

DETERMINING THE SMALL SOLUTIONS TO S -UNIT EQUATIONS

N. P. SMART

ABSTRACT. In this paper we generalize the method of Wildanger for finding small solutions to unit equations to the case of S -unit equations. The method uses a minor generalization of the LLL based techniques used to reduce the bounds derived from transcendence theory, followed by an enumeration strategy based on the Fincke-Pohst algorithm. The method used reduces the computing time needed from MIPS years down to minutes.

The main computational problem when solving a diophantine equation is usually the location of the “small” solutions. In this paper we assume we are given the generators of two finitely generated multiplicative subgroups of some number field K . In what follows we shall denote these subgroups by G_1 and G_2 . We also assume that we are given two fixed algebraic numbers $\alpha_1, \alpha_2 \in K^*$. In [4] the author gave a practical algorithmic solution to the determination of all the solutions to the equation

$$(1) \quad \alpha_1 \tau_1 + \alpha_2 \tau_2 + 1 = 0 \text{ with } (\tau_1, \tau_2) \in G_1 \times G_2.$$

That there are finitely many solutions to such an equation follows from work of Siegel. An effective proof of the finiteness of the number of solutions was first given by Györy, [3], using Baker’s method of linear forms in logarithms.

Using an adaption of Györy’s method combined with the reduction techniques of de Weger [9], one can reduce the solution of (1) to the determination of the “small” solutions. The technique used in [4] to determine such solutions was a sieving technique which lent itself to implementation on a parallel computer or a network of workstations. For further discussion of this sieving technique see [5].

Recently Wildanger [10] has given a much more efficient technique of determining the small solutions in the case where $G_1 = G_2 = \mathcal{O}_K^*$. In this paper we extend Wildanger’s method to the general case. The main problem that one encounters is the presence of finite places in the support of the two groups.

Wildanger makes use of the Fincke-Pohst algorithm [2]. We try to avoid the use of this algorithm for as long as possible. This is because we feel that applying Fincke-Pohst to lattices generated by real vectors with very large coefficients held to very high precision can lead to floating point errors. This is due to rounding errors in the algorithm for Cholesky decomposition and in the LLL algorithm itself. Indeed, rounding errors introduced in the floating point version of the LLL algorithm can lead to the production of a basis which is not even LLL reduced. Below we make use

Received by the editor December 1, 1997.

1991 *Mathematics Subject Classification*. Primary 11Y50, 11D61.

Key words and phrases. S -unit equations.

©1999 American Mathematical Society

of the LLL algorithm on lattices generated by vectors with integer entries. We can therefore make use of the integer version of the LLL algorithm due to de Weger, [8], which does not suffer from numerical instability. We only apply the Fincke-Pohst algorithm and the floating point version of LLL when we have considerably reduced the precision needed in the calculations.

1. NOTATION

We shall let S_1 and S_2 denote the set of primes (places), both finite and infinite, in the support of the groups G_1 and G_2 respectively. In other words

$$S_i = \{\mathfrak{p} \in M_K : |\alpha|_{\mathfrak{p}} \neq 1 \text{ for some } \alpha \in G_i\} = \text{Supp}(G_i).$$

We let t_i denote the rank of the group G_i . We suppose that G_i has independent generators of infinite order given by $\beta_{1,i}, \dots, \beta_{t_i,i}$. We can then write

$$\tau_i = \zeta_i \prod_{j=1}^{t_i} \beta_{j,i}^{a_{j,i}},$$

where $\zeta_i \in \text{Tors}(\mathcal{O}_K^*)$ and $a_{j,i} \in \mathbb{Z}$. We let $H = \max |a_{j,i}|$ and choose b and $\mathfrak{p}_g \in S_b$ such that

$$H = |a_{k,b}| \text{ for some } k \text{ and } |\log |\tau_b|_{\mathfrak{p}_g}| = \max_{\mathfrak{p}} |\log |\tau_b|_{\mathfrak{p}}|.$$

Now by Lemma 1 of [4] we can determine a constant c_1 such that

$$(2) \quad H \leq c_1 |\log |\tau_b|_{\mathfrak{p}_g}|.$$

Using the method of Baker and the computational reduction techniques of de Weger, see [4], we can find a constant, H_0 , of “reasonable” size such that $H \leq H_0$. By “reasonable” we mean “reasonable” when compared with the initial bound which can be derived from the application of Baker’s method alone. However the value of H_0 is usually still too large to allow direct enumeration of the solutions. It is common to refer to the solutions such that $H \leq H_0$ as the “small” solutions to the equation. This is because any “large” solutions are eliminated by Baker’s method and any “medium” sized solutions are eliminated by the application of the method of de Weger.

Let S denote a finite set of places of K , including all the infinite ones. We let S_f denote the subset of finite places of S . As the set of finite places of K and the set of prime ideals of \mathcal{O}_K are equivalent, we shall also refer to S_f as being a set of prime ideals. The order of $S_\infty = M_K^\infty$ is given by $r + 1$, where r is the rank of the group \mathcal{O}_K^* . We have $r + 1 = r_1 + r_2$, where r_1 is the number of real places and r_2 is the number of complex conjugate places. We place an order on S in the following way:

$$|\alpha|_{\mathfrak{p}_i} = \begin{cases} |\alpha^{(i)}|, & 1 \leq i \leq r_1, \\ |\alpha^{(i)}|^2, & r_1 + 1 \leq i \leq r_1 + r_2, \\ p_i^{-f_i \text{ord}_{\mathfrak{p}_i}(\alpha)}, & i > r_1 + r_2, \mathfrak{p}_i \in S_f, \end{cases}$$

where the conjugates $\alpha^{(i)}$ of K are ordered in the usual way, see [7, page 225], and p_i and f_i denote the rational prime lying below \mathfrak{p}_i and its residual degree respectively. The ramification index of \mathfrak{p}_i will be denoted by e_i . We of course assume some implicit fixed order for the places in $M_K \setminus M_K^\infty$, a fixed order for the real places and a fixed order for the complex places. This gives an order on any finite set S of places.

