Continued fractions and Parallel SQUFOF

S. McMath, F. Crabbe, D. Joyner

7-2-2005

Contents

1 Introduction

Integer factoriztion has long been an important problem in number theory, and with developments in computing and cryptography, its importance continues to rise. Though there are many fast algorithms for factoring numbers, this paper focuses on the square forms factorization (SQUFOF) algorithm (see Algorithm 4 below for a precise description). Daniel Shanks developed

SQUFOF in the 1970's, and it is still the fastest known algorithm for factoring integers in the 20- to 30-digit range. SQUFOF is currently used in conjunction with more recent algorithms that need to factor 20-digit numbers in order to generate their results.

Most of the Shanks' original work on SQUFOF was not published (see however, [Sh1]) and his notes are incomplete[1]. One purpose of this paper is to present Shanks's original SQUFOF algorithm in its entirety for the first time. The paper goes on to present several interesting results concerning both traditional SQUFOF and its parallelization.

1.1 Main results

This paper contains 3 new results:

1. A proof that the two-sided continued fraction of the normalized square root (an important part of the SQUFOF algorithm) has several very attractive properties - periodicity, a symmetry point corresponding to a factorization of N, and so on (see Theorems 6, 8, and 9 for details).

2. A proof of the infrastructure distance formula, Theorem 11 below, which is also an important part of SQUFOF. This is in some sense well-known but a proof has not, as far as we can see, appeared in the literature.

3. Empirical results comparing two techniques for parallelization of SQUFOF, showing that while the *multipliers* method is superior for small numbers of processors, it becomes less efficient per processor as the number of processors increases. The *segments* method maintains its efficiency per procesor as the number of processors increases, and thus is predicted to be superior for large numbers of processors.

2 Continued Fractions and Quadratic Forms

The stepping stone for SQUFOF is the continued fraction expansion for the square root of N. (We slightly simplify matters by instead using the "normalized square root (4) here.) The terms of this continued fraction expansion give rise to a sequence of quadratic forms of discriminant N via (5). We

[1] These notes have been typed in and are available on the web [Sh2], [Sh3], [Sh4].

shall describe SQUFOF in terms of the "cycle" of continued fractions in the periodic expansion of (4) and the corresponding quadratic forms.

2.1 Integral binary quadratic forms

There is a "dictionary" between certain aspects of

- indefinite integral binary quadratic forms,
- ideals in a real quadratic number field,
- the simple continued fraction of quadratic surds.

The reader will be assumed to be familiar with at least the basic aspects of this correspondence. For details, see for example, Buell [Bu], Lenstra [Len], Williams [W] (especially pp. 641-645), Cohen [Coh] and the references found there, or [M].

A **binary quadratic form** (or simply a "form") is a homogeneous form of degree two in two variables x, y,

$$f(x,y) = ax^2 + bxy + cy^2 = (x,y) \cdot \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix},$$

for some constants a, b, c. This form shall also be denoted by the triple (a, b, c). The **discriminant**[2] of f is $D = disc(f) = b^2 - 4ac$. In this note, we shall focus on the case $D > 0$, in which case the form is called **indefinite**. *From now on, we assume without further mention that $D > 0$ is a non-square such that $D \equiv 0 \pmod 4$ or $D \equiv 1 \pmod 4$.*

If $a, b, c \in \mathbb{Z}$ then we say f is **integral**. If moreover $gcd(a, b, c) = 1$ then we say the form is **primitive**. Let $F(D)$ denote the set of all integral forms of discriminant D and let $F(D)_p$ denote the subset of primitive ones.

The groups

$$GL_2(\mathbb{Z}) = \{\gamma = \begin{pmatrix} s & t \\ u & v \end{pmatrix} \mid s, t, u, v \in \mathbb{Z},\ \det(\gamma) = \pm 1\},$$

and

$$SL_2(\mathbb{Z}) = \{\gamma \in GL_2(\mathbb{Z}) \mid \det(\gamma) = 1\}$$

act on the polynomials $\mathbb{Z}[x, y]$ via

[2] Sometimes also called the **determinant** of f.

$$\gamma = \begin{pmatrix} s & t \\ u & v \end{pmatrix} : (x,y) \longmapsto (sx+ty, ux+vy).$$

Therefore, they also act on the set of integral forms via

$$(\gamma^* f)(x,y) = f(sx+ty, ux+vy),$$

for $\gamma \in GL_2(\mathbb{Z})$. In terms of the symmetric matrix $A = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}$ associated to the form f, this action may be epressed as

$$\gamma^*(A) = {}^t\gamma \cdot A \cdot \gamma.$$

We say that two forms f_1, f_2 are **equivalent** if $f_2 = \gamma^* f_1$, for some $\gamma \in GL_2(\mathbb{Z})$. We say that two forms f_1, f_2 are **properly equivalent**, written $f_1 \sim f_2$, if $f_2 = \gamma^* f_1$, for some $\gamma \in SL_2(\mathbb{Z})$. For $f \in F(D)$, we let

$$F(D)_f = [f] = \{f' \in F(D) \mid f \sim f'\}$$

denote the proper equivalence class of f. An element $\gamma \in GL_2(\mathbb{Z})$ is called an **automorph** of f if $\gamma^* f = f$. A form f is called **ambiguous** if it has an automorph in $GL_2(\mathbb{Z}) - SL_2(\mathbb{Z})$. Note that if $f \in F(D)$ is ambiguous then each $f' \in [f]$ is also ambiguous.

