MATHEMATICS OF COMPUTATION
VOLUME 39, NUMBER 160
OCTOBER 1982, PAGES 709-723

On the Number of Markoff Numbers Below a Given Bound

By Don Zagier

Abstract. According to a famous theorem of Markoff, the indefinite quadratic forms with exceptionally large minima (greater than $\frac{1}{3}$ of the square root of the discriminant) are in 1 : 1 correspondence with the solutions of the Diophantine equation $p^2 + q^2 + r^2 = 3pqr$. By relating Markoff's algorithm for finding solutions of this equation to a problem of counting lattice points in triangles, it is shown that the number of solutions less than x equals $C \log^2 3x + O(\log x \ \log\log^2 x)$ with an explicitly computable constant $C = 0.18071704711507\ldots$. Numerical data up to 10^{1300} is presented which suggests that the true error term is considerably smaller.

1. By a *Markoff triple* we mean a solution (p, q, r) of the *Markoff equation*

$$(1) \qquad p^2 + q^2 + r^2 = 3pqr \qquad (p, q, r \in \mathbf{Z},\ 1 \leqslant p \leqslant q \leqslant r);$$

a *Markoff number* is a member of such a triple. These numbers, of which the first few are

$$1, 2, 5, 13, 29, 34, 89, 169, 194, 233, 433, 610, 985, \ldots,$$

play a role in a famous theorem of Markoff [10] (see also Frobenius [6], Cassels [2]): the $\mathrm{GL}_2(\mathbf{Z})$-equivalence classes of real indefinite binary quadratic forms Q of discriminant 1 for which the invariant

$$\mu(Q) = \min_{(x,y) \in \mathbf{Z}^2 - \{(0,0)\}} |Q(x, y)|$$

is greater than $\frac{1}{3}$ are in one-to-one correspondence with the Markoff triples, the invariant $\mu(Q)$ for the form corresponding to (p, q, r) being $(9 - 4r^{-2})^{-1/2}$. Thus the part of the *Markoff spectrum* (the set of all $\mu(Q)$) lying above $\frac{1}{3}$ is described exactly by the Markoff numbers. An equivalent theorem is that, under the action of $\mathrm{SL}_2(\mathbf{Z})$ on $\mathbf{R} \cup \{\infty\}$ given by $x \to (ax + b)/(cx + d)$, the $\mathrm{SL}_2(\mathbf{Z})$-equivalence classes of real numbers x for which the approximation measure

$$\mu(x) = \limsup_{q \to \infty} \Big(q \cdot \min_{p \in \mathbf{Z}} |qx - p| \Big)$$

is $> \frac{1}{3}$ are in 1 : 1 correspondence with the Markoff triples, the spectrum being the same as above (e.g. $\mu(x) = 5^{-1/2}$ for x equivalent to the golden ratio and $\mu(x) \leqslant 8^{-1/2}$ for all other x). Thus the Markoff numbers are important both in the theory of quadratic forms and in the theory of Diophantine approximation. They have also arisen in connection with problems in several other branches of mathematics, e.g. the

Received July 15, 1981; revised December 16, 1981.
1980 *Mathematics Subject Classification.* Primary 10B10; Secondary 10A21, 10C25, 10J25, 10F99.

word enumeration problem in the free semigroup on two generators [3] and the calculation of signature invariants of certain 4-dimensional manifolds [9].

Two obvious questions that one can ask about the Markoff spectrum are whether its terms all have multiplicity one (i.e. whether the equivalence class of Q or x with a given invariant $\mu > \frac{1}{3}$ is unique) and how rapidly the terms tend to the limiting value $\frac{1}{3}$, or equivalently (denoting by $\mathfrak{M}$ the set of Markoff triples)

1. Can there exist distinct triples (p, q, r), $(p', q', r') \in \mathfrak{M}$ with $r = r'$?
2. What is the asymptotic behavior of the function

$$\mathsf{M}(x) = \#\{(p, q, r) \in \mathfrak{M} \mid r \leq x\}$$

as $x \to \infty$?

The answer to 1. is almost certainly negative, but no correct proof of this conjecture has ever been advanced. As to the second—which, if the uniqueness conjecture is true, is equivalent to asking how many Markoff numbers lie below x—one is led easily by numerical data to the conjecture that $\mathsf{M}(x)$ grows like a constant times $\log^2 x$ (see Figure 1; the data on $\mathsf{M}(x)$ for $x \leq 10^{15}$ was obtained on a table computer in a few minutes using the tree algorithm described below). The purpose of this article is to show that this is the case. More precisely, we will prove the following result.

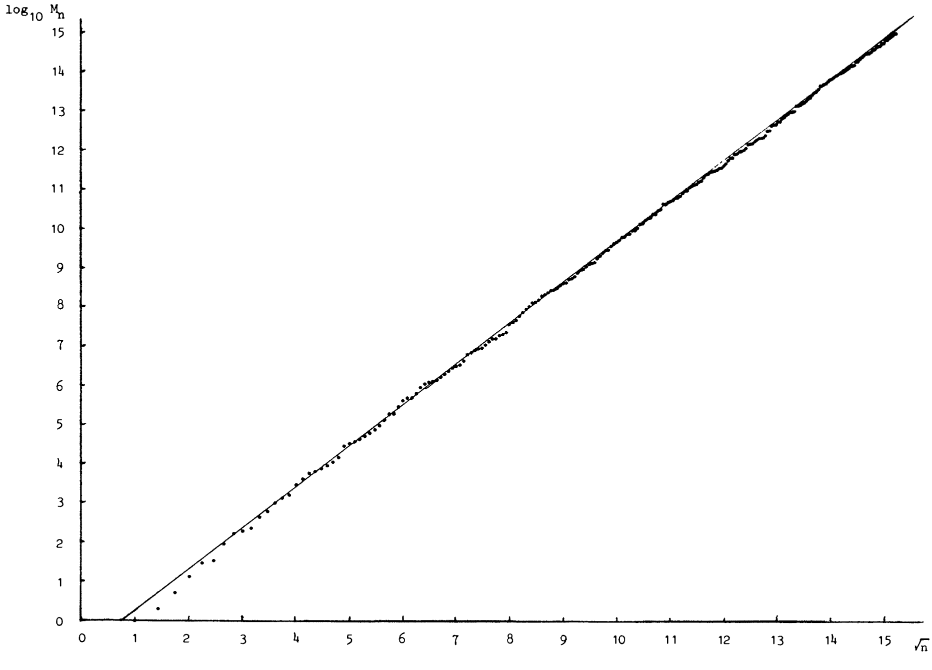


FIGURE 1

Graph of $\log M_n$ (M_n = *nth Markoff number*) *against* $\sqrt{n}$ *for* $M_n \leq 10^{15}$ *and least-squares linear fit*