Let $R \in \mathbb{R}_{>1}$ and S a finite set of places of K . We define

$$\langle\langle R, S \rangle\rangle = \left\{ \alpha \in K : \frac{1}{R} \leq |\alpha|_{\mathfrak{p}} \leq R \text{ for all } \mathfrak{p} \in S \right\}.$$

As a last bit of notation we let \mathcal{L} denote the set of solutions that we wish to determine, i.e.

$$\mathcal{L} = \{(\tau_1, \tau_2) \in G_1 \times G_2 : \alpha_1\tau_1 + \alpha_2\tau_2 + 1 = 0\}.$$

We then let

$$\mathcal{L}_{H_i} = \{(\tau_1, \tau_2) \in \mathcal{L} : H \leq H_i\},$$

by which we mean those solutions whose maximum exponent is H_i ; hence $\mathcal{L} = \mathcal{L}_{H_0}$.

We also define

$$\mathcal{L}_{H_i}(R) = \{(\tau_1, \tau_2) \in \mathcal{L}_{H_i} : \tau_1 \in \langle\langle R, S_1 \rangle\rangle\}.$$

Now let

$$R_0 = \max_{\mathfrak{p} \in S_1} \exp \left(H_0 \sum_{j=1}^{t_1} |\log |\beta_{j,1}|_{\mathfrak{p}}| \right).$$

Lemma 1. $\mathcal{L} = \mathcal{L}_{H_0}(R_0)$.

Proof. We need to show that for all $\mathfrak{p} \in S_1$ we have

$$\frac{1}{R_0} \leq |\tau_1|_{\mathfrak{p}} \leq R_0.$$

Let $\mathfrak{p} \in S_1$; then we have, as $|a_{i,j}| \leq H_0$,

$$\begin{aligned} |\log |\tau_1|_{\mathfrak{p}}| &= \left| \sum_{j=1}^{t_1} a_{j,1} \log |\beta_{j,1}|_{\mathfrak{p}} \right| \leq \sum_{j=1}^{t_1} H |\log |\beta_{j,1}|_{\mathfrak{p}}| \\ &\leq \max_{\mathfrak{p} \in S_1} \left(H_0 \sum_{j=1}^{t_1} |\log |\beta_{j,1}|_{\mathfrak{p}}| \right) = \log R_0. \end{aligned}$$

Hence $-\log R_0 \leq \log |\tau_1|_{\mathfrak{p}} \leq \log R_0$, from which the result follows. □

2. DECOMPOSING THE SOLUTION SPACE

We set

$$s_i = \max_{\mathfrak{p} \in S_1 \cup S_2} \max (|\alpha_i|_{\mathfrak{p}}, |\alpha_i^{-1}|_{\mathfrak{p}}) \text{ for } i = 1 \text{ and } 2$$

and

$$s_3 = \max_{\mathfrak{p} \in S_1 \cup S_2} \min (|\alpha_2^{-1}|_{\mathfrak{p}}).$$

Now let $R_i, R_{i+1} \in \mathbb{R}_{>1}$ with $s_1, s_2, s_3 < R_{i+1} < R_i$, and let $H_i \in \mathbb{Z}$. We clearly have $R_{i+1} > 1$, as $s_1, s_2 > 1$. We shall also assume that $R_{i+1} > (s_3 - 1)/s_1$. The idea is then to find an integer $H_{i+1} < H_i$ and then decompose the space $\mathcal{L}_{H_i}(R_i)$ into the union of $\mathcal{L}_{H_{i+1}}(R_{i+1})$ and a union of sets which we can then show to have either no non-trivial elements or a few elements which can be easily determined.

If we can repeat this process, eventually we will be left with enumerating a set of the form $\mathcal{L}_{H_j}(R_j)$ for small values of H_j and R_j . In a later section we shall explain how the sets are shown to have either no non-trivial elements or a small set of easily determined elements. In this section we shall be content with showing how the solution space decomposes.

We define the following sets:

$$\begin{aligned} T_{1,\mathfrak{p}}(H_i, R_i, R_{i+1}) &= \left\{ (\tau_1, \tau_2) \in \mathcal{L}_{H_i}(R_i) : \left| -\alpha_1\tau_1 - 1 \right|_{\mathfrak{p}} < \frac{1}{1 + s_1 R_{i+1}} \right\}, \\ T_{2,\mathfrak{p}}(H_i, R_i, R_{i+1}) &= \left\{ (\tau_1, \tau_2) \in \mathcal{L}_{H_i}(R_i) : \left| -\frac{1}{\alpha_1\tau_1} - 1 \right|_{\mathfrak{p}} < \frac{1}{1 + s_1 R_{i+1}} \right\}, \\ T_{3,\mathfrak{p}}(H_i, R_i, R_{i+1}) &= \left\{ (\tau_1, \tau_2) \in \mathcal{L}_{H_i}(R_i) : \begin{array}{l} \left| -\alpha_2\tau_2 - 1 \right|_{\mathfrak{p}} < \frac{s_1}{R_{i+1}}, \\ \alpha_2\tau_2 \in \langle\langle 1 + s_1 R_i, S_2 \rangle\rangle \end{array} \right\}, \\ T_{4,\mathfrak{p}}(H_i, R_i, R_{i+1}) &= \left\{ (\tau_1, \tau_2) \in \mathcal{L}_{H_i}(R_i) : \begin{array}{l} \left| -\frac{\alpha_2\tau_2}{\alpha_1\tau_1} - 1 \right|_{\mathfrak{p}} < \frac{s_1}{R_{i+1}}, \\ \frac{\alpha_2\tau_2}{\alpha_1\tau_1} \in \langle\langle 1 + s_1 R_i, S_1 \cup S_2 \rangle\rangle \end{array} \right\}. \end{aligned}$$

We then define the sets

$$\begin{aligned} T_1(H_i, R_i, R_{i+1}) &= \bigcup_{\mathfrak{p} \in S_2} T_{1,\mathfrak{p}}(H_i, R_i, R_{i+1}), \\ T_2(H_i, R_i, R_{i+1}) &= \bigcup_{\mathfrak{p} \in S_1 \cup S_2} T_{2,\mathfrak{p}}(H_i, R_i, R_{i+1}), \\ T_3(H_i, R_i, R_{i+1}) &= \bigcup_{\mathfrak{p} \in S_1} T_{3,\mathfrak{p}}(H_i, R_i, R_{i+1}), \\ T_4(H_i, R_i, R_{i+1}) &= \bigcup_{\mathfrak{p} \in S_1} T_{4,\mathfrak{p}}(H_i, R_i, R_{i+1}). \end{aligned}$$