We say that two forms $(a_1, b_1, c_1), (a_2, b_2, c_2) \in F(D)$ are **adjacent** if $c_1 = a_2$ and $b_1 + b_2 \equiv 0 \pmod{2a_2}$. In this case, we say that (a_2, b_2, c_2) is to the **right of** (a_1, b_1, c_1) ((a_1, b_1, c_1) is to the **left of** (a_2, b_2, c_2)).

2.1.1 Reduction

A form (a, b, c) is called **reduced** if $|D^{1/2} - 2|a|| < b < D^{1/2}$. Let $F(D)_{red}$ denote the subset of reduced forms of discriminant D.

Lemma 1 *(a) Given any $f \in F(D)_{red}$ there is a unique $f' \in F(D)_{red}$ adjacent to the right of f and a unique $f'' \in F(D)_{red}$ adjacent to the left of f.*

(b) There are exactly two reduced ambiguous forms in a cycle of reduced forms in an ambiguous class.

For (a) see Buell [Bu], page 23; for (b), see [Bu], Theorem 9.12. Lemma 1 allows us to define the **cycle** of reduced forms associated to $f \in F(D)_{red}$: it is the set of all $f' \in F(D)_{red}$ which is adjacent to the left or right of f. This cycle is denoted $F(D)_{red,f}$.

Lemma 2 *An ambiguous equivalence class contains two points of symmetry, that is, pairs of reduced adjacent forms, (c, b, a) to the left of (a, b, c), in the cycle that are the symmetric reverse of each other. In that case, either a divides the determinant, or $a/2$ divides the determinant.*

This follows from Theorem 9 below.

It is evident that if a form is ambiguous, then each form in its equivalence class is also ambiguous.

Proposition 3 *The set $F(D)_{red}$ of reduced forms can be partitioned into cycles of adjacent forms.*

Consider the action of

$$T_m = \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix}$$

on a form (a, b, c): $T_m(a, b, c) = (a', b', c')$, where $a' = a$, $b' = b + 2am$, $c' = \frac{(b')^2 - D}{4a'}$. This defines a map $T_m : F(D) \to F(D)$, for each $m \in \mathbb{Z}$.

Consider the action of

$$W = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

on a form (a, b, c): $W(a, b, c) = (a', b', c')$, where $a' = c$, $b' = -b$, $c' = a$. This defines a map $W : F(D) \to F(D)$.

Algorithm 1 *(Reduction)*

Input: $f \in F(D)$.

Output: $f' \in F(D)_{red}$ *with* $f \sim f'$.

Let $f(x, y) = ax^2 + bxy + cy^2$ *and let*

$$J_{a,D} = \{x \mid -|a| < x < |a|, \text{ if } |a| \geq D^{1/2},\ -2|a| < x < D^{1/2}, \text{ if } |a| < D^{1/2}\}.$$

1. *Apply T_m to (a, b, c) to obtain a form (a, b', c'), where $b' \in J_{a,D}$ and c' is chosen so that the new form has discriminant D.*

2. *If (a, b', c') is reduced then return $f'(x, y) = ax^2 + b'xy + c'y^2$. Otherwise, replace (a, b', c') by $W(a, b', c') = (c', -b', a)$ and go to step 1.*

According to Lagarias [L1], this has complexity $O(\log(max(|a|, |b|, |c|)))$.

Define the **adjacency map** $\rho : F(D) \to F(D)$ by

$$\rho(a, b, c) = (a', b', c'), \tag{1}$$

where $a' = c$, $b' \in J_{c,D}$, and $b' \equiv -b \pmod{2c}$, and c' is determined by the condition $disc(a', b', c') = D$. This defines a bijection $\rho : F(D)_{red} \to F(D)_{red}$.

Unfortunately, given $f \in F(D)$ with $D > 0$ there are usually several $f' \in F(D)_{red}$ which are properly equivalent to f. In other words, the cycle

$$F(D)_{red,f} = \{f' \in F(D)_{red} \mid f \sim f'\} = \{f' = \rho^n f \mid n \in \mathbb{Z}\}$$

can be rather large. Indeed, it is known that $|F(D)_{red,f}| = O(D^{1/2+\epsilon})$, for each $\epsilon > 0$, where the exponent $1/2$ is best possible (Lagarias [Len, L2]) and where the O-constant depends on ϵ.