THEOREM. *The number of Markoff triples* (p, q, r) *with* $p \leqslant q \leqslant r \leqslant x$ *is given by*

$$\text{(2)} \qquad \mathsf{M}(x) = C(\log x)^2 + O\big(\log x(\log\log x)^2\big) \qquad (x \to \infty),$$

where

$$\text{(3)} \qquad C = \frac{3}{\pi^2} \sum_{(p,q,r)\in\mathfrak{M}}^{*} \frac{f(p)+f(q)-f(r)}{f(p)f(q)f(r)} \approx 0.18071704711507;$$

here Σ^* *means that the two Markoff triples* $(1, 1, 1)$ *and* $(1, 1, 2)$ *with* $p = q$ *are to be counted with multiplicity* $\frac{1}{2}$, *and* $f(x)$ *is the function*

$$\text{(4)} \qquad f(x) = \log\frac{3x+\sqrt{9x^2-4}}{2} = \operatorname{arc\,cosh}\frac{3x}{2} \qquad \left(x \geqslant \frac{2}{3}\right).$$

The fact that $C_1 \log^2 x \leqslant \mathsf{M}(x) \leqslant C_2 \log^2 x$ for some positive constants C_1 and C_2 was proved by Harvey Cohn [4], who gave the lower bound

$$\liminf \frac{\mathsf{M}(x)}{\log^2 x} \geqslant \frac{3}{\pi^2}\frac{1}{f(1)f(2)} \approx 0.17917.$$

The existence of $\lim \mathsf{M}(x)/\log^2 x$ was proved by C. Gurwood in an as yet unpublished thesis [7]; however, his method of proof, which is quite different from ours, did not lead to a formula for the value of this limit and also gave a less precise error term than that in (2) (namely $O(\log^{13/7}x \log\log x)$).

We will give the proof of the theorem in the next section, while Section 3 describes numerical calculations of $\mathsf{M}(x)$ up to $x = 10^{1300}$. These suggest strongly that the asymptotic formula (2) can be replaced by

$$\text{(5)} \qquad \mathsf{M}(x) = C(\log 3x)^2 + o(\log x)$$

(equivalently, if M_n denotes the nth Markoff number, counting multiplicities if the unicity conjecture mentioned above is false, then $M_n \sim \frac{1}{3}A^{\sqrt{n}}$ where $A = e^{1/\sqrt{C}} \approx 10.5101504$), but do not seem to be conclusive enough to warrant a guess as to the order of magnitude of the true error term in (5).

The work described in this paper was done during a visit to the Istituto di Matematica in Pisa in 1979, with support from the Centro Nazionale di Ricerca and the Sonderforschungsbereich Theoretische Mathematik of the University of Bonn. I would like to thank the members of the Institute at Pisa, and in particular Professor Carlo Viola, for their hospitality during this visit, and Dr. F. Romani of the I.E.I. of the C.N.R., Pisa, for his help with the computer calculations. I would also like to thank the referee for several suggestions, in particular that of formulating the "Stokes' theorem" identity (24) to codify an argument used several times in the proof.

2. Given a Markoff triple $\mathfrak{m} = (p, q, r)$, three other triples can be found by fixing two elements of the triple and taking the other root of the quadratic equation satisfied by the third, i.e., by replacing (p, q, r) by $(p, q, 3pq - r)$, $(p, r, 3pr - q)$, or $(q, r, 3qr - p)$. Repeating the process, we obtain at each stage two new triples (two rather than three because one of three triples generated from $\mathfrak{m}$ is the one from which $\mathfrak{m}$ itself was obtained). In this way an infinite "tree" of triples is generated starting from the triple $(1, 1, 1)$ (see Figure 2); we will call this tree the *Markoff tree*.

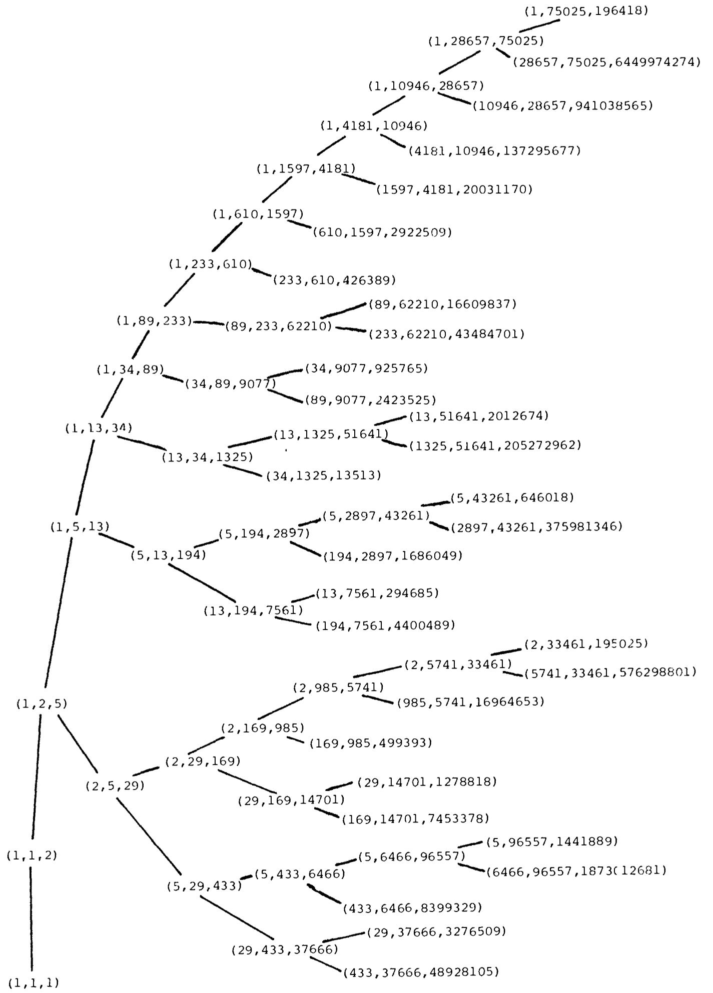


FIGURE 2

Markoff triples (p, q, r) *with* $\max(p, q) \leqslant 100000$

Conversely, given a Markoff triple (p, q, r) with $r > 1$, one checks easily that $3pq - r < r$; and from this it follows by induction that all Markoff triples occur, and occur only once, on this tree (for a fuller discussion of this and other properties of the Markoff tree, see [2]).