Lemma 2. *Let R_i, R_{i+1} and H_i be as above. We define*

$$c_2 = \max \left(\log \left(\frac{s_1 R_{i+1} + 1}{s_2} \right), \log \left(\frac{s_1 R_{i+1} + 1}{s_3} \right), \log(R_{i+1}) \right)$$

and set $H_{i+1} = c_1 c_2$. Then

$$\mathcal{L}_{H_i}(R_i) = \mathcal{L}_{H_{i+1}}(R_{i+1}) \cup \bigcup_{j=1}^4 T_j(H_i, R_i, R_{i+1}).$$

Proof. We assume that $(\tau_1, \tau_2) \in \mathcal{L}_{H_i}(R_i)$ and that $(\tau_1, \tau_2) \notin \mathcal{L}_{H_i}(R_{i+1})$. If this is the case, then there exists a $\mathfrak{q} \in S_1$ such that either $|\tau_1|_{\mathfrak{q}} < 1/R_{i+1}$ or $|\tau_1|_{\mathfrak{q}} > R_{i+1}$. In the first case we deduce that

$$\left| -\alpha_2\tau_2 - 1 \right|_{\mathfrak{q}} = |\alpha_1\tau_1|_{\mathfrak{q}} < \frac{s_1}{R_{i+1}}.$$

Now if $(\tau_1, \tau_2) \notin T_1(H_i, R_i, R_{i+1})$, then for all $\mathfrak{p} \in S_2$ we have

$$|\alpha_2\tau_2|_{\mathfrak{p}} = \left| -\alpha_1\tau_1 - 1 \right|_{\mathfrak{p}} \geq \frac{1}{1 + s_1 R_{i+1}}.$$

We also have that

$$|\alpha_2\tau_2|_{\mathfrak{p}} = \left| -\alpha_1\tau_1 - 1 \right|_{\mathfrak{p}} \leq 1 + |\alpha_1\tau_1|_{\mathfrak{p}} \leq \begin{cases} 1 + s_1 R_i, & \mathfrak{p} \in S_1, \\ 1 + s_1, & \mathfrak{p} \in S_2 \setminus S_1. \end{cases}$$

Hence, for all $\mathfrak{p} \in S_2$,

$$\begin{aligned} \left| \log |\alpha_2\tau_2|_{\mathfrak{p}} \right| &\leq \max\{\log(1 + s_1), \log(1 + s_1 R_{i+1}), \log(1 + s_1 R_i)\} \\ &= \log(1 + s_1 R_i). \end{aligned}$$

It therefore follows that if $|\tau_1|_{\mathfrak{q}} < 1/R_{i+1}$ and $(\tau_1, \tau_2) \notin T_1(H_i, R_i, R_{i+1})$, then $(\tau_1, \tau_2) \in T_3(H_i, R_i, R_{i+1})$.

We now consider the case that $|\tau_1|_q > R_{i+1}$. This means

$$\left| -\frac{\alpha_2\tau_2}{\alpha_1\tau_1} - 1 \right|_q = \left| \frac{1}{\alpha_1\tau_1} \right|_q < \frac{s_1}{R_{i+1}}.$$

If $(\tau_1, \tau_2) \notin T_2(H_i, R_i, R_{i+1})$, then for all $\mathfrak{p} \in S_1 \cup S_2$ we have

$$\left| -\frac{\alpha_2\tau_2}{\alpha_1\tau_1} \right|_{\mathfrak{p}} = \left| -\frac{1}{\alpha_1\tau_1} - 1 \right|_{\mathfrak{p}} \geq \frac{1}{1 + s_1R_{i+1}}.$$

In addition,

$$\left| -\frac{\alpha_2\tau_2}{\alpha_1\tau_1} \right|_{\mathfrak{p}} = \left| -\frac{1}{\alpha_1\tau_1} - 1 \right|_{\mathfrak{p}} \leq \left| \frac{1}{\alpha_1\tau_1} \right|_{\mathfrak{p}} + 1 \leq \begin{cases} 1 + s_1R_i, & \mathfrak{p} \in S_1, \\ 1 + s_1, & \mathfrak{p} \in S_2 \setminus S_1. \end{cases}$$

Hence $(\tau_1, \tau_2) \in T_4(H_i, R_i, R_{i+1})$.

So we have

$$\mathcal{L}_{H_i}(R_i) = \mathcal{L}_{H_i}(R_{i+1}) \cup \bigcup_{j=1}^4 T_j(H_i, R_i, R_{i+1}).$$

Now assume that $(\tau_1, \tau_2) \in \mathcal{L}_{H_i}(R_{i+1})$. Then for all $\mathfrak{p} \in S_1 \cup S_2$ we have $|\log |\tau_1|_{\mathfrak{p}}| \leq \log R_{i+1}$ and

$$|\tau_2|_{\mathfrak{p}} = \left| \frac{-\alpha_1\tau_1 - 1}{\alpha_2} \right|_{\mathfrak{p}} \leq \frac{|\alpha_1\tau_1|_{\mathfrak{p}} + 1}{|\alpha_2|_{\mathfrak{p}}} \leq \frac{s_1R_{i+1} + 1}{s_2}.$$

Now if $(\tau_1, \tau_2) \notin T_1(H_i, R_i, R_{i+1})$, then for all $\mathfrak{p} \in S_2$ we have

$$|\tau_2|_{\mathfrak{p}} = \left| \frac{-\alpha_1\tau_1 - 1}{\alpha_2} \right|_{\mathfrak{p}} \geq \frac{s_3}{s_1R_{i+1} + 1}.$$

This last inequality clearly also holds for $\mathfrak{p} \in S_1 \setminus S_2$, so we deduce for all $\mathfrak{p} \in S_1, S_2$ that if $(\tau_1, \tau_2) \in \mathcal{L}_{H_i}(R_{i+1}) \setminus T_1(H_i, R_i, R_{i+1})$, then

$$|\log |\tau_1|_{\mathfrak{p}}| \leq c_2, \quad |\log |\tau_2|_{\mathfrak{p}}| \leq c_2.$$

But from equation (2) we then deduce that we must have $H \leq H_{i+1}$. Hence the solution space decomposes as stated. \square

Clearly when applying this result we need to choose a value of R_{i+1} such that the methods of the following sections allow us to deduce that the sets $T_{j,\mathfrak{p}}(H_i, R_i, R_{i+1})$ are either trivial or easy to determine. Wildanger gives a heuristic method to determine the best values for R_{i+1} in the case where $G_1 = G_2 = \mathcal{O}_K^*$. The analysis in the general case appears similar, and Wildanger’s choices of R_{i+1} seem to suffice.