2.1.2 Composition

The *composition* of forms has important properties for SQUFOF. The rules of composition are fairly general. A binary quadratic form F is called **a composition** of $f, g \in F(D)$ if it satisfies an equation such as

$$f(x, y)g(u, v) = F(B_1(x, y, u, v), B_2(x, y, u, v)), \tag{2}$$

where B_1 and B_2 are quadratic forms in x, y, u, v of a certain type. The exact conditions B_1, B_2 satisfy do not concern us here (see Cox [Cox] if you are curious and Gauss [G] if you are really curious). The point is that there may be more than one pair B_1, B_2 satisfying (2), so that the composition F is not unique. (However, the conditions on B_1, B_2 specified by Gauss do imply that, for a given $f, g \in F(D)$ any two such compositions must be equivalent to each other.) One way around this ambiguity is to specify a choice of B_1, B_2 and hence define F uniquely.

The idea described below was known in some form to Dirichlet and possibly Gauss.

Algorithm 2 *Input:* $(a_1, b_1, c_1), (a_2, b_2, c_2) \in F(D)$.

Output: A composition $(\frac{a_1a_2}{m^2}, B, \frac{(B^2-D)m^2}{4a_1a_2}) \in F(D)$.

1. *Compute* $m = gcd(a_1, a_2, \frac{b_1+b_2}{2})$. *(Since* $D = b_i^2 - 4a_ic_i$*, for* $i = 1, 2$*,* b_1 *and* b_2 *have the same parity.)*
2. *Solve the congruences*

$$\begin{aligned} a_2 mB &\equiv mb_1a_2 \pmod{2a_1a_2}, \\ a_1 mB &\equiv mb_2a_1 \pmod{2a_1a_2}, \\ \tfrac{b_1+b_2}{2} mB &\equiv m\tfrac{b_1b_2+D}{2} \pmod{2a_1a_2}, \end{aligned}$$

   simultaneously an integer B*. Choose the solution with smallest absolute value.*

See [Sh1] or [Bu] for a proof of the correctness of this algorithm. Buell [Bu] also provides the substitutions that would be needed for Gauss's definition of composition.

In other words, we define the **composition** of $(a_1, b_1, c_1), (a_2, b_2, c_2) \in F(D)$ to be the form resulting from the above algorithm:

$$(a_1, b_1, c_1) * (a_2, b_2, c_2) = (\frac{a_1a_2}{m^2}, B, \frac{(B^2 - D)m^2}{4a_1a_2}).$$

Remark 1 *The binary operation* $* : F(D) \times F(D) \to F(D)$ *is associative but not its "restriction"* $\# : F(D)_{red} \times F(D)_{red} \to F(D)_{red}$ *(where* $\#$ *is composition algorithm 2 followed by reduction algorithm 1).*

Let $f, g \in F(D)_{red}$ *be elements in the principal cycle of discriminant* D*. It was observed by D. Shanks (see §5 in Lenstra [Len]) that cycles enjoy a "coset-like property"* $\rho^k f \# \rho^\ell g = \rho^{a_{k,\ell}}(f \# g)$*, for some* $a_{k,\ell} \in \mathbb{Z}$ *. In particular, the principal cycle is closed under composition. Therefore, the the set of complete quotients of the continued fraction of such an* α *can be identified with a set closed under* $\#$*.*

For further discussion of this, see Lenstra [Len] (5.1).

The "structure" of a cycle has been termed the "infrastructure" of $F(D)$ *by D. Shanks.*

If $f, f', g, g', h \in F(D)$ then Gauss showed

(a) $(f * g) * h \sim f * (g * h)$, and

(b) $f \sim f'$ and $g \sim g'$ imply $f * g \sim f' * g'$.

These imply that the set of equivalence classes of forms of discriminant D is a group $C(D)$, called the **class group** of D. From the construction, it is clear that $f * g \sim g * f$, so $C(D)$ is abelian.

The following Theorem was known to Shanks, since SQUFOF depends essentially on it.

Theorem 4 *An equivalence class has order* 2 *or* 1 *in the class group if and only if it is ambiguous.*

Any form $(1, b, c) \in F(D)$ acts as the identity for $*$. The cycle of the identity is the **principal cycle** of forms. Any form f whose square $f^2 = f * f$ belongs to the principal cycle is an ambiguous form ([Bu], Corollary 4.9).

2.2 Continued fractions

Throughout, assume that $N \equiv 1 \pmod 4$ and is not a perfect square.

We shall only consider simple continued fractions here. In other words, if $\alpha \in \mathbb{R}$ is the number we want to compute the continued fraction of, let $x_0 = \alpha$, $b_0 = \lfloor \alpha \rfloor$, where $\lfloor x \rfloor$ denotes the **floor** of x, and, for $i > 0$, let

$$x_i = \frac{1}{x_{i-1} - b_{i-1}}, \qquad b_i = \lfloor x_i \rfloor . \tag{3}$$

The term x_i is called the i^{th} **complete quotient** of α and b_i is called the i^{th} **partial quotient** of α. The **simple continued fraction** of α is ([HW]):

$$\alpha = b_0 + \frac{1}{b_1 + \frac{1}{b_2 + \ldots}},$$

also written $[b_0, b_1, b_2, \ldots]$. We are only concerned with continued fractions of an irrational $\alpha \in K = \mathbb{Q}(\sqrt{N})$. In this case, the sequence $b_0, b_1, b_2, \ldots$ is eventually periodic.