To prove the theorem we must analyze the asymptotic behavior of the Markoff tree. From the Markoff equation (1) we find that $3r^2 \geqslant 3pqr$ or $r \geqslant pq$; if p is large (which will happen for all but a small portion of the tree, contributing $O(\log x)$ to $\mathsf{M}(x)$), then this implies that r is much larger than q and hence (1) gives $r^2 < 3pqr < r^2 + o(r^2)$ or $r \sim 3pq$. Multiplying both sides of this equation by 3 and taking logarithms gives

$$\log(3p) + \log(3q) = \log(3r) + o(1) \qquad (p \text{ large}).$$

In other words, the map $x \to \log(3x)$ maps a Markoff triple onto an approximate solution (a, b, c) of the equation

$$a + b = c, \tag{6}$$

and to the same degree of approximation the branching process

$$\begin{tikzcd}[row sep=tiny] & (p,r,3pr-q) \approx (p,r,3pr) \\ (p,q,r) \arrow[ur] \arrow[dr] & \\ & (q,r,3qr-p) \approx (q,r,3qr) \end{tikzcd}$$

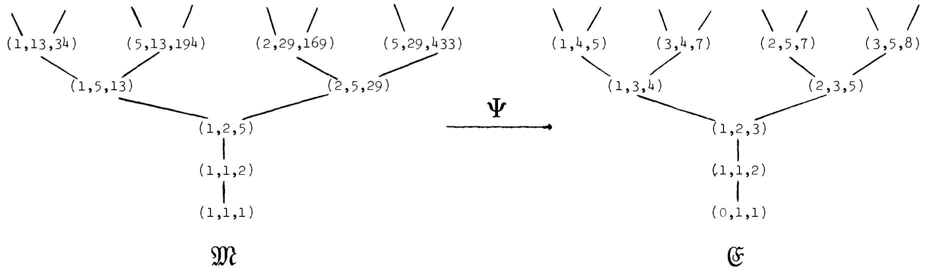
which defines the Markoff tree maps to the algorithm

$$\begin{tikzcd}[row sep=tiny] & (a,c,a+c) \\ (a,b,c) \arrow[ur] \arrow[dr] & \\ & (b,c,b+c) \end{tikzcd} \tag{7}$$

for passing from one solution of (6) to two larger ones. But this algorithm has a well-known effect: starting from the solution $(a, b, c) = (0, 1, 1)$ of (6) we generate by means of (7) all solutions of

$$a + b = c, \qquad 0 \leq a \leq b \leq c, \quad (a, b) = 1 \tag{8}$$

(Euclidean algorithm). Thus we obtain a "Euclid tree" $\mathfrak{E}$ parallel to the Markoff tree $\mathfrak{M}$ and a bijective correspondence Ψ: $\mathfrak{M} \to \mathfrak{E}$ sending the root $(1, 1, 1)$ of $\mathfrak{M}$ to the root $(0, 1, 1)$ of $\mathfrak{E}$ and then at each branching mapping the smaller branch of the Markoff tree to the smaller branch of the Euclid tree, as shown in Figure 3 (this pairing $\mathfrak{M} \leftrightarrow \mathfrak{E}$ was also observed by Harvey Cohn [5]). The proof of the theorem will be based on refinements of this correspondence. That this gives the right order of magnitude $C \log^2 x$ for $\mathsf{M}(x)$ can be seen from the fact that the counting function



p	1	2	5	13	29	34	89	169	194	233	433	610	985
$\alpha(p)$	1	2	3	4	5	5	6	7	7	7	8	8	9

FIGURE 3
The Markoff tree and the Euclid tree

$$\mathsf{E}(x) = \#\{(a, b, c) \in \mathfrak{E} \mid c \leq x\}$$

is given asymptotically by

$$\mathsf{E}(x) = 1 + \frac{1}{2} \sum_{c \leq x} \varphi(c) \sim \frac{3}{2\pi^2} x^2 \tag{9}$$

(φ = Euler function) and, by the remarks above, the correspondence Ψ is given roughly by $x \to \log(3x)$. A somewhat more careful argument leads to the following statement, which is weaker than the theorem and will be used in its proof.

LEMMA 1. *The ratio* $\mathsf{M}(x)/\log^2 x$ *is bounded and bounded away from zero. More precisely, we have*

$$\liminf \frac{\mathsf{M}(x)}{\log^2 x} \geqslant \frac{3}{2\pi^2 \log^2 3} \approx .126, \tag{10}$$

$$\limsup \frac{\mathsf{M}(x)}{\log^2 x} \leqslant \frac{6}{\pi^2 \log^2 5} \approx .235. \tag{11}$$

Proof. The correspondence Ψ clearly has the form

$$\Psi(p,q,r) = (\alpha(p), \alpha(q), \alpha(r)) \qquad (\text{if } p > 1),$$

where α is a map from the set of Markoff numbers onto the set of natural numbers, the first few values of which are given in the table in Figure 3. (Actually, since the conjecture on the unicity of Markoff numbers is not known, one should more properly say that α is a function of Markoff triples and write $\Psi(\mathfrak{m}) = (\alpha(\mathfrak{m}_p), \alpha(\mathfrak{m}_q), \alpha(\mathfrak{m}))$, where $\mathfrak{m}_p$ and $\mathfrak{m}_q$ denote the triples below $\mathfrak{m}$ with largest elements p and q, respectively; from now on this point will not be mentioned again and it will be understood that by "Markoff number" we mean a Markoff number marked by a given position of appearance as the largest element of a Markoff triple.) The map α is surjective but not injective: each integer $n > 2$ has $\frac{1}{2}\varphi(n)$ preimages. We claim that

$$\alpha(p) \geqslant \frac{\log 3p}{\log 3} \tag{12}$$

for all Markoff numbers p; indeed, this holds for $p = 1$ and $p = 2$ by inspection and then follows inductively because of the inequality $r < 3pq$ or $\log 3r < \log 3p + \log 3q$ for Markoff triples (p, q, r). Clearly (10) follows from (9) and (12). The reverse direction is similar: for a Markoff triple (p, q, r) other than $(1, 1, 1)$ or $(1, 1, 2)$ we have

$$(r - \tfrac{5}{2}pq)(r - \tfrac{1}{2}pq) = q^2(p^2 - 1) + \tfrac{1}{4}p^2(q^2 - 4) \geqslant 0,$$

and hence (since $r - \frac{1}{2}pq \geqslant \frac{1}{2}pq > 0$) $r \geqslant \frac{5}{2}pq$. From this we obtain by induction the estimate

$$\alpha(p) \leqslant \frac{\log 5p/2}{\frac{1}{2}\log 5},$$

which in conjunction with (9) implies Eq. (11).

To go further, we make two modifications:

A. Replace the function $x \to \log 3x$ by the function f defined in (4).

B. Break up the Markoff tree into a union of subtrees and apply the same comparison method to each one.

A. The function (4) arises for the following reason. Any Markoff number p occurs in an infinite chain (which Cassels [2], punning on a more famous discovery of the same mathematician, calls a Markoff chain) of triples (p, q_{i-1}, q_i) $(i \in \mathbf{Z})$ with

$$q_{i+1} = 3pq_i - q_{i-1}.$$

(This is not quite correct since in (1) we adopted the convention of always writing the elements of a Markoff triple in ascending order; thus if (q_0, q_1, p) is the first

Markoff triple containing p the chain actually has the form

$$\ldots,(p,q_{-1},q_{-2}),(q_0,p,q_{-1}),(q_0,q_1,p),(q_0,p,q_2),(p,q_2,q_3),\ldots$$

rather than $\{(p,q_{i-1},q_i)\}_{i\in\mathbf{Z}}$.) This is a Fibonacci-type recursion—indeed, if $p=1$, then the q_i *are* Fibonacci numbers—and by a standard argument the numbers q_i grow like the $|i|$th power of the larger eigenvalue of the matrix $\left(\begin{smallmatrix}3p & -1\\ 1 & 0\end{smallmatrix}\right)$, i.e.,

$$\lim_{|i|\to\infty}\frac{\log q_i}{|i|}=f(p).$$

As a result, the function f provides a much closer correspondence between the equations (1) and (6) than does the function $x\to\log 3x$: the latter function transforms (6) into the equation $r^2=3pqr$, which is a rather poor approximation to the Markoff equation, but the equation $f(p)+f(q)=f(r)$ obtained by applying the function f to (6) can be rewritten as

$$p^2+q^2+r^2=3pqr+\tfrac{4}{9}\tag{13}$$

(to see this, write $\varepsilon_p=(3p+\sqrt{9p^2-4})/2$ and similarly for q and r; then $3p=\varepsilon_p+\varepsilon_p^{-1}$, $3q=\varepsilon_q+\varepsilon_q^{-1}$, $3r=\varepsilon_p\varepsilon_q+\varepsilon_p^{-1}\varepsilon_q^{-1}$ and (13) follows), which is very close to (1). The error made in approximating (1) by (13) is bounded explicitly by the following lemma.

