l-1}$, which is a contradiction. Thus we can find some n, with $2 \leqslant n \leqslant v$, so that $\mathbf{z}_n$ is not in the $\mathbb{R}$-span of $\mathbf{w}_1, \ldots, \mathbf{w}_{l-1}$. Then, setting $\mathbf{w}_l = \mathbf{z}_n$, we see that the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_l$ are linearly independent, as desired, and therefore we have completed the inductive step.

We now use the inductive result in the case $l = m$. Then for each j with $1 \leqslant j \leqslant m$, we set $\mathbf{x}_j = \mathbf{w}_j$. Observe that because we have chosen $\mathbf{L}(\mathbf{y}_i)$ and $\mathbf{L}(\mathbf{y}_1)$ to be in the same box for $1 \leqslant i \leqslant v$, we must have

$$\|\mathbf{L}(\mathbf{z}_i)\| = \|\mathbf{L}(\mathbf{y}_i) - \mathbf{L}(\mathbf{y}_1)\| \leqslant N^{(m+h-s)/h} \quad \text{for all } i \text{ with } 1 \leqslant i \leqslant v.$$

As we have chosen $\mathbf{x}_1, \ldots, \mathbf{x}_m$ from among $\mathbf{z}_2, \ldots, \mathbf{z}_v$, this completes the proof of Lemma 1. □

3. *Reduction to an almost diagonal form*

We now give a definition in order to recast Lemma 1 in a more convenient form.

DEFINITION 1. Suppose that m is a nonnegative integer, h is a positive integer and E is a positive real number. Let $w_1^{(m)}(h, E)$ be the *smallest positive integer* t such that, given any integer $s \geqslant 1+t$, there exists a constant $C(h, s)$ so that, given any real linear forms $L_1, \ldots, L_h$ in s variables, and given any real number $N \geqslant C(h, s)$, there exist $m+1$ linearly independent integral vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{m+1}$ such that

$$\|\mathbf{x}_j\| \leqslant N$$

for all j with $1 \leqslant j \leqslant m+1$, and

$$|L_i(\mathbf{x}_j)| \leqslant N^{-E}|L_i|$$

for all i with $1 \leqslant i \leqslant h$ and for all j with $1 \leqslant j \leqslant m+1$.

It is not *a priori* clear that any such integer t exists. The proof of the following corollary of Lemma 1 shows that one does indeed exist, in the course of giving an upper bound for $w_1^{(m)}(h, E)$.

COROLLARY 1. *Suppose that m is a nonnegative integer, h is a positive integer, and E is a positive real number. Then*

$$w_1^{(m)}(h, E) \leqslant [h(E+1)+m].$$

Proof. By Lemma 1, it suffices to have $t = w_1^{(m)}(h, E)$ large enough so that

$$1 - \frac{1+t}{h} + \frac{m+1}{h} \leqslant -E,$$

which occurs if and only if

$$h - t + m \leqslant -hE.$$

However, this condition is equivalent to

$$t \geqslant m + h(E+1).$$

This yields our result. □

Now we give another definition.

DEFINITION 2. Suppose that n is a positive integer and E is a positive real number. Let $\tilde{w}_3^{(n)}(E)$ be the *smallest positive integer* t such that, given any integer $s \geqslant 1+t$, there exists a constant C_1, which may depend only on s and E, so that given any real cubic form F in s variables, and given any real number $N \geqslant C_1$, there exist n linearly independent vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{Z}^s$ with $\|\mathbf{x}_j\| \leqslant N$ for all j with $1 \leqslant j \leqslant n$, and so that for any $\mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n$,

$$F(u_1 \mathbf{x}_1 + \ldots + u_n \mathbf{x}_n) = F(\mathbf{x}_1)u_1^3 + \ldots + F(\mathbf{x}_n)u_n^3 + O(N^{-E}|F|\,\|\mathbf{u}\|^3).$$

Here the implicit constant in Landau's O-notation depends only on n and E.