For example, let

$$\alpha = \begin{cases} \frac{\sqrt{N} + \lfloor \sqrt{N} \rfloor - 1}{2}, & \lfloor \sqrt{N} \rfloor \text{ even,} \\ \frac{\sqrt{N} + \lfloor \sqrt{N} \rfloor}{2}, & \lfloor \sqrt{N} \rfloor \text{ odd.} \end{cases} \tag{4}$$

We call this α the **normalized square root of** N. The continued fraction sequence $b_0, b_1, \ldots$ is (purely) periodic. In general, the period of α is the size of the cycle associated to the identity in the class group (Buell [Bu], Theorem 3.18 (a)).

At each step in the continued fraction expansion, it is possible to simplify $x_i - b_i$ to the form $\frac{\sqrt{N} - P_i}{Q_i} \in [0, 1)$, where $P_i, Q_i \in \mathbb{Z}$ satisfy $P_i^2 \equiv N \pmod{Q_i}$. In general, if P, Q are positive integers and $x = \frac{\sqrt{N} + P}{Q}$ satisfies

$P^2 \equiv N \pmod{Q}$, $0 < P < \sqrt{N}$, $|\sqrt{N} - Q| < P$, then we say that x is **reduced**. It is known that if x, y are two such reduced numbers and $y = \gamma(x)$ (where $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})$ acts on $\hat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ by $\gamma(x) = \frac{ax+b}{cx+d}$) then y occurs in the simple continued fraction expansion of x as a complete quotient (and x occurs in the simple continued fraction expansion of y as a complete quotient). See Buell [Bu], Proposition 3.20 for a proof.

If P, Q are positive integers and $x = \frac{\sqrt{N}+P}{Q}$ then we associate to x the quadratic forms

$$f_- = (-Q/2, P, -\frac{P^2 - N}{2Q}), \quad f_+ = (Q/2, P, \frac{P^2 - N}{2Q}), \tag{5}$$

which have discriminant N. (We implicitly assume here that $\frac{P^2-N}{2Q} \in \mathbb{Z}$ and Q is even. Note that if x is reduced then so are $f_\pm$, and conversely.)

Lemma 5 *(H. Cohen [Coh], §5.7.1) The continued fraction expansion of the quadratic irrational corresponding to the unit reduced form is not only periodic but symmetric.*

What is the continued fraction analog of "adjacency" of forms? Applying the adjacency map (1) is roughly analogous to the "stepping" process of going from one complete quotient to the next in a continued fraction. See Williams §5 for a discussion of the the ideal-theoretic analog, at least for the case of the simple continued fraction of $\frac{-1+\sqrt{N}}{2}$.

One tool used by many different algorithms is the continued fraction expression for (4), where N is the number to be factored. This expression is calculated recursively: $x_0 = \alpha$, $b_0 = \lfloor x_0 \rfloor$, and using (3) in general. Observe that solving equation (3) for x_{i-1} gives $x_{i-1} = b_{i-1} + \frac{1}{x_i}$.

The recursive formulas are, for $i \geq 0$,

$$\begin{aligned} x_{i+1} &= \frac{1}{x_i - b_i} \\ &= \frac{Q_i}{\sqrt{N} - P_i} \\ &= \frac{\sqrt{N} + P_i}{Q_{i+1}} \\ &= b_{i+1} + \frac{\sqrt{N} - P_{i+1}}{Q_{i+1}}, \\ b_i &= \lfloor x_i \rfloor. \end{aligned} \tag{6}$$

Theorem 6 provides some well-known fundamental properties and identities of continued fractions.

Theorem 6 *([Ri])*

In the continued fraction expansion of (4), with $x_0 = \alpha$*, each* x_i *reduces to the form* $\frac{\sqrt{N}+P_{i-1}}{Q_i}$*, with unique* $Q_i, P_i \in \mathbb{Z}$ *satisfying*

(a) $N = P_i^2 + Q_i Q_{i+1}$*,*

(b) $P_i = b_i Q_i - P_{i-1}$*,*

(c) $b_i = \left\lfloor \frac{\lfloor\sqrt{N}\rfloor+P_{i-1}}{Q_i} \right\rfloor \geq 1$*,*

(d) $0 < P_i < \sqrt{N}$*,*

(e) $|\sqrt{N} - Q_i| < P_{i-1}$*,*

(f) Q_i *is an integer,*

(g) $Q_{i+1} = Q_{i-1} + b_i(P_{i-1} - P_i)$*.*

(h) This sequence is eventually periodic.

(i) $\left\lfloor \frac{\sqrt{N}+P_i}{Q_i} \right\rfloor = \left\lfloor \frac{\sqrt{N}+P_{i-1}}{Q_i} \right\rfloor = b_i$*.*

These denominators $\{Q_i\}$ will be referred to as **pseudo-squares**. (Indeed, for $i \geq 0$, if we write $[b_0, b_1, ...b_i] = \frac{A_i}{B_i}$ then $A_{i-1}^2 - B_{i-1}^2 N = (-1)^i Q_i$ and so $A_{i-1}^2 \equiv (-1)^i Q_i \pmod N$.)