LEMMA 2. *Let (p,q,r) be a Markoff triple. Then*

$$f(r)<f(p)+f(q)<f(r)+\frac{c}{r^2},\tag{14}$$

where c is an absolute constant.

Proof. Let $r_1=f^{-1}(f(p)+f(q))$; this makes sense because f is monotone. Then (p,q,r_1) satisfies the equation (13), and, subtracting this from (1), we find

$$\tfrac{4}{9}=r_1^2-3pqr_1-r^2+3pqr=(r_1-r)(r_1+r-3pq)>\tfrac{4}{5}r(r_1-r),$$

where in the last step we have used $r_1>r$ and the inequality $r\geqslant\tfrac{5}{2}pq$ obtained earlier. Thus

$$r<r_1<r+\frac{5}{9r},$$

and applying f to these equations gives (14). We do not care much about the exact value of c, but remark that one can choose $c=.97$ or, if one excludes the two Markoff triples $(1,1,1)$ and $(1,1,2)$, $c=.54$, and this is more or less optimal since the inequality (14) is false with $c=.52$ for all Markoff triples with $p=1$.

B. Let $\mathfrak{M}'$ be a finite connected subset of $\mathfrak{M}$ which contains the "trunk" $\{(1,1,1),(1,1,2)\}$, and let $\mathfrak{S}$ be the set of triples $\mathfrak{m}\in\mathfrak{M}\setminus\mathfrak{M}'$ such that $\mathfrak{M}'\cup\{\mathfrak{m}\}$ is connected. It is clear by induction that $|\mathfrak{S}|=|\mathfrak{M}'|-1$ and that $\mathfrak{M}$ can be written as a disjoint union

$$\mathfrak{M}=\mathfrak{M}'\cup\bigcup_{\mathfrak{m}\in\mathfrak{S}}\mathfrak{M}_{\mathfrak{m}},\tag{15}$$

where $\mathfrak{M}_{\mathfrak{m}}$ is the infinite tree with root $\mathfrak{m}$, i.e. the set of all triples in $\mathfrak{M}$ lying above the triple $\mathfrak{m}$. Hence, for sufficiently large x (larger than r for all $(p,q,r)\in\mathfrak{M}'$), we have with the obvious notation

$$\mathsf{M}(x)=|\mathfrak{M}'|+\sum_{\mathfrak{m}\in\mathfrak{S}}\mathsf{M}_{\mathfrak{m}}(x)=1+\sum_{\mathfrak{m}\in\mathfrak{S}}(\mathsf{M}_{\mathfrak{m}}(x)+1).\tag{16}$$

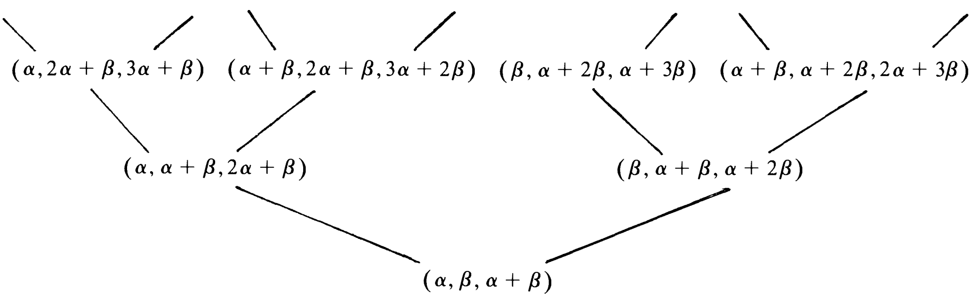
We will make the special choice

$$\mathfrak{M}' = \{(p,q,r) \in \mathfrak{M} \mid r \leqslant y\}, \qquad \mathfrak{S} = \{(p,q,r) \in \mathfrak{M} \mid q \leqslant y < r\},$$

where y is large but much smaller than x (Figure 2 illustrates $\mathfrak{M}' \cup \mathfrak{S}$ for this choice of $\mathfrak{M}'$ with $y = 10^5$); then (16) becomes

$$(17)\quad \mathsf{M}(x) = \mathsf{M}(y) + \sum_{\substack{(p,q,r)\in\mathfrak{M}\\ q\leqslant y<r}} \mathsf{M}_{(p,q,r)}(x) = 1 + \sum_{\substack{(p,q,r)\in\mathfrak{M}\\ q\leqslant y<r}} \big(\mathsf{M}_{(p,q,r)}(x) + 1\big).$$

For y large, the term $\mathsf{M}(y)$ in (17) can be estimated by Lemma 1. To study the terms in the sum, we use Lemma 2. For α, β positive real numbers we denote by $\mathfrak{E}_{\alpha,\beta}$ the tree


$(\alpha, 2\alpha+\beta, 3\alpha+\beta)$
$(\alpha+\beta, 2\alpha+\beta, 3\alpha+2\beta)$
$(\beta, \alpha+2\beta, \alpha+3\beta)$
$(\alpha+\beta, \alpha+2\beta, 2\alpha+3\beta)$
$(\alpha, \alpha+\beta, 2\alpha+\beta)$
$(\beta, \alpha+\beta, \alpha+2\beta)$
$(\alpha, \beta, \alpha+\beta)$

of solutions of (6) generated from $(\alpha, \beta, \alpha+\beta)$ using the algorithm (7) and by

$$\mathsf{E}_{\alpha,\beta}(x) = \#\{(a,b,c) \in \mathfrak{E}_{\alpha,\beta} \mid c \leqslant x\}$$

the counting function of $\mathfrak{E}_{\alpha,\beta}$. Then Lemma 2 implies that, for a Markoff triple (p,q,r) with r large, the function f maps the tree $\mathfrak{M}_{(p,q,r)}$ approximately onto the tree $\mathfrak{E}_{f(p),f(q)}$, so $\mathsf{M}_{(p,q,r)}(x) \approx \mathsf{E}_{f(p),f(q)}(f(x))$. More precisely, we have:

LEMMA 3. *Let* $\mathfrak{m} = (p,q,r)$ *be a Markoff triple, and set* $\alpha = f(p)$, $\beta = f(q)$, $\alpha' = f(p) - c/r^2$, $\beta' = f(q) - c/r^2$ *with c as in Lemma* 2. *Then for all x*

$$(18)\qquad \mathsf{E}_{\alpha,\beta}(f(x)) \leqslant \mathsf{M}_{\mathfrak{m}}(x) \leqslant \mathsf{E}_{\alpha',\beta'}(f(x)).$$