Again, it is not *a priori* clear that any such integer t exists. However, the following lemma shows that one does exist, in the course of providing a method to give bounds for these quantities. It is a partial analogue for inequalities of a lemma given by Wooley. (See [**16**, Lemma 2.1].)

LEMMA 3. *Fix δ with $0 < \delta < 1$. Suppose as well that E_1, E_2 and E_3 are positive real numbers, and that n is a positive integer. Define*

$$E = \min\{E_1\,\delta + \delta - 3, E_2 - E_2\,\delta - 3\delta, E_3 - 2\},$$
$$M = \tilde{w}_3^{(n)}(E_2),$$
$$s = 1 + w_1^{(0)}\left(\binom{n+1}{2}, E_3\right).$$

Then if $E > 0$, one has

$$\tilde{w}_3^{(n+1)}(E) \leqslant s + w_1^{(M)}\left(\binom{s+1}{2}, E_1\right). \tag{8}$$

Proof. We note first of all that the proof is very similar to the proof of the aforementioned lemma given by Wooley. We also observe that it is possible to prove a closer analogue to this lemma, but we do not require the full generality such a lemma would provide.

For convenience, we first set $c_2 = 1/(M+1)$. Then we take any integer B with

$$B > s + w_1^{(M)}\left(\binom{s+1}{2}, E_1\right).$$

Then fix any N with N sufficiently large, where the meaning of this phrase will be indicated later. Consider any cubic form F with real coefficients in B variables. Now define the forms G_0, G_1, H_0, H_1 by

$$F(\mathbf{y} + t\mathbf{x}) = G_0(\mathbf{y}, \mathbf{x})t + G_1(\mathbf{y}, \mathbf{x})t^3 + H_0(\mathbf{y}, \mathbf{x})t^2 + H_1(\mathbf{y}, \mathbf{x}). \tag{9}$$

We observe that $G_1(\mathbf{y}, \mathbf{x}) = F(\mathbf{x})$ and $H_1(\mathbf{y}, \mathbf{x}) = F(\mathbf{y})$.

Now we let $\{\mathbf{e}_1, \ldots, \mathbf{e}_B\}$ be the standard unit basis for $\mathbb{R}^B$. Define the real subspaces $T = \langle \mathbf{e}_1, \ldots, \mathbf{e}_s \rangle$ and $U = \langle \mathbf{e}_{s+1}, \ldots, \mathbf{e}_B \rangle$. Consider any element $\mathbf{y} \in T \cap \mathbb{Z}^B$, say $\mathbf{y} = u_1 \mathbf{e}_1 + \ldots + u_s \mathbf{e}_s$, where $\mathbf{u} = (u_1, \ldots, u_s) \in \mathbb{Z}^s$. Upon substituting this $\mathbf{y}$, the polynomial $G_0(\mathbf{y}, \mathbf{x})$ becomes a form of total degree 2 in $\{u_1, \ldots, u_s\}$, and degree 1 in $\mathbf{x}$. Thinking of G_0 as a form in $\{u_1, \ldots, u_s\}$ of degree 2, we see that G_0 is a sum of at most $\binom{s+1}{2}$ monomials, whose coefficients are each forms of degree 1 in $\mathbf{x}$. Here we will apply the definition of $w_1^{(M)}(\binom{s+1}{2}, E_1)$ to the monomials of G_0, say $u_k u_l L_{kl}(\mathbf{x})$, which are linear in $\mathbf{x}$ for all k and l with $1 \leqslant k \leqslant l \leqslant s$. To do so, we restrict $\mathbf{x}$ to lie in U. However, U has affine dimension

$$B - s > w_1^{(M)}\left(\binom{s+1}{2}, E_1\right).$$

Thus, assuming that N is large enough in terms of M and δ so that $c_2 N^{\delta}$ is larger than the constant C implicit within the definition of $w_1^{(M)}(\binom{s+1}{2}, E_1)$, we may use the defining property of $w_1^{(M)}$. In this manner we find linearly independent integral vectors $\mathbf{d}_1, \ldots, \mathbf{d}_{M+1}$, with $\|\mathbf{d}_i\| \leqslant c_2 N^{\delta}$ and $\mathbf{d}_i \in U$ for all i with $1 \leqslant i \leqslant M+1$, so that each of the $\binom{s+1}{2}$ coefficients is small enough that we have

$$|G_0(\mathbf{y}, \mathbf{d}_i)| \ll \|\mathbf{u}\|^2 (c_2 N^{\delta})^{-E_1} |F| \qquad \text{for } 1 \leqslant i \leqslant M+1.$$