Remark 2 *The fact that each* x_i *reduces to the form* $\frac{\sqrt{N}+P_{i-1}}{Q_i}$ *is important for computational efficiency because this together with (c) imply that floating point arithmetic is not necessary for any of these calculations. Also, by use of (b) and (g), the arithmetic used in this recursion is on integers* $< 2\sqrt{N}$*.*

Since the continued fraction is eventually periodic, it is reasonable to consider that when it loops around on itself, the terms being considered may have come from some terms "earlier" in the recursion. Lemma 7 shows that by exchanging these two related expressions, the direction is reversed. The algorithm for stepping a continued fraction expansion in the opposite direction will be precisely the same as the one for the forward direction, except that the numerator is changed first. Note that this same change (with the exception of c_0) could be achieved by merely changing the sign of P_{i-1}.

Lemma 7 *Let N, and, for $i \geq 0$, let x_i, b_i, P_i, Q_i be as in Theorem 6. Let $y_0 = \frac{\sqrt{N}+P_{i+1}}{Q_{i+1}}$ and let $c_0 = \lfloor y_0 \rfloor$. If we define, for $j \geq 1$, $y_j = \frac{1}{y_{j-1}-c_{j-1}}$, $c_{j-1} = \lceil y_{j-1} \rfloor$ then $c_0 = b_{i+1}$ and $y_j = \frac{\sqrt{N}+P_{i-j+1}}{Q_{i-j+1}}$, when $0 \leq j \leq i$.*

Using Lemma 7 to go backwards in the continued fraction expansion, denote the terms before x_0 as $x_{-1}, x_{-2}, \ldots$. The sequence $\{x_i \mid i \in \mathbb{Z}\}$ will be called the **two-sided continued fraction** of x_0. Define Q_{-i} and P_{-i} similarly, $i \geq 0$.

Theorem 8 *(a) With these conventions on the negative indices, Theorem 6 applies for all $i \in \mathbb{Z}$.*

(b) Define x_i as in Theorem 6, $i \in \mathbb{Z}$. There exists a positive integer π such that for all $i \in \mathbb{Z}$, $x_i = x_{i+\pi}$.

(c) Let $x_0 = \alpha$ such that $Q_0 \mid 2P_{-1}$ (as in equation (4)). The sequence of pseudo-squares is symmetric about Q_0, so that for all $i \in \mathbb{Z}$, $Q_i = Q_{-i}$.

This follows easily from the lemma above so the proof is omitted.

This demonstrates an important fact about continued fractions: that the direction of the sequences of pseudo-squares and residues can be reversed (i.e. the indices decrease) by making a slight change and applying the same recursive mechanism. The presence of one point of symmetry allows a proof that another point of symmetry exists and that a factorization of N may be obtained from this symmetry[3]:

Theorem 9 *Let $s = \lfloor \frac{\pi}{2} \rfloor$, where π is the period from Theorem 8. If π is even then (a) $Q_{s+i} = Q_{s-i}$, (b) $Q_s \neq Q_0$, (c) $P_s = P_{s-1}$, and (d) $Q_s \mid 2N$, for all $i \in \mathbb{Z}$. If π is odd then, for all $i \in \mathbb{Z}$,*

- *$Q_{s+i+1} = Q_{s-i}$, and*
- *either (a) $\gcd(Q_s, N)$ is a nontrivial factor of N, or (b) -1 is a quadratic residue of N.*

The argument for the first statement is in [W], pages 641-642. For an elementary proof of both statements, see [M].

[3] This was actually discovered in the opposite order. It was clear that ambiguous forms that met this criteria provided a factorization but was later realized that these same forms produced symmetry points. This was first noticed by Gauss [G] and first applied by Shanks [Sh4].

2.3 Infrastructure distance formula

For $m < n$, and for $\{x_i\}_{i\in\mathbb{Z}}$, the terms in the continued fraction in (6), Shanks defined **infrastructure distance** by

$$D(x_m, x_n) = \log\left(\prod_{k=m+1}^{n} x_k\right). \tag{7}$$

We abuse notation and write $D(F_m, F_n)$ as well for this quantity, where a form F corresponds to a term x in the continued fraction via the map $x \longmapsto f_+$ (5). Lenstra [Len] adds a term of $\frac{1}{2}\log(Q_n/Q_m)$ to this (where Q denotes the pseudo-square term of x), with the effect that the resulting formulas are slightly simplified but the proofs are more complicated and less intuitive. Definition 7 is also used by Williams in [W].

Since the quadratic forms are cyclic, in order for the distance between two forms to be measured consistently, it must be considered modulo the distance around the principal cycle.

Definition 10 Let π be the period of the principal cycle. The **regulator** R of the class group is the distance around the principal cycle, that is, $R = D(F_0, F_\pi)$.

Therefore, distance must be considered modulo R, so that D is a map from pairs of forms to the interval $[0, R) \subset \mathbb{R}$. The addition of two distances must be reduced modulo R as necessary.