Proof. As in the comparison between $\mathfrak{M}$ and $\mathfrak{E}$, we see that the map $\Psi_{\mathfrak{m}}$: $\mathfrak{M}_{\mathfrak{m}} \to \mathfrak{E}_{\alpha,\beta}$, obtained by superposing one tree on the other in an order-preserving way, has the form $\Psi_{\mathfrak{m}}(p_1, q_1, r_1) = (\alpha_{\mathfrak{m}}(p_1), \alpha_{\mathfrak{m}}(q_1), \alpha_{\mathfrak{m}}(r_1))$, where $\alpha_{\mathfrak{m}}$ is the function from Markoff numbers occurring in $\mathfrak{M}_{\mathfrak{m}}$ to the set of positive integral linear combinations of α and β defined by $\alpha_{\mathfrak{m}}(p) = \alpha$, $\alpha_{\mathfrak{m}}(q) = \beta$ and $\alpha_{\mathfrak{m}}(r_1) = \alpha_{\mathfrak{m}}(p_1) + \alpha_{\mathfrak{m}}(q_1)$ for $(p_1, q_1, r_1) \in \mathfrak{M}_{\mathfrak{m}}$. Since $f(r_1) < f(p_1) + f(q_1)$ (Lemma 2), we see by induction that $\alpha_{\mathfrak{m}}(n) \geqslant f(n)$ for all n and hence that each triple $(a,b,c) \in \mathfrak{E}_{\alpha,\beta}$ with $c \leqslant f(x)$ corresponds under $\Psi_{\mathfrak{m}}$ to a triple $(p_1, q_1, r_1) \in \mathfrak{M}_{\mathfrak{m}}$ with $r_1 \leqslant x$. This gives the first inequality in (18). For the second, we define maps $\Psi'_{\mathfrak{m}}$ and $\alpha'_{\mathfrak{m}}$ in the same way as $\Psi_{\mathfrak{m}}$ and $\alpha_{\mathfrak{m}}$ but with α, β replaced by α', β'. Then we claim that $\alpha'_{\mathfrak{m}}(n) \leqslant f(n) - c/r^2$ for all n. Indeed, this is true (with equality) for $n = p$ and $n = q$, and if the assertion has been proved for the first two elements of a triple $(p_1, q_1, r_1) \in \mathfrak{M}_{\mathfrak{m}}$, then Lemma 2 implies

$$\begin{aligned}\alpha'_{\mathfrak{m}}(r_1) &= \alpha'_{\mathfrak{m}}(p_1) + \alpha'_{\mathfrak{m}}(q_1) \leqslant f(p_1) + f(q_1) - \frac{2c}{r^2}\\ &\leqslant f(r_1) + \frac{c}{r_1^2} - \frac{2c}{r^2} \leqslant f(r_1) - \frac{c}{r^2},\end{aligned}$$

whence the claim follows by induction. We deduce that

$$\mathsf{M}_{\mathrm{m}}(x) \leqslant \mathsf{E}_{\alpha',\beta'}\big(f(x) - c/r^2\big);$$

the inequality in the lemma is a slight weakening of this.

To use this result we need an asymptotic formula for the function $\mathsf{E}_{\alpha,\beta}(x)$. Since the numbers α, β will vary, it must be uniform in α, β and x. Such an estimate is provided by the following lemma.

LEMMA 4. *Let x, α, β be positive real numbers, $\beta \geqslant \alpha$. Then*

$$\mathsf{E}_{\alpha,\beta}(x) = \frac{3}{\pi^2}\frac{x^2}{\alpha\beta} + O\left(\frac{x}{\alpha}\right) + O\left(\frac{x}{\beta}\log\frac{x}{\beta}\right), \tag{19}$$

where the constants implied by the $O(\ )$ are absolute and effectively computable.

Proof. The triples occurring in the tree $\mathfrak{C}_{\alpha,\beta}$ are seen by induction to have the form $(m''\alpha + n''\beta,\ m'\alpha + n'\beta,\ m\alpha + n\beta)$, where m, n are coprime positive integers and m'', n'', m', n' are the integers determined uniquely from m and n by the conditions

$$0 \leqslant m'' < m', \quad 0 \leqslant n'' < n', \quad m'' + m' = m, \quad n'' + n' = n, \quad m'n - mn' = \pm 1.$$

Conversely, every pair of relatively prime positive integers m and n gives rise to an element of $\mathfrak{C}_{\alpha,\beta}$ in this way (if we map $\mathfrak{C}_{\alpha,\beta}$ to $\mathfrak{C}\setminus\{(0,1,1),(1,1,2)\} = \mathfrak{C}_{1,2}$ in the obvious way, then the element of $\mathfrak{C}_{\alpha,\beta}$ in question is sent to the triple $(m, m+n, 2m+n) \in \mathfrak{C}$). Hence

$$\mathsf{E}_{\alpha,\beta}(x) = \#\{(m,n) \in \mathbf{Z}^2 \mid m, n > 0, (m,n) = 1, m\alpha + n\beta \leqslant x\}. \tag{20}$$

Define

$$\mathsf{N}_{\alpha,\beta}(x) = \#\{(m,n) \in \mathbf{N}^2 \mid m\alpha + n\beta \leqslant x\}. \tag{21}$$

Then

$$\mathsf{N}_{\alpha,\beta}(x) = \sum_{d=1}^{\infty} \mathsf{E}_{\alpha,\beta}\left(\frac{x}{d}\right) = \sum_{d=1}^{[x/\beta]} \mathsf{E}_{\alpha,\beta}\left(\frac{x}{d}\right),$$

so by the Möbius inversion formula

$$\mathsf{E}_{\alpha,\beta}(x) = \sum_{d=1}^{\infty} \mu(d)\mathsf{N}_{\alpha,\beta}\left(\frac{x}{d}\right) = \sum_{d=1}^{[x/\beta]} \mu(d)\mathsf{N}_{\alpha,\beta}\left(\frac{x}{d}\right).$$