As G_0 is linear in the second argument, we see that for any $\mathbf{t} = (t_1, \ldots, t_{M+1}) \in \mathbb{R}^{M+1}$, we have

$$|G_0(\mathbf{y}, t_1 \mathbf{d}_1 + \ldots + t_{M+1} \mathbf{d}_{M+1})| \ll \|\mathbf{u}\|^2 \|\mathbf{t}\| (c_2 N^{\delta})^{-E_1} |F|. \tag{10}$$

Now we will concern ourselves with the form $G_1(\mathbf{y}, \mathbf{x})$, that is, $F(\mathbf{x})$. To that end, consider the form R defined by

$$R(t_1, \ldots, t_{M+1}) = F(t_1 \mathbf{d}_1 + \ldots + t_{M+1} \mathbf{d}_{M+1}). \tag{11}$$

As R is a cubic form in $M+1$ variables, we can find linearly independent $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{Z}^{M+1}$ with $\|\mathbf{a}_i\| \leqslant N^{1-\delta}$ for all i with $1 \leqslant i \leqslant n$, and such that for all $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$,

$$\begin{aligned} R(v_1 \mathbf{a}_1 + \ldots + v_n \mathbf{a}_n) = {} & v_1^3 R(\mathbf{a}_1) + \ldots + v_n^3 R(\mathbf{a}_n) \\ & + O((N^{1-\delta})^{-E_2} (|R|) (\|\mathbf{v}\|)^3). \end{aligned} \tag{12}$$

Now, for each i with $1 \leqslant i \leqslant n$, let $\mathbf{a}_i = (a_{i1}, \ldots, a_{i(M+1)})$. Then we have

$$\begin{aligned} R(v_1 \mathbf{a}_1 + \ldots + v_n \mathbf{a}_n) &= R\left(\sum_{i=1}^n v_i a_{i1}, \ldots, \sum_{i=1}^n v_i a_{i(M+1)}\right) \qquad (13) \\ &= F\left(\sum_{j=1}^{M+1} \left(\sum_{i=1}^n v_i a_{ij}\right) \mathbf{d}_j\right) \\ &= F\left(\sum_{i=1}^n v_i \left(\sum_{j=1}^{M+1} a_{ij} \mathbf{d}_j\right)\right) \\ &= F\left(\sum_{i=1}^n v_i \mathbf{b}_i\right), \qquad (14) \end{aligned}$$

where we define, for each i with $1 \leqslant i \leqslant n$,

$$\mathbf{b}_i = \sum_{j=1}^{M+1} a_{ij} \mathbf{d}_j.$$

Observe that, for each i and j,

$$R(\mathbf{a}_i) = F(a_{i1}\mathbf{d}_1 + \ldots + a_{i(M+1)}\mathbf{d}_{M+1}) = F(\mathbf{b}_i). \tag{15}$$

Upon inserting (14) and (15) into (12), we have, for all $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$,

$$F(v_1\mathbf{b}_1 + \ldots + v_n\mathbf{b}_n) = v_1^3 F(\mathbf{b}_1) + \ldots + v_n^3 F(\mathbf{b}_n) + O((N^{1-\delta})^{-E_2}(|R|)(\|\mathbf{v}\|)^3).$$