Theorem 11 *(infrastructure distance formula) If $F_1 \sim F_k$ are equivalent forms and $G_1 \sim G_\ell$ are equivalent forms and $D_{\rho,1}$ is the reduction distance for $F_1 * G_1$ and $D_{\rho,2}$ is the reduction distance for $F_k * G_\ell$ and m_1 and m_k are the factors cancelled in each respective composition (Algorithm 2), then*

$$D(F_1\#G_1, F_k\#G_\ell) = D(F_1, F_k) + D(G_1, G_\ell) + D_{\rho,2} - D_{\rho,1} + \log(m_2/m_1)$$

proof: Here is a sketch. (For more details, see Theorem A.5.2 in [M].)

As each quadratic form is associated with a reduced lattice, an analysis of distance requires a connection between reduced lattices (see §3 of [W] for the definition of reduced lattice). *We use the notation of Williams [W] without further mention.*

If $\mathcal{L}$ denotes lattice in $\mathbb{Q}(\sqrt{N})$, let $L(\mathcal{L})$ denote the least positive integer contained in it.

Lemma 12 *(Lemma A.4.2 of [M]) Let I be a primitive ideal and let $\mathcal{L}$ denote the lattice corresponding to I. If $\mathcal{L}'$ is a lattice with basis $\{1, \xi\}$ and for some θ, $\theta\mathcal{L}' = \mathcal{L}$, then the ideal J corresponding to the lattice $\mathcal{L}'$ is a primitive ideal and*

$$(L(I)\theta)J = (L(J))I \tag{8}$$

The method of Voronoi (see for example [W]) is used to obtain a sequence of adjacent minima, corresponding to a sequence of reduced lattices. Consider a sequence of lattices $\mathcal{L}_1$, $\mathcal{L}_2, \cdots$ corresponding to ideals $K_1, K_2, \cdots$ corresponding to binary quadratic forms $F_1, F_2, \cdots$, corresponding to terms $x_1, x_2, \cdots$ in a continued fraction expansion (6). If, for two adjacent lattices in the sequence, ξ_i is defined by $\mathcal{L}_{i+1} = 1/\xi_i\mathcal{L}_i$, then the chain of adjacent minima of $\mathcal{L}_1$ are defined by $\theta_k = \prod_{i=1}^{k-1} \xi_i$, so $\theta_k\mathcal{L}_k = \mathcal{L}_1$ (see [W], §3). Distance between such lattices is then defined by

$$D(\mathcal{L}_k, \mathcal{L}_\ell) = \log(\theta_k/\theta_\ell) \tag{9}$$

and this definition of distance corresponds exactly to the definition given for quadratic forms (see [W], §6).

Although this definition has so far only been applied to reduced ideals (for the definition of reduced ideal, see for example [W] §2) and lattices, the reduction of ideals and lattices corresponding to quadratic form and continued fraction reduction is well known:

Lemma 13 *(Lemma A.5.1 in [M]) Let I be any primitive ideal in $\mathbb{Z}[\sqrt{N}]$. There exists a reduced ideal I_k and a $\theta_k \in I$ such that $(L(I)\theta_k)I_n = (L(I_k))I$.*

Here θ_k may be efficiently computed by Voronoi's method or by continued fractions. Then the reduction distance is defined by $D_\rho = -\log(\theta_k)$ and may be considered as the distance from I to I_k.

Let I_1 denote the ideal corresponding to the form F_1 in the usual way (as in [Len]), let J_1 be the ideal corresponding to G_1, and let K_1 denote the ideal corresponding to $F_1 * G_1$. We have that $(s)K_1 = I_1J_1$, for some s. Let K_j be a reduced ideal and $\lambda \in K_1$ such that

$$\lambda K_j = K_1. \tag{10}$$

Then K_j is the ideal corresponding to $F_1 \# G_1$.

Similarly, let I_k denote the ideal corresponding to the quadratic form F_k and J_ℓ be the ideal corresponding to the form G_ℓ. If H_1 denotes the ideal corresponding to the composition $F_k * G_\ell$, then $(t)H_1 = I_k J_\ell$, for some t. Let H be a reduced ideal and choose $\eta \in H_1$ such that $\eta H = H_1$. Then H corresponds to $F_k \# G_\ell$.

Let μ and ϕ be such that $\mu I_k = I_1$ and $\phi J_\ell = J_1$. Combining these equations, gives

$$K_j = K_1/\lambda = I_1 J_1/\lambda s = (\frac{\mu\phi}{\lambda s}) I_k J_\ell = (\frac{s\mu\phi}{\lambda t}) H_1 = (\frac{s\mu\phi\eta}{\lambda t}) H.$$

Set $\psi = \frac{s\mu\phi\eta}{\lambda t}$ and then $\psi H = K_j$, so that by (9),

$$\begin{aligned} D(K_j, H) &= -\log(\psi) = -\log(\mu) - \log(\phi) - \log(\eta) + log(\lambda) - \log(s/t) \\ &= D(I_1, I_k) + D(J_1, J_\ell) + D(H_1, H_j) - D(K_1, K_j) + \log(t/s), \end{aligned}$$

as desired. $\square$

Remark 3 *Shanks stated Square Forms Factorization has an expected run-time of $O(\sqrt[4]{N})$ (see Gower [Go] for a detailed discussion of this).*

We explain a related idea remarked on by H. Lenstra [Len], page 148.