On the other hand,

$$\begin{aligned}\mathsf{N}_{\alpha,\beta}(x) &= \sum_{n=1}^{[x/\beta]} \left[\frac{x - n\beta}{\alpha}\right] = \sum_{n=1}^{[x/\beta]} \left(\frac{x - n\beta}{\alpha} + O(1)\right)\\ &= \frac{x}{\alpha}\left[\frac{x}{\beta}\right] - \frac{\beta}{2\alpha}\left(\left[\frac{x}{\beta}\right]^2 + \left[\frac{x}{\beta}\right]\right) + O\left(\frac{x}{\beta}\right)\\ &= \frac{x^2}{2\alpha\beta} - \frac{x}{2\alpha} + O\left(\frac{\beta}{\alpha}\right) + O\left(\frac{x}{\beta}\right),\end{aligned}$$

and hence

$$\begin{aligned}\mathsf{E}_{\alpha,\beta}(x) &= \sum_{d=1}^{[x/\beta]} \mu(d)\left(\frac{x^2}{2\alpha\beta d^2} - \frac{x}{2\alpha d} + O\left(\frac{\beta}{\alpha}\right) + O\left(\frac{x}{\beta d}\right)\right)\\ &= \frac{x^2}{2\alpha\beta}\sum_{d=1}^{[x/\beta]}\frac{\mu(d)}{d^2} - \frac{x}{2\alpha}\sum_{d=1}^{[x/\beta]}\frac{\mu(d)}{d} + O\left(\frac{x}{\alpha}\right) + O\left(\frac{x}{\beta}\sum_{d=1}^{[x/\beta]}\frac{1}{d}\right)\\ &= \frac{x^2}{2\alpha\beta}\left(\frac{6}{\pi^2} + O\left(\frac{\beta}{x}\right)\right) - \frac{x}{2\alpha}o(1) + O\left(\frac{x}{\alpha}\right) + O\left(\frac{x}{\beta}\log\frac{x}{\beta}\right).\end{aligned}$$

This proves the lemma.

We are now ready to estimate the various terms in (17). Lemmas 2, 3, and 4 give a uniform and effective estimate

$$\begin{aligned}\mathsf{M}_{(p,q,r)}(x) = \frac{3}{\pi^2}\,&\frac{f(x)^2}{\left(f(p)+O(1/r^2)\right)\left(f(q)+O(1/r^2)\right)}\\ &+O\left(\frac{f(x)}{f(p)}\right) + O\left(\frac{f(x)}{f(q)}\log\frac{f(x)}{f(q)}\right).\end{aligned}$$

We replace $\log(f(x)/f(q))$ in the last term by $\log f(x)$ since, with our final choice of y, it will be of this order of magnitude anyway. We can absorb the various " +1" in (17) into the error term $O(f(x)/f(p))$ since p will always be smaller than x. Equation (17) then gives

$$\mathsf{M}(x) = C_y f(x)^2 + O\left(D_y f(x)^2 + E_y f(x) + F_y f(x)\log f(x)\right),$$

where

$$C_y = \sum\frac{1}{f(p)f(q)},\quad D_y = \sum\frac{1}{r^2 f(p)^2 f(q)},\quad E_y = \sum\frac{1}{f(p)},\quad F_y = \sum\frac{1}{f(q)},$$

the sum in each case being over all Markoff triples (p, q, r) with $q \leqslant y < r$. We claim that

$$(22)\quad C_y = C + O\left(\frac{1}{y^2}\right),\quad D_y = O\left(\frac{1}{y^2}\right),\quad E_y = O(\log y),\quad F_y = O(\log y),$$

where C is the constant defined in (3). From this it will follow that

$$\mathsf{M}(x) = Cf(x)^2 + O\left(\frac{f(x)^2}{y^2} + f(x)\log y\ \log f(x)\right)$$

and hence with $y = (\log x)^{1/2}/\log\log x$ the assertion of the theorem.

We start with C_y. For $\mathfrak{M}'$, $\mathfrak{S}$ as in (15) we have the identity

$$(23)\qquad \sum_{(p,q,r)\in\mathfrak{S}}\frac{1}{f(p)f(q)} = \sum_{(p,q,r)\in\mathfrak{M}'}^{*}\frac{f(p)+f(q)-f(r)}{f(p)f(q)f(r)},$$

where Σ^* has the same meaning as in (3). This is a special case of the identity

$$(24)\qquad \sum_{(p,q,r)\in\mathfrak{S}} g(p,q) = \sum_{(p,q,r)\in\mathfrak{M}'}^{*}\left(g(p,r)+g(q,r)-g(p,q)\right)$$

(where g is an arbitrary function), which can be proved easily by induction on M', starting with $M' = \{(1,1,1), (1,1,2)\}$ and adding one vertex at a time. (The case $g = 1$ of (24) is the formula $|\mathfrak{S}| = |\mathfrak{M}'| - 1$ used earlier.) Equation (23) and Lemma 2 give

$$C_y = \sum_{\substack{(p,q,r)\in\mathfrak{M}\\ r\leqslant y}}^{*} \frac{f(p)+f(q)-f(r)}{f(p)f(q)f(r)} = C + \sum_{\substack{(p,q,r)\in\mathfrak{M}\\ r>y}} O\left(\frac{1}{r^2 f(p)f(q)f(r)}\right).$$

Since $f(p)$ is bounded from below and $f(q)$ and $f(r)$ are greater than a constant times $\log r$ (because $r < 3pq \leqslant 3q^2$), the last term is $O(\Sigma_r 1/r^2\log^2 r)$, where the sum extends over all Markoff numbers r (counting multiplicity if the uniqueness conjecture is false) greater than y. But this can be estimated using Lemma 1 and partial summation:

$$\begin{aligned}\sum_{M_n>y} \frac{1}{M_n^2\log^2 M_n} &= \sum_{r=y+1}^{\infty} \frac{\mathsf{M}(r)-\mathsf{M}(r-1)}{r^2\log^2 r}\\ &= \frac{-\mathsf{M}(y)}{(y+1)^2\log^2(y+1)} + \sum_{r>y} \mathsf{M}(r)\left(\frac{1}{r^2\log^2 r} - \frac{1}{(r+1)^2\log^2(r+1)}\right)\\ &= O\left(\frac{\mathsf{M}(y)}{y^2\log^2 y}\right) + \sum_{r>y} O\left(\frac{\mathsf{M}(r)}{r^3\log^2 r}\right) = O\left(\frac{1}{y^2}\right).\end{aligned}$$

This proves the first of the estimates (21). For the others, we observe that $r > y$ and hence $D_y < y^{-2}C_y = O(y^{-2})$, while $F_y \leqslant E_y$, so that we need only prove the estimate $E_y = O(\log y)$. For this we apply the identity (24) with $g(p,q) = 1/f(p)$ to obtain

$$E_y = \sum_{\substack{(p,q,r)\in\mathfrak{M}\\ r\leqslant y}}^{*} \frac{1}{f(q)}.$$

Since $r < 3q^2$ for a Markoff triple, we have $1/f(q) = O(1/\log r)$ and hence—again using Lemma 1 and Abel summation—

$$\begin{aligned}E_y &\ll \sum_{M_n\leqslant y} \frac{1}{\log M_n}\\ &= \sum_{p=1}^{y} \frac{\mathsf{M}(p)-\mathsf{M}(p-1)}{\log p} = \frac{\mathsf{M}(y)}{\log y} + \sum_{p=1}^{y-1} \mathsf{M}(p)\left(\frac{1}{\log p} - \frac{1}{\log(p+1)}\right)\\ &= O\left(\frac{\mathsf{M}(y)}{\log y}\right) + \sum_{p=1}^{y-1} O\left(\frac{\mathsf{M}(p)}{p\log^2 p}\right) = O(\log y) + \sum_{p=1}^{y-1} O\left(\frac{1}{p}\right) = O(\log y).\end{aligned}$$

This completes the proof of the theorem.