3. SHOWING THAT $T_{j,\mathfrak{p}}(H_i, R_i, R_{i+1})$ IS TRIVIAL

In all four cases our problem can be phrased as trying to show that there are no non-trivial solutions to the following problem. Let $\rho \in K^*$, and let $\epsilon_1, \dots, \epsilon_t$ denote multiplicatively independent elements of K^* . Set

$$\gamma = \zeta^{b_0} \prod_{i=1}^t \epsilon_i^{b_i}$$

with $\langle \zeta \rangle = \text{Tors}(\mathcal{O}_K^*)$. Let $\delta \in (0, 1)$. Then we wish to show that there are no solutions to the inequality

$$|\rho\gamma - 1|_{\mathfrak{p}} < \delta,$$

where \mathfrak{p} is some place of K and $|b_i| \leq B$ some given positive constant. There are two methods which we employ, depending on whether \mathfrak{p} is an infinite or finite place.

3.1. \mathfrak{p} infinite. Note that for any $z \in \mathbb{C}$, $|z - 1| < \delta$ implies that $|\log |z|| < \log(1/(1 - \delta))$. Hence we can immediately deduce that

$$|\log |\rho\gamma|_{\mathfrak{p}}| \leq \delta' = \begin{cases} \log \frac{1}{1-\delta}, & \mathfrak{p} \text{ real,} \\ \frac{1}{2} \log \frac{1}{1-\sqrt{\delta}}, & \mathfrak{p} \text{ complex.} \end{cases}$$

Notice that if δ is very small, then δ' will also be very small. We choose an integer constant C of the size of 10^t and then look at the lattice Λ generated by the columns of the matrix

$$A = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ 0 & & 1 & \\ [C \log |\epsilon_1|_{\mathfrak{p}}] & \dots & [C \log |\epsilon_{t-1}|_{\mathfrak{p}}] & [C \log |\epsilon_t|_{\mathfrak{p}}] \end{pmatrix} \in \mathbb{Z}^{t \times t},$$

where $[\cdot]$ denotes the nearest integer function, with some fixed convention for numbers of the form $(2m+1)/2$. We also define the vector $\vec{y} = (0, \dots, 0, -[C \log |\rho|_{\mathfrak{p}}])^t \in \mathbb{Z}^t$. Using the integer version of the LLL-algorithm and [9, Lemmata 3.4 and 3.5] we can compute a lower bound c_4 on

$$\ell(\Lambda, \vec{y}) = \begin{cases} \min_{\vec{x} \in \Lambda} \|\vec{x} - \vec{y}\|, & \vec{y} \notin \Lambda, \\ \min_{\vec{x} \in \Lambda} \|\vec{x}\|, & \vec{y} \in \Lambda. \end{cases}$$

We can then hopefully eliminate the set under consideration using the following lemma. If the following lemma does not work then we need to either increase C , or increase R_{i+1} , or use the technique of the next section.

Lemma 3. *Let*

$$Q = (t - 1)B^2 + \left(\frac{tB + 1}{2} + C\delta' \right)^2.$$

Then, if $c_4^2 > Q$, we can conclude that there are no non-trivial elements γ such that $|\rho\gamma - 1|_{\mathfrak{p}} < \delta$.

Proof. Put

$$\Phi = [C \log |\rho|_{\mathfrak{p}}] + \sum_{i=1}^t b_i [C \log |\epsilon_i|_{\mathfrak{p}}].$$

Then

$$\left| \Phi - C \left(\log |\rho|_{\mathfrak{p}} + \sum_{i=1}^t b_i \log |\epsilon_i|_{\mathfrak{p}} \right) \right| \leq \frac{tB + 1}{2}.$$

Hence

$$\begin{aligned} |\Phi| &\leq |\Phi - C \log |\rho\gamma|_{\mathfrak{p}}| + |C \log |\rho\gamma|_{\mathfrak{p}}| \\ &\leq \frac{tB + 1}{2} + C\delta'. \end{aligned}$$

Now consider the lattice point $\vec{x} = A\vec{z}$, where $\vec{z} = (b_1, \dots, b_t)^t$. For it we have

$$\vec{x} - \vec{y} = (b_1, \dots, b_{t-1}, \Phi)^t.$$

So, if $\vec{y} \neq \vec{x}$, we must have

$$c_4^2 \leq \ell(\Lambda, \vec{y})^2 \leq (t-1)B^2 + \Phi^2 \leq Q.$$

But as $c_4^2 > Q$ we see that $\vec{y} = \vec{x}$ and $b_1 = \dots = b_{t-1} = 0$ and $[C \log |\rho|_{\mathfrak{p}}] + b_t[C \log |\epsilon_t|_{\mathfrak{p}}] = 0$. If such a solution exists it is easy to determine and will therefore be called trivial. \square

3.2. \mathfrak{p} finite. Let p denote the rational prime lying below \mathfrak{p} and let e and f denote the ramification index and residue degree of \mathfrak{p} . We shall assume that

$$\delta' = -\frac{\log \delta}{ef \log p} \geq 1,$$

which is not a large restriction as we are assuming that δ is very small.