The idea is to first compute the regulator R. This has complexity $O(N^{\frac{1}{5}+\epsilon})$, assuming the Riemann hypothesis [Len]. Now use the "baby-step giant-step" method (as discussed in §13 of [Len]) to get close to the symmetry point:

Algorithm 3 *(Baby-step giant-step)*
Input: N and R
Output: Factorization of N

1. *Compute the form F associated to the first or second steps of the continued fraction algorithm of the normalized square root of N, (4).*
2. ***while** F is not within R/4 of the symmetry point (where distance is judged using the distance formula in Theorem 11).*
   (a) *Store F in a Collection F_c*
   (b) *$F = F \# F$ (These are the "giant-steps")*
3. *Use the intermediate forms in F_c to compose with F until within $\log N$ of the symmetry point.*
4. *Using the forward and backward steps (see Theorem 8) of the continued fraction algorithm ("baby steps"), locate the symmetry point.*
5. *using Lemma 2 find a factorization of N.*

Steps 2, 3, and 4, each take $O(\log N)$, so that the factorization takes $O(N^{\frac{1}{5}+\epsilon})$.

3 SQUFOF

Formally, here is the algorithm for factoring N:

Algorithm 4 *(SQUFOF)*
Input: N.
Output: A factor of N

1. $Q_0 \leftarrow 1, P_0 \leftarrow \lfloor\sqrt{N}\rfloor, Q_1 \leftarrow N - P_0^2$
2. $r \leftarrow \lfloor\sqrt{N}\rfloor$
3. **while** $Q_i \neq$ *perfect square for some* i *even*
   (a) $b_i \leftarrow \left\lfloor \frac{r+P_{i-1}}{Q_i} \right\rfloor$
   (b) $P_i \leftarrow b_i Q_i - P_{i-1}$
   (c) $Q_{i+1} \leftarrow Q_{i-1} + b_i(P_{i-1} - P_i)$
   (d) **if** $i = 2^n$ *for some* n *Store* $(Q_i, 2 \cdot P_i)$
4. $F_0 = (\sqrt{Q_i}, 2 \cdot P_{i-1}, \frac{P_{i-1}^2 - N}{Q_i})$
5. *Compose* F_0 *with stored forms according to the binary representation of* $i/2$ *and store result to* F_0.
6. $F_0 = (A, B, C)$
7. $Q_0 \leftarrow |A|, P_0 \leftarrow B/2, Q_1 \leftarrow |C|$
8. $q_0 \leftarrow Q_1, p_0 \leftarrow P_0, q_1 \leftarrow Q_0$
9. **while** $P_i \neq P_{i-1}$ *and* $p_i \neq p_{i-1}$
   (a) *Apply same recursive formulas to* (Q_0, P_0, Q_1) *and* (q_0, p_0, q_1)
10. *If* $P_i = P_{i-1}$, *either* Q_i *or* $Q_i/2$ *is a nontrivial factor of* N.
11. *If* $p_i = p_{i-1}$, *either* q_i *or* $q_i/2$ *is a nontrivial factor of* N.

3.1 Proof

Let N, the number to be factored, not be a perfect square. Expanding the continued fraction for $\sqrt{N}$, let Q be the first square pseudo-square found on an even index. Let $r = \sqrt{Q}$. Let $F = (r^2, b, c)$ be the associated quadratic form. Then (r, b, rc), which reduces with reduction distance $D_\rho = 0$ to $G = (r, b', c')$ is a reduced quadratic form whose square is F. Therefore, by Theorem 4, G is ambiguous and thus has a symmetry point in its cycle.

Since by Theorem 11, $2D(G_s, G) = D(F_s, F) \pmod R$ where F_s is the symmetry point of the principal cycle with coefficient 1, $D(G_s, G) = D(F_s, F)/2 \pmod{R/2}$. Since the two points of symmetry are $R/2$ away from each other, this means that there is a symmetry point at distance

$D(F_s, F)/2$ behind G. Therefore, a point of symmetry may be found by reversing G and traveling this short distance. Now if the coefficient at this symmetry point is ± 1, then there would have been a pseudo-square in the continued fraction expansion equal to r somewhere before F. If the coefficient is 2, then this symmetry point could be composed with G to find $2r$ at an earlier point in the principal cycle. Therefore, if neither r nor $2r$ were encountered before F in the continued fraction expansion, then the symmetry point provides a nontrivial factor for N.

4 Parallel SQUFOF

With the large amount of computation required for factorization, the efficiency of a parallel implementation is especially important for factorization algorithms (see Brent [Br] for a survey and some terminology).

There have been proposed two ways to parallelize SQUFOF: using multipliers and using segments. We will discuss the segments method here. More information on the multipliers method can be found in Gower [Go].

4.1 Segments

The segments technique depends upon the ability to use composition to jump to arbitrary locations in the principal cycle. The cycle can be divided into multiple equal-sized sub-sequences and each sub-sequence can be searched by one of the processors. As recently as ANTS 2004, Pomerance suggested investigating parallel SQUFOF (personal communication; see also [W] page 645).