3. In this section we discuss the error term in the asymptotic formula (2) from both a theoretical and an experimental point of view.

It is clear from the proof of the theorem that the main source of this error term is the rather crude estimate of $\mathsf{E}_{\alpha,\beta}(x)$ given in Lemma 4. For instance, eliminating the $\log x/\beta$ in (19) would replace the error term in (2) by $O(\log x \log\log x)$, and replacing the error term in (19) by $(x/\alpha)^{1-\varepsilon}$ for some $\varepsilon > 0$ would permit one to reduce the error term in (2) to $O((\log x)^{1-\varepsilon}(\log\log x)^{1+\varepsilon})$. One can therefore ask whether a better result than that given in Lemma 4 can be obtained by using more sophisticated methods.

The function $\mathsf{E}_{\alpha,\beta}(x)$ is given by Eq. (20) as the number of visible lattice points in the triangle $m, n \leqslant 0$, $\alpha m + \beta n \leqslant x$, where by "visible" we mean a point which is the nearest point to the origin on a straight line. It is related to the total number $\mathsf{N}_{\alpha,\beta}(x)$ of lattice points in this triangle by a Möbius inversion formula. Now the trivial estimate $x^2/2\alpha\beta - x/2\alpha + O(x/\beta)$ which we gave for $\mathsf{N}_{\alpha,\beta}(x)$ can be improved considerably. The function $\mathsf{N}_{\alpha,\beta}$ was studied in two famous memoirs of Hardy and Littlewood [8], in which they prove in particular:

$$\frac{\alpha}{\beta}\text{ irrational} \Rightarrow \mathsf{N}_{\alpha,\beta}(x) = \frac{x^2}{2\alpha\beta} - \frac{x}{2\alpha} - \frac{x}{2\beta} + o(x),$$

$$t\left(\frac{\alpha}{\beta}\right) \geqslant h \Rightarrow \mathsf{N}_{\alpha,\beta}(x) = \frac{x^2}{2\alpha\beta} - \frac{x}{2\alpha} - \frac{x}{2\beta} + O(x^c) \qquad \left(\forall c > 1 - \frac{1}{h}\right);$$

here $t(\lambda)$, the approximation type of a real number λ, is defined by

$$t(\lambda) \geqslant h \quad \text{if} \left|\lambda - \frac{p}{q}\right| > \frac{\text{const}}{q^h} \text{ for all } p, q \in \mathbf{Z}, q > 0.$$

They show that these estimates are essentially best possible. Now in our application the number $\alpha/\beta = f(p)/f(q)$ is always irrational (since otherwise the numbers ε_p and ε_q defined after Eq. (13) would be units of norm 1 in the same real quadratic field, and hence the number $3r_1 = \varepsilon_p\varepsilon_q + \varepsilon_p^{-1}\varepsilon_q^{-1}$ would be an integer, contradicting the estimate $r < r_1 < r + 5/9r$ given in the proof of Lemma 2), and then the results of Baker [1, p. 22] show that it in fact has finite type. Thus one might hope that the error term in (19) could be improved considerably by combining the deep results of Hardy-Littlewood and of Baker. Unfortunately, this fails for several reasons:

(i) The results of Hardy and Littlewood are neither effective nor uniform with respect to α and β.

(ii) In passing from $\mathsf{N}_{\alpha,\beta}$ to $\mathsf{E}_{\alpha,\beta}$ by the Möbius formula, an error term $O(x^c)$ with $c < 1$ gives rise to an error

$$\sum_{d<x} \mu(d) O\left(\left(\frac{x}{d}\right)^c\right) = O\left(x^c \sum_{d<x} d^{-c}\right) = O\left(x^c \frac{x^{1-c}}{1-c}\right) = O(x)$$

essentially independent of c, so that the huge improvement of $O(x)$ to $O(x^c)$ for $\mathsf{N}_{\alpha,\beta}$ leads only to the modest improvement of $O(x \log x)$ to $O(x)$ for $\mathsf{E}_{\alpha,\beta}$.

(iii) Even if these problems could be resolved and we had an effective error estimate $O(x^{1-1/t(\alpha/\beta)})$ for $\mathsf{E}_{\alpha,\beta}(x)$, we could not improve our final result, because the value of $t(f(p)/f(q))$ following from the theorem of Baker quoted, namely $O(\log p \log q \log\log q)$ with an absolute and effective $O(\ )$-constant, is much too big for our purposes: in our application p and q run up to $y \approx \log x$, so Baker's result gives $t = O((\log\log x)^2 \log\log\log x)$, which is much larger than $\log f(x)$, so that $f(x)^{1-1/t}$ is of the same order of magnitude as $f(x)$.

Thus it does not seem possible to get a serious improvement of (19) in this way. However, a small improvement (say, getting rid of the logarithm) may well be possible; indeed, even in the case $\alpha = \beta$, which leads to the worst error term (namely $>$ const $\cdot\, x$) for the Hardy-Littlewood function $\mathsf{N}_{\alpha,\beta}(x)$, the error term in (19) can be improved at least from $O(x \log x)$ to $O(x \log^{2/3} x \log\log^{4/3} x)$, as was shown by Walfisz [11] (note that $\mathsf{E}_{1,1}(x)$ is essentially the summatory function of the Euler φ-function), and one can hope that for α/β irrational or of finite approximation type this can be improved still further. In any case, the problem of finding good estimates for $\mathsf{E}_{\alpha,\beta}(x)$ seems to be nontrivial and of some interest quite apart from the application to Markoff numbers. More generally, I would propose to analytic number theorists the problem of extending the classical results on lattice points in convex regions to the problem of counting visible lattice points in such regions.*

So much for the theoretical possibilities of finding the best error term in the asymptotic formula for $\mathsf{M}(x)$. From the practical point of view, one can simply try to calculate $\mathsf{M}(x)$ for some extremely large x and thus determine empirically whether the error term given in (2) is in the right ball park. The number $\mathsf{M}(x)$ can be calculated numerically by the same method as was used for the proof of the theorem, i.e., a combination of Eqs. (17) and (18) for a suitable choice of y. In the proof of the theorem we took a fairly small value of y (roughly $\sqrt{\log x}$) in order to balance the relatively large error in our formula for $\mathsf{E}_{\alpha,\beta}(f(x))$. In a numerical calculation we will compute $\mathsf{E}_{\alpha,\beta}(x)$ exactly and hence should choose a much larger y—ideally, so large that the estimate (18) gives an exact value for $\mathsf{M}_{\mathfrak{m}}(x)$. This is what was done in the actual computation: Eq. (17) was used with $y = 100000$ (i.e., with $\mathfrak{M}' \cup \mathfrak{S}$ as in Figure 2) and x ran up to 10^{1300}, still sufficiently small that the upper and lower bounds in (18) always agreed (not surprising since $f(10^{1300})$ is only about 3000, while α', β' and α, β differ by at most 10^{-10}). The computer calculated the difference

$$\varepsilon(x) = | \mathsf{M}(x) - C(\log 3x)^2 |$$

for each power of 10 up to 10^{1300} and printed this error every time it exceeded the largest previous error. The results are shown in Table 1.