Now, recall that $\|\mathbf{d}_i\| \leqslant c_2 N^\delta$ for $1 \leqslant i \leqslant M+1$, whence we have $|R| \ll |F|(c_2 N^\delta)^3$. Therefore

$$F(v_1\mathbf{b}_1 + \ldots + v_n\mathbf{b}_n) = v_1^3 F(\mathbf{b}_1) + \ldots + v_n^3 F(\mathbf{b}_n) + O((N^{1-\delta})^{-E_2}(c_2^3 N^{3\delta})(|F|)(\|\mathbf{v}\|)^3). \tag{16}$$

We now handle the form H_0. Here we substitute $v_1\mathbf{b}_1 + \ldots + v_n\mathbf{b}_n$ for $\mathbf{x}$. Then H_0 becomes a form of degree 2 in $\{v_1, \ldots, v_n\}$, whose coefficients are forms of degree 1 in $\mathbf{y}$. There will be $\binom{n+1}{2}$ of these coefficients. We consider each monomial of the type $M_{ab}(\mathbf{y})v_a v_b$ with $1 \leqslant a \leqslant b \leqslant n$. We think of each such monomial as a form in $\mathbf{y}$ and presently bound the size of the coefficients. To do so, we first recall that each $\mathbf{b}_i = \sum_{j=1}^{M+1} a_{ij}\mathbf{d}_j$, so for $1 \leqslant i \leqslant n$,

$$\|\mathbf{b}_i\| \leqslant (M+1)(\|\mathbf{a}_i\|)(\max_{1 \leqslant j \leqslant M+1} \|\mathbf{d}_j\|) \leqslant (M+1)(N^{1-\delta})(c_2 N^\delta) = N, \tag{17}$$

as $c_2 = 1/(M+1)$. Thus our coefficients have size

$$\ll (\max_{1 \leqslant i \leqslant n} |v_i \mathbf{b}_i|)^2 |F| \ll N^2(\|\mathbf{v}\|)^2 |F|.$$

Now we restrict $\mathbf{y}$ to lie in T. Since T has affine dimension

$$s > w_1^{(0)}\left(\binom{n+1}{2}, E_3\right),$$

it follows for N sufficiently large that we can find a nonzero integral $\mathbf{f}$ with $\mathbf{f} \in T$, $\|\mathbf{f}\| \leqslant N$, and

$$|H_0(\mathbf{f}, v_1\mathbf{b}_1 + \ldots + v_n\mathbf{b}_n)| \ll N^{-E_3} N^2 (\|\mathbf{v}\|)^2 |F|.$$

Using the fact that H_0 is linear in the first entry, we have, for any $h \in \mathbb{R}$,

$$|H_0(h\mathbf{f}, v_1\mathbf{b}_1 + \ldots + v_n\mathbf{b}_n)| \ll N^{2-E_3}|h|(\|\mathbf{v}\|)^2 |F|. \tag{18}$$

We turn now to the form H_1. Recall that for any $\mathbf{y}$ and any $\mathbf{x}$, we have $H_1(\mathbf{y}, \mathbf{x}) = F(\mathbf{y})$. Thus, in particular, we have for any h, $v_1, \ldots, v_n \in \mathbb{R}$,

$$H_1(h\mathbf{f}, v_1\mathbf{b}_1 + \ldots + v_n\mathbf{b}_n) = h^3 F(\mathbf{f}). \tag{19}$$

Finally, we set

$$\mathbf{x} = v_1\mathbf{b}_1 + \ldots + v_n\mathbf{b}_n \quad \text{and} \quad \mathbf{y} = h\mathbf{f}.$$

Then, using (9) to combine (10), (16), (18) and (19), we see that, for any $\mathbf{v} \in \mathbb{R}^n$ and any $h \in \mathbb{R}$, one has

$$F(h\mathbf{f} + v_1\mathbf{b}_1 + \ldots + v_n\mathbf{b}_n) = v_1^3 F(\mathbf{b}_1) + \ldots + v_n^3 F(\mathbf{b}_n) + h^3 F(\mathbf{f}) + \Delta,$$

where

$$\Delta \ll \|\mathbf{u}\|^2 \|\mathbf{t}\| (c_2 N^\delta)^{-E_1} |F| + (N^{1-\delta})^{-E_2}(c_2^3 N^{3\delta})(|F|)(\|\mathbf{v}\|)^3 + N^{2-E_3}(|h|)(\|\mathbf{v}\|)^2 |F|. \tag{20}$$