Our method proceeds using the p -adic analogue to the previous method for infinite places. If $|\rho\gamma - 1|_{\mathfrak{p}} < \delta < 1$, then we must have $\text{ord}_{\mathfrak{p}}(\rho\gamma) = 0$. Using the method explained in [4, Lemma 3], we can replace $\rho, \epsilon_1, \dots, \epsilon_t$ with a finite set of possibilities for $\mu_0, \mu_1, \dots, \mu_s \in K^*$ such that

$$\rho\gamma = \rho\zeta \prod_{i=1}^t \epsilon_i^{b_i} = \mu_0 \prod_{i=1}^s \mu_i^{m_i}$$

with $\text{ord}_{\mathfrak{p}}(\mu_i) = 0$ for $i = 0, \dots, s$ and $\max |m_i| \leq \max |b_i| \leq B$. So we are left with trying to determine if there are any solutions m_i to the inequality

$$|\mu_0 \prod_{i=1}^s \mu_i^{m_i} - 1|_{\mathfrak{p}} < \delta$$

with $|m_i| \leq B$. For $\alpha \in K^*$ we let $\log_p(\alpha)$ denote the p -adic logarithm of α when we consider α as an element of $K_{\mathfrak{p}}$. If we set

$$\Delta = \log_p \mu_0 + \sum_{i=1}^s m_i \log_p \mu_i \in K_{\mathfrak{p}},$$

then, as $\delta' \geq 1$,

$$\text{ord}_p(\Delta) = \text{ord}_p(\log_p(\rho\gamma)) = \text{ord}_p(\rho\gamma - 1) \geq \frac{-\log \delta}{ef \log p} = \delta'.$$

Let $n = [K_{\mathfrak{p}} : \mathbb{Q}_p]$ and $K_{\mathfrak{p}} = \mathbb{Q}_p(\phi)$; then we can write

$$\Delta = \sum_{i=0}^{n-1} \Delta_i \phi^i,$$

where

$$\Delta_i = \nu_{0,i} + \sum_{j=1}^s m_j \nu_{j,i}, \quad \text{with } \nu_{j,i} \in \mathbb{Q}_p \text{ and } 0 \leq i \leq n-1.$$

It then follows, see [7, Page 257], that, for all i ,

$$(3) \quad \text{ord}_p(\Delta_i) \geq \delta' - \frac{1}{2}D_p(\phi) = c_5,$$

where $D_p(\phi) = \text{ord}_p(\text{Disc}_{K_{\mathfrak{p}}/\mathbb{Q}_p}(\phi))$. We then choose $\lambda \in \mathbb{Q}_p$ such that

$$\text{ord}_p(\lambda) = \min_{0 \leq i \leq s} \left(\min_{0 \leq j \leq n-1} (\text{ord}_p(\nu_{i,j})) \right) = c_6.$$

We set

$$\Delta_i/\lambda = \kappa_{0,i} + \sum_{j=1}^s m_j \kappa_{j,i} \text{ for } i = 0, \dots, n-1,$$

and so $\kappa_{i,j} \in \mathbb{Z}_p$. We then have that

$$\text{ord}_p(\Delta_i/\lambda) \geq c_5 - c_6 = c_7.$$

We let $u \in \mathbb{N}$ be such that p^u is roughly the size of $B^{1+s/n}$ and such that $u \leq c_7$. The constant u plays a similar role to the constant C in the method for infinite places. For $\alpha \in \mathbb{Z}_p$ we let $\alpha^{(u)}$ denote the unique rational integer such that $\alpha \equiv \alpha^{(u)} \pmod{p^u}$ and $0 \leq \alpha^{(u)} < p^u$. We then let Λ denote the lattice generated by the columns of the matrix

$$A = \begin{pmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 0 & & 1 & \\ & & \kappa_{1,0}^{(u)} & \dots & \kappa_{s,0}^{(u)} & p^u & 0 \\ & & \vdots & & \vdots & & \ddots \\ \kappa_{1,n-1}^{(u)} & \dots & \kappa_{s,n-1}^{(u)} & & & & p^u \end{pmatrix} \in \mathbb{Z}^{(n+s) \times (n+s)}.$$

We also define the vector $\vec{y} = (0, \dots, 0, -\kappa_{0,0}^{(u)}, \dots, -\kappa_{0,n-1}^{(u)})^t \in \mathbb{Z}^{n+s}$. Using the integer version of the LLL-algorithm, we can compute a lower bound c_8 on $\ell(\Lambda, \vec{y})$. If the following lemma does not work, then we need to either increase u , or increase R_{i+1} , or use the method of the next section, just as we did when considering infinite places. However we must remember that we must always satisfy $u \leq c_7$; this is a severe restriction of the method in practice, especially when the ideal \mathfrak{p} is ramified. If \mathfrak{p} is ramified then c_7 can become very small, due to inequality (3).

Lemma 4. *If $c_8^2 > sB^2$, then there are no non-trivial solutions to the inequality*

$$|\rho\gamma - 1|_{\mathfrak{p}} < \delta.$$

Proof. As $\text{ord}_p(\Delta_i/\lambda) \geq c_7 \geq u$ for $i = 0, \dots, n-1$, we have, for all i ,

$$z_i = \frac{\kappa_{0,i}^{(u)} + \sum_{j=1}^s m_j \kappa_{j,i}^{(u)}}{p^u} \in \mathbb{Z}.$$

Therefore, we can consider the lattice point $\vec{x} = A\vec{z}$, where

$$\vec{z} = (m_1, \dots, m_s, -z_0, \dots, -z_{n-1})^t \in \mathbb{Z}^{n+s}.$$

Hence

$$\vec{x} - \vec{y} = (m_1, \dots, m_s, 0, \dots, 0)^t.$$

So either $c_8^2 \leq \ell(\Lambda, \vec{y})^2 \leq sB^2$ or $\vec{x} = \vec{y}$. The first possibility is ruled out by assumption, which leaves us to deduce that $m_1 = \dots = m_s = 0$. \square

4. ENUMERATING $T_{j,\mathfrak{p}}(H_i, R_i, R_{i+1})$

After application of the above techniques we will reach a space $\mathcal{L}_{H_i}(R_i)$ which we cannot decompose any further, as Lemma 2 gives rise to sets $T_{j,\mathfrak{p}}(H_i, R_i, R_{i+1})$ which we cannot show have only trivial elements.

We need to enumerate all the possible elements in $T_{j,\mathfrak{p}}(H_i, R_i, R_{i+1})$. It is at this stage that we make use of the Fincke-Pohst algorithm. However we hope that,

as we at least have a reduced value of R_i compared to our initial value R_0 , we can handle any numerical instability which occurs.

As before, we write

$$\gamma = \zeta^{b_0} \prod_{i=1}^t \epsilon_i^{b_i},$$

where $b_i \in \mathbb{Z}$. We can assume that $c_0 \in \{0, \dots, w - 1\}$, where w denotes the number of elements of finite order in \mathcal{O}_K^* . We have, for some $R \in \mathbb{R}_{>1}$, $\rho \in K^*$ and $\delta \in (0, 1)$,

$$\frac{1}{R} \leq |\rho\gamma|_{\mathfrak{q}} \leq R$$

for all $\mathfrak{q} \in S = \text{Supp}(\langle \epsilon_1, \dots, \epsilon_t \rangle)$ and

$$|\rho\gamma - 1|_{\mathfrak{p}} < \delta.$$

We have two cases to consider: \mathfrak{p} is finite or infinite.