TABLE 1

N	$\varepsilon(10^N)$	N	$\varepsilon(10^N)$	N	$\varepsilon(10^N)$	N	$\varepsilon(10^N)$	N	$\varepsilon(10^N)$
1	0.9	7	2.4	93	6.8	373	13.8	567	25.3
2	1.1	12	2.8	109	8.4	404	15.0	832	26.0
3	1.4	22	3.1	111	10.0	429	16.8	952	27.9
4	1.8	38	3.5	112	10.4	498	18.0	1015	37.8
5	2.3	44	6.6	255	13.3	534	23.1	1116	38.6

As can be seen, the agreement between $\mathsf{M}(x)$ and $C \log^2 3x$ is extremely good, the worst case found being

$$\mathsf{M}(10^{1116}) = 1,194,385, \qquad C\left(\log 3 \cdot 10^{1116}\right)^2 = 1,194,346.4$$

with an error of 38.6. This is so much smaller than the number $\log x (\log\log x)^2$ in our theorem, which for $x = 10^{1116}$ is about 1.2×10^6, that one's first thought is that the error term in (2) is of completely the wrong order of magnitude and that the correct exponent of $\log x$ must be less than 1, perhaps $\frac{1}{2}$ (since 38.6 is about the

*See note added in proof.

square-root of $\log(10^{1116})$). That this conclusion is not warranted by the numerical evidence is demonstrated in Figure 4, where the information of Table 1 is shown graphically in comparison with the three functions

$$(25) \qquad \frac{1}{10}(\log x)^{1/2}\log\log x, \quad \left(\frac{1}{10}\log x\right)^{2/3}, \quad \frac{\log x}{(\log\log x)^2}.$$

As can be seen, the growth of $\log x$ and $\log\log x$ is so slow that even for the huge values of x we are considering the three functions (25) have exactly the same order of magnitude (the function $\sqrt[6]{\log x}/\log\log x$ changes by less than 5% in the range $10^{30} \leqslant x \leqslant 10^{1300}$!). Thus the empirical evidence does not support the conclusion that the exponent of $\log x$ in (2) can be reduced, and after analyzing the errors in the proof of the theorem heuristically I would guess that the true error term in (2) probably is in fact $(\log x)^{1+o(1)}$, so that our result is in a crude sense best possible.

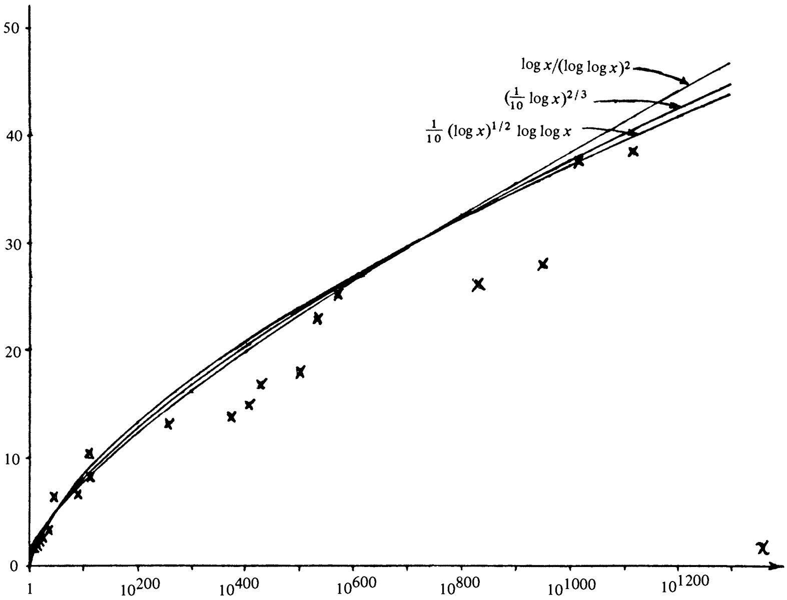


FIGURE 4

Comparison of $\varepsilon(x)$ with some functions of slow growth

Note Added in Proof. The problem of counting lattice points with coprime coordinates has been looked at, at my suggestion, by B. Z. Moroz (to appear). Assuming the Riemann hypothesis, he shows that an estimate of the form $N(x) = \lambda x^2 + \mu x + O(x^\gamma)$ with $\gamma < 1$ for the number of lattice points in xA (A a region in $\mathbf{R}^2$) implies an estimate of the form $E(x) = \zeta(2)^{-1}\lambda x^2 + O(x^{\gamma'})$ with $\gamma' < 1$ (in fact, with any $\gamma' > (4-\gamma)/(5-2\gamma)$) for the corresponding problem for visible points. Thus the difficulty mentioned in (ii) above can be circumvented if one assumes the Riemann hypothesis. However, this leads to no apparent improvement in the result on Markoff numbers, since the problems mentioned in (i) and (iii) remain.

Sonderforschungsbereich Theoretische Mathematik
Univeristät Bonn
Beringstrasse 4
D-5300 Bonn 1, West Germany

1. A. BAKER, *Transcendental Number Theory*, Cambridge Univ. Press, Cambridge, 1975.

2. J. W. S. CASSELS, *An Introduction to Diophantine Approximation*, Cambridge Univ. Press, Cambridge, 1957, Chapter II.

3. H. COHN, "Markoff numbers and primitive words," *Math. Ann.*, v. 196, 1972, pp. 8–22.

4. H. COHN, "Minimal geodesics on Fricke's torus-covering," in *Riemann Surfaces and Related Topics: Proceedings of the* 1978 *Stony Brook Conference*, Ann. of Math. Studies No. 97, Princeton Univ. Press, Princeton, N.J., 1981, pp. 73–86.

5. H. COHN, "Growth types of Fibonacci and Markoff," *Fibonacci Quart.* (2), v. 17, 1979, pp. 178–183.

6. G. FROBENIUS, *Über die Markoffschen Zahlen*, Preuss. Akad. Wiss. Sitzungsberichte, 1913, pp. 458–487.

7. C. GURWOOD, *Diophantine Approximation and the Markoff Chain*, Thesis, New York University, 1976, Section VI.

8. G. H. HARDY & J. E. LITTLEWOOD, "Some problems of Diophantine approximation: The lattice points of a right-angled triangle," (1st memoir), *Proc. London Math. Soc.* (2), v. 20, 1922, pp. 15–36, (2nd memoir), *Abh. Math. Sem. Univ. Hamburg*, v. 1, 1921, pp. 212–249. In *Collected Papers of G. H. Hardy*, Vol. I, pp. 136–158, 159–196, Clarendon Press, Oxford, 1966.

9. F. HIRZEBRUCH & D. ZAGIER, *The Atiyah-Singer Theorem and Elementary Number Theory*, Publish or Perish, Boston, Mass., 1974, §8.

10. A. A. MARKOFF, "Sur les formes binaires indéfinies," *Math. Ann.*, v. 17, 1880, pp. 379–399.

11. A. WALFISZ, *Weylsche Exponentialsummen in der neueren Zahlentheorie*, Math. Forschungsberichte XV, Deutscher Verlag d. Wiss., Berlin, 1963, Kap. IV.