Note that, by (11) and (13), we have chosen $t_j = \sum_{i=1}^{n} v_i a_{ij}$ for $1 \leqslant j \leqslant M+1$, whence we see that

$$\|\mathbf{t}\| \ll (\|\mathbf{v}\|)\,(\max_{1 \leqslant i \leqslant n} \|\mathbf{a}_i\|) \ll (\|\mathbf{v}\|)N^{1-\delta}.$$

Also, as $\mathbf{y} \in T$, one has $(u_1, \ldots, u_s, 0, \ldots, 0) = \mathbf{y} = h\mathbf{f}$, so that

$$\|\mathbf{u}\| \ll |h|(\|\mathbf{f}\|) \ll N|h|.$$

Upon rearranging and simplifying (20), we obtain, for any $\mathbf{v} \in \mathbb{R}^n$ and any $h \in \mathbb{R}$,

$$\Delta \ll \|\mathbf{v}\|(|h|)^2 N^{-E_1\delta-\delta+3}|F| + N^{-E_2+E_2\delta+3\delta}(|F|)\,(\|\mathbf{v}\|)^3 + N^{-E_3+2}|h|(\|\mathbf{v}\|)^2|F|.$$

It follows from (17) that $\|\mathbf{b}_i\| \leqslant N$ for $1 \leqslant i \leqslant n$. Also, we have chosen $\mathbf{f}$ so that $\|\mathbf{f}\| \leqslant N$. Moreover, by construction, $\mathbf{b}_i \in U$ for $1 \leqslant i \leqslant n$, and $\mathbf{f} \in T$, so the vectors $\mathbf{b}_1, \ldots, \mathbf{b}_n, \mathbf{f}$ are linearly independent. Recalling the definition of E, we see that we have found $n+1$ vectors of the desired type. □

4. *Completion of the proof of Theorem* 1

We simplify Lemma 3 in order to streamline our work when we 'almost diagonalize' our cubic form.

COROLLARY 2. *Fix δ with $0 < \delta < 1$. Fix a real number $E > 0$ and a positive integer n. Define*

$$s = 1 + \left\lceil \binom{n+1}{2}(E+3) \right\rceil.$$

Then

$$\tilde{w}_3^{(n+1)}(E) \leqslant s + \left\lceil \binom{s+1}{2}\left(\frac{E+3}{\delta}\right) \right\rceil + \tilde{w}_3^{(n)}\left(\frac{E+3\delta}{1-\delta}\right).$$

Proof. Define E_1, E_2, and E_3 so that

$$E_1 = \frac{E+3-\delta}{\delta}, \quad E_2 = \frac{E+3\delta}{1-\delta}, \quad \text{and} \quad E_3 = E+2.$$

Then note that we have

$$E = \min\{E_1\,\delta + \delta - 3, E_2 - E_2\,\delta - 3\delta, E_3 - 2\}.$$

Thus we may apply Lemma 3 with these choices for E_1, E_2, and E_3. By Corollary 1,

$$1 + w_1^{(0)}\left(\binom{n+1}{2}, E_3\right) \leqslant 1 + \left\lceil \binom{n+1}{2}(E_3+1) \right\rceil.$$

Therefore, we may take

$$s = \left\lceil \binom{n+1}{2}(E+3) \right\rceil + 1. \tag{21}$$

Now, to complete the proof, we merely note that Corollary 1 yields

$$w_1^{(M)}\left(\binom{s+1}{2}, E_1\right) \leqslant \left\lceil \binom{s+1}{2}\left(\frac{E+3}{\delta}\right) + M \right\rceil.$$

Thus, as $M = \tilde{w}_3^{(n)}((E+3\delta)/(1-\delta))$ is an integer, we obtain

$$w_1^{(M)}\left(\binom{s+1}{2}, E_1\right) \leqslant \left\lceil \binom{s+1}{2}\left(\frac{E+3}{\delta}\right)\right\rceil + \tilde{w}_3^{(n)}\left(\frac{E+3\delta}{1-\delta}\right).$$

Inserting this bound and equation (21) into equation (8) establishes Corollary 2. □

Now we state another theorem, which is basically a rewording of a result due to Brüdern [**4**, p. 2]. Brüdern's work is an improvement of a theorem of Pitman and Ridout [**14**, Theorem 2].