4.1. \mathfrak{p} infinite. Just as before we deduce that

$$|\log |\rho\gamma|_{\mathfrak{p}}| \leq \delta' = \begin{cases} \log \frac{1}{1-\delta}, & \mathfrak{p} \text{ real,} \\ \frac{1}{2} \log \frac{1}{1-\sqrt{\delta}}, & \mathfrak{p} \text{ complex.} \end{cases}$$

We also have, with obvious notation,

$$|\text{Arg}((\rho\gamma)^{(\mathfrak{p})})| \leq \arccos \sqrt{1 - \delta} = \delta''.$$

As $\mathfrak{p} \in S$, we can write $\mathfrak{p} = \mathfrak{q}_j$ for some value of j . Consider the sublattice of $\mathbb{R}^{\#S+1}$ generated by the columns of the matrix A which is obtained from the matrix

$$\frac{1}{\log R} \begin{pmatrix} \log |\epsilon_1|_{\mathfrak{q}_1} & \dots & \log |\epsilon_t|_{\mathfrak{q}_1} & 0 \\ \vdots & & \vdots & \vdots \\ \log |\epsilon_1|_{\mathfrak{q}_{\#S}} & \dots & \log |\epsilon_t|_{\mathfrak{q}_{\#S}} & 0 \\ 0 & \dots & 0 & 0 \end{pmatrix} \in \mathbb{R}^{(\#S+1) \times (t+1)}.$$

by replacing the j^{th} row by

$$\frac{1}{\delta'} (\log |\epsilon_1|_{\mathfrak{q}_j}, \dots, \log |\epsilon_t|_{\mathfrak{q}_j}, 0),$$

and the last row by

$$\frac{1}{\delta''} (\text{Arg}(\epsilon_1^{(\mathfrak{p})}), \dots, \text{Arg}(\epsilon_t^{(\mathfrak{p})}), \text{Arg}(\zeta^{(\mathfrak{p})})).$$

Note that we expect the j^{th} and last row of A to have much larger entries than the other rows. Also consider the vector \vec{y} obtained from the vector

$$\frac{1}{\log R} (-\log |\rho|_{\mathfrak{q}_1}, \dots, -\log |\rho|_{\mathfrak{q}_t}, 0)^t \in \mathbb{R}^{t+1}.$$

by replacing the j^{th} element by $-\log |\rho|_{\mathfrak{q}_j} / \delta'$ and the last element by

$$\text{Arg}((1/\rho)^{(\mathfrak{p})}) / \delta''.$$

We then have, if \vec{x} is the lattice vector $A(c_1, \dots, c_t, c_0)^t$,

$$\|\vec{x} - \vec{y}\|^2 = \frac{\log^2 |\rho\gamma|_{\mathfrak{p}}}{\delta'^2} + \frac{\text{Arg}^2((\rho\gamma)^{(\mathfrak{p})})}{\delta''^2} + \sum_{\mathfrak{q} \in S, \mathfrak{q} \neq \mathfrak{p}} \frac{\log^2 |\rho\gamma|_{\mathfrak{q}}}{\log^2 R} \leq \#S + 1.$$

We can then combine a variant of the Fincke-Pohst algorithm [2] with the sieving ideas of [4] to determine all elements in $T_{j,\mathfrak{p}}(H_i, R_i, R_{i+1})$.

4.2. \mathfrak{p} finite. We proceed as before, but now the ideal \mathfrak{p} allows us to alter the generators we are using. As before, we have

$$|\rho\gamma - 1|_{\mathfrak{p}} < \delta < 1,$$

and so $\text{ord}_{\mathfrak{p}}(\rho\gamma) = 0$. So we again can reduce to one of a finite set of similar problems where $\mu_0 \prod_{i=1}^s \mu_i^{m_i} = \rho\gamma$, with $\text{ord}_{\mathfrak{p}}(\mu_i) = 0$. Suppose \mathfrak{p} has residue degree f and lies above the rational prime p . We put $q = p^f$ and choose n to be an integer such that

$$\delta \leq q^{-n}.$$

As $|\mu_0 \prod_{i=1}^s \mu_i^{m_i} - 1|_{\mathfrak{p}} = |\rho\gamma - 1|_{\mathfrak{p}} < \delta$, we have

$$\mu_0 \prod_{i=1}^s \mu_i^{m_i} \equiv 1 \pmod{\mathfrak{p}^n}.$$

Let M denote the subgroup of K^* generated by μ_0, \dots, μ_s . Now, as $\text{ord}_{\mathfrak{p}}(\mu_i) = 0$ for all i , we can consider the group M/\mathfrak{p}^n . Using an algorithm like the ones in [1] or [6], one can determine the group structure of the M/\mathfrak{p}^n as a product of cyclic groups $C_1 \times \dots \times C_g$. These two algorithms are based on the Baby-Step/Giant-Step strategy of Shanks and Pollard’s Rho method respectively. However these algorithms are far too general to work in a fast and efficient manner in our problem.

Instead we first compute the orders of μ_i in M/\mathfrak{p}^n . This can be done very quickly, assuming p is “small”, as the orders must be equal to a p^{th} power times a divisor of $q - 1$. All that is then required, to determine the group structure, is a lattice enlarging procedure to determine the full lattice of relations given the sublattice given by the relations $\mu_i^{h_i} \equiv 1 \pmod{\mathfrak{p}^n}$ for some h_i . Such a lattice enlarging procedure is given in [6], as algorithm **MINIMIZE**. It seems to work very well in practice although its complexity is worse than $O(|M/\mathfrak{p}^n|)$, but for smooth group orders the method works very fast.