THEOREM 2 [**4**]. *Fix $\theta > 0$. Suppose that $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_9) \in \mathbb{R}^9$. Then the inequality*

$$|\lambda_1 x_1^3 + \ldots + \lambda_9 x_9^3| < 1 \tag{22}$$

has a nonzero solution $\mathbf{x} \in \mathbb{Z}^9$ with $\|\mathbf{x}\| \ll \max\{1, \|\boldsymbol{\lambda}\|^{8/3+\theta}\}$.

Proof. This is an immediate deduction from the statement of Brüdern at the end of [**4**, Section 1]. We note that if $|\lambda_i| < 1$ for any i with $1 \leqslant i \leqslant 9$, then we may easily find a nonzero solution satisfying the desired bound by setting $x_j = 1$ if $j = i$ and 0 otherwise, for $1 \leqslant j \leqslant 9$. Therefore, we may assume to the contrary. In this case, Brüdern's result shows that we may find a nontrivial solution to (22) with $\sum_{i=1}^{9} |\lambda_i x_i^3| \ll |\lambda_1 \ldots \lambda_9|^{1+\theta}$. Thus we see that for $1 \leqslant i \leqslant 9$, we have $|\lambda_i x_i^3| \ll |\lambda_1 \ldots \lambda_9|^{1+\theta}$. Therefore $|x_i^3| \ll \|\boldsymbol{\lambda}\|^{8+8\theta}$ for all i, that is, $\|\mathbf{x}\| \ll \|\boldsymbol{\lambda}\|^{8/3+8\theta/3}$. □

Proof of Theorem 1. Suppose that $F(\mathbf{x})$ is a real cubic form in s variables. Fix $\epsilon > 0$, and take any $s \geqslant 1 + \tilde{w}_3^{(9)}(24+\epsilon)$. Then for any sufficiently large N we may find, by definition of $\tilde{w}_3^{(9)}(24+\epsilon)$, linearly independent integral $\mathbf{x}_1, \ldots, \mathbf{x}_9$ with $\|\mathbf{x}_i\| \leqslant N$ for $1 \leqslant i \leqslant 9$, and

$$F(t_1 \mathbf{x}_1 + \ldots + t_9 \mathbf{x}_9) = t_1^3 F(\mathbf{x}_1) + \ldots + t_9^3 F(\mathbf{x}_9) + O(|F|(\|\mathbf{t}\|)^3 N^{-24-\epsilon}) \tag{23}$$

for any $\mathbf{t} = (t_1 \ldots, t_9) \in \mathbb{R}^9$. Now consider the inequality

$$|2(t_1^3 F(\mathbf{x}_1) + \ldots + t_9^3 F(\mathbf{x}_9))| < 1.$$

Choose $\theta > 0$ with $9\theta < \epsilon$. For $1 \leqslant i \leqslant 9$, one has $|F(\mathbf{x}_i)| \ll N^3|F|$. Thus, by Theorem 2, we can find an integral $\mathbf{t}$ with $0 < \|\mathbf{t}\| \ll (|F|N^3)^{8/3+\theta}$ solving the above inequality, that is, with

$$|(t_1^3 F(\mathbf{x}_1) + \ldots + t_9^3 F(\mathbf{x}_9))| < \tfrac{1}{2}.$$

Then the error term in (23) is

$$\ll |F|(|F|N^3)^{8+3\theta} N^{-24-\epsilon} \ll |F|^{9+3\theta} N^{24+9\theta-24-\epsilon}.$$

By choosing N sufficiently large in terms of $|F|$, we can force the magnitude of this error to be less than $\frac{1}{2}$, whence we see that $t_1 \mathbf{x}_1 + \ldots + t_9 \mathbf{x}_9$ is a solution to our original inequality. It is nontrivial because $\mathbf{x}_1, \ldots, \mathbf{x}_9$ are linearly independent and $(t_1, \ldots, t_9) \neq \mathbf{0}$. It now remains only to bound $\tilde{w}_3^{(9)}(24+\epsilon)$.

To do so, it is enough to use Corollary 2 eight times in succession, with well-chosen parameters δ. We note that $\tilde{w}_3^{(1)}(E) = 0$ for any positive real number E. Thus,

making use of Mathematica, starting with $n = 8$, and continuing to $n = 1$, the choices 0.127, 0.140 89, 0.166 66, 0.197 03, 0.238 66, 0.322 27, 0.430 6, and 0.999 99 for δ yield the bound (for a sufficiently small choice of ϵ)

$$\tilde{w}_3^{(9)}(24+\epsilon) \leqslant 359\,551\,881.$$

Thus one general homogeneous cubic diophantine inequality in s variables is non-trivially soluble, provided that $s \geqslant 359\,551\,882$, which is our desired result. □

Acknowledgements. I would like to thank Professor Wooley for his helpful suggestions. This work forms part of my PhD thesis for the University of Michigan.

References

1. R. C. Baker, J. Brüdern and T. D. Wooley, 'Cubic diophantine inequalities', *Mathematika* 42 (1995) 264–277.
2. B. J. Birch and H. Davenport, 'Indefinite quadratic forms in many variables', *Mathematika* 5 (1958) 8–12.
3. R. Brauer, 'A note on systems of homogeneous algebraic equations', *Bull. Amer. Math. Soc.* 51 (1945), 749–755.
4. J. Brüdern, 'Cubic diophantine inequalities II', *J. London Math. Soc.* (2) 53 (1996) 1–18.
5. H. Davenport, 'Cubic forms in thirty-two variables', *Philos. Trans. Roy. Soc. London Ser. A* 251 (1959) 193–232.
6. H. Davenport, 'Cubic forms in sixteen variables', *Proc. Roy. Soc. London Ser. A* 272 (1963) 285–303.
7. D. R. Heath-Brown, 'Cubic forms in ten variables', *Proc. London Math. Soc.* (3) 47 (1983) 225–257.
8. C. Hooley, 'On nonary cubic forms', *J. Reine Angew. Math.* 386 (1988) 32–98.
9. C. Hooley, 'On nonary cubic forms II', *J. Reine Angew. Math.* 415 (1991) 95–165.
10. C. Hooley, 'On nonary cubic forms III', *J. Reine Angew. Math.* 456 (1994) 53–63.
11. D. J. Lewis and R. Schulze-Pillot, 'Linear spaces on the intersection of cubic hypersurfaces', *Monatsh. Math.* 97 (1984) 277–285.
12. G. A. Margulis, 'Discrete subgroups and ergodic theory', *Number theory, trace formulas and discrete groups*, Oslo, 1987 (Academic Press, Boston, MA, 1989) 377–398.
13. J. Pitman, 'Cubic inequalities', *J. London Math. Soc.* 43 (1968) 119–126.
14. J. Pitman and D. Ridout, 'Diagonal cubic equations and inequalities', *Proc. Roy. Soc. London A* 297 (1967) 476–502.
15. W. M. Schmidt, 'Diophantine inequalities for forms of odd degree', *Adv. Math.* 38 (1980) 128–151.
16. T. Wooley, 'Forms in many variables, *Analytic number theory. Proceedings of the* 39*th Taniguchi International Symposium, Kyoto, May* 1996, London Mathematical Society Lecture Note Series 247 (ed. Y. Motohashi, Cambridge University Press, 1997) 361–376.

Department of Mathematics
University of Michigan
East Hall
525 East University Avenue
Ann Arbor
MI 48109-1109
USA