We can then map the equation

$$\mu_0 \prod_{i=1}^s \mu_i^{m_i} \equiv 1 \pmod{\mathfrak{p}^n}$$

to an equivalent equation in $C_1 \times \dots \times C_g$. We therefore generate a set of congruence conditions on the exponents m_i modulo the orders of the groups C_i . Using these congruence conditions we can now write

$$\rho\gamma = \mu'_0 \prod_{i=1}^s \mu_i'^{n_i}$$

for some new values $\mu'_i \in K^*$. Let S' denote the support of the elements μ'_1, \dots, μ'_s . Clearly $S' \subset S$. We now proceed in a similar manner to the case of infinite places; Consider the sublattice of $\mathbb{R}^{\#S'}$ generated by the columns of the matrix

$$A = \frac{1}{\log R} \begin{pmatrix} \log |\mu'_1|_{\mathfrak{q}_1} & \dots & \log |\mu'_s|_{\mathfrak{q}_1} \\ \vdots & & \vdots \\ \log |\mu'_1|_{\mathfrak{q}_{\#S'}} & \dots & \log |\mu'_s|_{\mathfrak{q}_{\#S'}} \end{pmatrix} \in \mathbb{R}^{\#S' \times s}.$$

Also consider the vector

$$\vec{y} = \frac{1}{\log R} \left(-\log |\mu'_0|_{\mathfrak{q}_1}, \dots, -\log |\mu'_0|_{\mathfrak{q}_{\#S'}} \right)^t \in \mathbb{R}^{\#S'}$$

We then have, if \vec{x} is the lattice vector $A(n_1, \dots, n_s)^t$,

$$\|\vec{x} - \vec{y}\|^2 = \sum_{\mathfrak{q} \in S'} \frac{\log^2 |\rho\gamma|_{\mathfrak{q}}}{\log^2 R} \leq \#S'$$

We can then determine as before all the elements in $T_{j,p}(H_i, R_i, R_{i+1})$ using the Fincke-Pohst algorithm.

5. EXAMPLE

We now consider one of the examples from [5]. Let K denote the number field generated by θ , where

$$\theta^8 + 1 = 0.$$

The unit rank of \mathcal{O}_K^* is three, and as generators of infinite order we can take

$$\eta_1 = \theta^2 + \theta^4 + \theta^6, \eta_2 = -(\theta^2 + \theta^3 + \theta^4), \eta_3 = 1 + \theta^3 - \theta^5.$$

The element $\xi = -\theta^7$ generates the sixteen roots of unity in K . There is one prime ideal, \mathfrak{t} , lying above (2), and it has ramification index eight. This ideal is principal, and as a generator we can take $\pi = 1 - \theta$. In [5], as part of a much larger computation, it was necessary to compute the 795 solutions to the unit equation

$$\tau_1 + \tau_2 + 1 = 0,$$

where

$$\tau_1 = \xi^{a_0} \eta_1^{a_1} \eta_2^{a_2} \eta_3^{a_3} \pi^{a_4}, \tau_2 = \xi^{b_0} \eta_1^{b_1} \eta_2^{b_2} \eta_3^{b_3} \pi^{b_4}.$$

Clearly we can assume that $0 \leq a_0, b_0 \leq 15$, whilst for $i = 1, \dots, 4$ we can determine that we must have $|a_i|, |b_i| \leq 1066 = H_0$. Using a sieving technique alone it took around 27 MIPS years to compute all the solutions to the S -unit equation. This meant having to run a network of workstations on this problem for nearly three weeks.

We apply the method of this paper and determine $R_0 = 10^{3598}$ and $c_1 = 1.63189$. Hence by Lemma 1 we have that $\mathcal{L} = \mathcal{L}_{H_0}(R_0)$. If we set $R_1 = 10^{90}$, then it is easy to determine, using Section 3, that the sets $T_{i,p}(H_0, R_0, R_1)$ are empty for $i = 1, \dots, 4$ and $\mathfrak{p} \in M_K^\infty$. A similar result holds for the finite place \mathfrak{t} once we compute that the 2-adic logarithms of our fundamental units are given by

$$\begin{aligned} \log_2(\eta_1) &= 186899879855629\theta^6 + 59390724766195\theta^2 \\ &\quad + 351843720888320 + O(2^{50}), \\ \log_2(\eta_2) &= 65657308478134\theta^7 + 195695972877837/2\theta^6 + 54554746468923\theta^5 \\ &\quad + 55396416308677\theta^3 + 24206352677363/2\theta^2 + 79478226388298\theta \\ &\quad + 43980465111040 + O(2^{50}), \\ \log_2(\eta_3) &= 90580788397509\theta^7 + 94575096855027/2\theta^6 + 162414331722422\theta^5 \\ &\quad + 57487993832778\theta^3 + 345229554255373/2\theta^2 + 129321537157691\theta \\ &\quad + 43980465111040 + O(2^{50}). \end{aligned}$$

Hence from Lemma 2 we conclude that $\mathcal{L} = \mathcal{L}_{H_1}(R_1)$, where $H_1 = 338$. The total time needed to compute this reduction was less than one second. We however have

a problem in carrying out this step again, using the LLL-based method of Section 3, to show that the sets $T_{i,t}(H_1, R_1, R_2)$ are trivial for some R_2 . This is because of inequality (3), which means that we must choose a constant u in the algorithm such that

$$u \leq c_7 = c_5 - c_6 = \frac{-\log \delta}{8 \log 2} - 11,$$

with $\delta \approx 1/R_2$. For the values of δ now under consideration this means that u must be chosen too small to be of any use. For the sets $T_{i,t}(H_1, R_1, R_2)$ we must therefore use the method based on the Fincke-Pohst algorithm.

For our second application of Lemma 2 we choose $R_2 = 10^{30}$. The LLL-based technique of Section 3 allows us to show that the sets $T_{i,p}(H_1, R_1, R_2)$ are trivial for the places not equal to \mathfrak{t} . However, for $T_{i,t}(H_1, R_1, R_2)$ we need to use the method of Section 4. This means that we must determine all the solutions to the following problem:

$$\begin{aligned} |\log |\rho|_{\mathfrak{p}}| &\leq \log(1 + R_1) \leq 208 \text{ for } \mathfrak{p} \in M_K^\infty, \\ |\rho - 1|_{\mathfrak{t}} &< 10^{-30}, \end{aligned}$$

where

$$\rho = \zeta^{c_0} \eta_1^{c_1} \eta_2^{c_2} \eta_3^{c_3} \pi_4^{c_4}$$

and $|c_1|, |c_2|, |c_3|, |c_4| \leq 2 \times H_1 = 676$. Clearly we must then have $c_4 = 0$ and $\rho \equiv 1 \pmod{\mathfrak{p}^{99}}$. This is much too high an exponent to work with, so we try to determine the larger set of all ρ with

$$\rho = \zeta^{c_0} \eta_1^{c_1} \eta_2^{c_2} \eta_3^{c_3} \equiv 1 \pmod{\mathfrak{p}^{44}}.$$

Using the group theoretic algorithms mentioned previously, we determine the structure of the group M/\mathfrak{p}^{44} , where $M = \langle \zeta, \eta_1, \eta_2, \eta_3 \rangle$. It is easy to determine that

$$\zeta^{16} \equiv \eta_1^{32} \equiv \eta_2^{64} \equiv \eta_3^{64} \equiv 1 \pmod{\mathfrak{p}^{44}},$$

and we can then determine the group structure in under two seconds using the lattice enlarging procedure of [6]. The group turns out to be isomorphic to $C_{64} \times C_{32} \times C_{32} \times C_{16}$, and we deduce that

$$\rho = (\eta_1^{32})^{d_1} (\eta_2^{64})^{d_2} (\eta_2^{32} \eta_3^{32})^{d_3},$$

for some integers d_1, d_2, d_3 . Using the four inequalities $|\log |\rho|_{\mathfrak{p}}| \leq \log(1 + R_1) \leq 208$ for $\mathfrak{p} \in M_K^\infty$, we can then determine that there are no non-trivial elements in $T_{i,t}(H_1, R_1, R_2)$ for $i = 1, \dots, 4$ using the Fincke-Pohst algorithm, which takes about a second of computing time.

We are therefore left, by Lemma 2, with determining the solutions in $\mathcal{L}_{H_2}(R_2)$ with $H_2 = 112$. We now choose $R_3 = 10^{15}$, and none of the LLL based methods now work. For the finite place the computation in the previous paragraph will suffice, as $2^{-44} > 10^{-15}$. For the infinite places the application of the Fincke-Pohst based method of Section 4 allows us to show, in about one second, that the sets $T_{i,p}(H_2, R_2, R_3)$ contain only the trivial elements. Hence we need only consider the solutions in $\mathcal{L}_{H_3}(R_3)$ where $H_3 = 56$.

We now set $R_4 = 10^6$, and when considering the finite place we now need to look at all solutions of

$$\begin{aligned} |\log |\rho|_{\mathfrak{p}}| &\leq \log(1 + R_3) \leq 35 \text{ for } \mathfrak{p} \in M_K^\infty, \\ \rho &\equiv 1 \pmod{\mathfrak{t}^{20}}, \end{aligned}$$

where

$$\rho = (\eta_1^4)^{d_1} (\eta_2^{16})^{d_2} (\eta_3^8)^{d_3} (\eta_2^{12} \eta_3^4 \zeta^4)^{d_0}$$

with $d_0 \in \{0, \dots, 3\}$ and $d_i \in \mathbb{Z}$. It then takes a few seconds to determine all the elements in $T_{i,t}(H_3, R_3, R_4)$ for $i = 1, \dots, 4$ using the Fincke-Pohst algorithm. For the infinite places we apply the method of Section 4 and determine in under five seconds that there are no non-trivial elements in $T_{i,p}(H_3, R_3, R_4)$ for $i = 1, \dots, 4$ and $\mathfrak{p} \in M_K^\infty$. So we can conclude that we need only consider the set $\mathcal{L}_{H_4}(R_4)$ where $H_4 = 22$.

Finally we perform the whole process again for $R_5 = 10^3$. Once again the sets $T_{i,p}(H_4, R_4, R_5)$ are empty for $i = 1, \dots, 4$ and $\mathfrak{p} \in M_K^\infty$. The sets for the finite place \mathfrak{t} are non-trivial but can be determined in a matter of seconds. We are finally left with enumerating the set $\mathcal{L}_{H_5}(R_5)$ for $H_5 = 11$. Enumerating this set can be accomplished using an adaption of the methods in Section 4.

We conclude that we can compute all the solutions to the S -unit equation above in a matter of minutes rather than MIPS years as was previously the case.

REFERENCES

- [1] J. Buchmann, M. Jacobson, and E. Teske. On some computational problems in finite abelian groups. *Math. Comp.*, 66:1663–1687, 1997. MR **98a**:11185
- [2] U. Fincke and M. Pohst. Improved methods for calculating vectors of short length in a lattice, including a complexity analysis. *Math. Comp.*, 44:463–471, 1985. MR **86e**:11050
- [3] K. Györy. On the number of solutions of linear equations in units of an algebraic number field. *Comment. Math. Helvetici*, 54:585–600, 1979. MR **81g**:11031
- [4] N.P. Smart. The solution of triangularly connected decomposable form equations. *Math. Comp.*, 64:819–840, 1995. MR **95f**:11110
- [5] N.P. Smart. S -unit equations, binary forms and curves of genus 2. *Proc. London Math. Soc.*, 75:271–307, 1997. MR **98d**:11072
- [6] E. Teske. A space efficient algorithm for group structure computation. *Math. Comp.*, 67:1637–1663, 1998. MR **99a**:11146
- [7] N. Tzanakis and B.M.M. de Weger. How to explicitly solve a Thue-Mahler equation. *Compositio Math.*, 84:223–288, 1992; 89 (1993), 241–242. MR **93k**:11025; MR **95a**:11030
- [8] B.M.M. de Weger. Solving exponential diophantine equations using lattice basis reduction algorithms. *J. Number Theory*, 26:325–367, 1987; 31 (1989), 88–89. MR **88k**:11097; MR **90a**:11040
- [9] B.M.M. de Weger. *Algorithms for Diophantine Equations*. Centre for Mathematics and Computer Science Amsterdam, 1989. CWI-Tract 65. MR **90m**:11205
- [10] K. Wildanger. *Über das Lösen von Einheiten- und Indexformgleichungen in algebraischen Zahlkörpern mit einer Anwendung auf die Bestimmung aller ganzen Punkte einer Mordellschen Kurve*. PhD thesis, Technischen Universität Berlin, 1997.

HEWLETT-PACKARD LABORATORIES, FILTON ROAD, STOKE GIFFORD, BRISTOL, BS12 6QZ, U.K.

E-mail address: nsma@hplb.hpl.hp.com