When factoring using SQUFOF parallelized by segments, we choose a quadratic form G several steps into the cycle and then square it several times (how many times is more an art than a science - it depends on the number of processors and their speed and wanting to have segments which finish fast but not too fast, say 20-30 in our case). Call the resulting form F. For $i \geq 1$, each F^{2i} is assigned to processor i as a beginning of another segment, $[F^{2i}$, $\rho(F^{2i})$, $\rho^2(F^{2i})$, .., $F^{2i+2}]$, where ρ is the adjacency map. When processor i finds a pseudo-square which is a perfect square, that form H may used to find the symmetry point as follows. Note $H = \rho^{2n}(F^{2i})$, for some n. First, take the square root of H and reverse it, call this H'. This is in a new cycle of quadratic forms. Next, compose H' with F^i, call it H''. Finally, compose

H'' with powers of G to bring it closer to the symmetry point.

Algorithm 5 *(Segment-based Parallel SQUFOF)*
Input: N
Output: A factor of N
Preparation:

1. $r \leftarrow \lfloor \sqrt{N} \rfloor$
2. $F_0 \leftarrow (1, 2r, N - r^2)$
3. *Cycle* F_0 *several steps forward.*
4. ***for*** $i = 1$ *to size (size is the logarithmic size of a segment.)*
   (a) $F_i \leftarrow F_{i-1} * F_{i-1}$
5. $F \leftarrow F_i$

Processor 0:

1. *Assign one processor to search from* F_0 *to* F_{size}.
2. $F_{start} \leftarrow F_{size}$,$F_{end} \leftarrow F_{size}^2$,$F_{rootS} \leftarrow F_{size-1}$,$F_{rootE} \leftarrow F_{size}$,$F_{step} \leftarrow F_{size-1}$
3. ***while*** *A factor hasn't been found*
   (a) *Wait for a processor to be free and send* F_{start}, F_{end}, *and* F_{rootS}.
   (b) $F_{start} \leftarrow F_{end}$,$F_{rootS} \leftarrow F_{rootE}$,$F_{rootE} \leftarrow F_{rootE} * F_{step}$,$F_{end} \leftarrow F_{rootE}^2$

Processor n*:*

1. *Receive* F_{start}, F_{end}, *and* F_{rootS}
2. *count* $\leftarrow 0$
3. $F_0 = (A, B, C)$
4. ***while*** *A factor is not found and* $F_{start} \neq F_{end}$
   (a) *Cycle* F_{start} *forward 2 steps.*
   (b) *count* $\leftarrow$ *count+1*
   (c) ***if*** A *is a perfect square*
      i. $F_{test} \leftarrow F_{start}^{-1/2}$
      ii. $F_{test} \leftarrow F_{test} * F_{rootS}$
      iii. ***for*** $j = size$ *to 1 (This loop composes* F_{test} *with the necessary forms to bring it close to the symmetry point.)*
         A. ***if*** *count* $> 2^j$
         B. $F_{test} \leftarrow F_{test} * F_j$
         C. *count* $\leftarrow$ *count* -2^j
         D. *Search in both directions from* F_{test} *for a symmetry point.*
         E. ***if*** *Factorization found at symmetry point, output and quit.*
5. ***if*** *A factor is still not found, receive new* F_{start}, F_{end}, *and* F_{rootS} *and start over.*

Since there is no overlap between the segments searched by the processors and since the perfect squares appear to be distributed evenly throughout the principal cycles, this parallelization should be efficient for any number of processors. There are two hazards when choosing selecting the size of the segment. If the segment size is too small, the processors will finish their segments so quickly that receiving new segments will become a bottleneck. Alternately, if the segments are too long, the processors may divide up more than the entire cycle, so that there is overlap. However, except for rare numbers that will factor fast regardless, there is significant room in between these two bounds.

Remark 4 *The segments based parallelization described here has been implemented in C using MPI and run on a 64 processor SGI Origin 2800. Detailed results and comparisons to the multipliers method can be found in McMath [M]. Initial results indicate that the segments method does indeed continue to be efficient when the number of processors is increased.*

5 Conclusion

This paper, aside from presenting SQUFOF in its entirety for the first time, has shown that the algorithm can be presented in terms of an elegent theoretical framework using two-sided continued fractions and class groups of quadratic forms over a real quadratic field. It further proved the infrastructure distance formula on the cycle of forms in the class group.

Acknowledgements

Daniel Shanks's hand-written notes were kindly made available to the authors by the executors of his papers (W. Adams, D. Buell, and H. Williams), to whom we are very grateful. We are also very grateful to S. Wagstaff and J. Gowers who kindly sent us Gower's recent PhD thesis [Go].

References

[Br] R. Brent, *Parallel algorithms for integer factorisation*, in **Number Theory and Cryptography** (edited by J. H. Loxton), pages 26-37, London Math. Soc. Lecture Note Series 154, Cambridge University Press, Cambridge, 1990.

[Bu] D. Buell, **Binary quadratic forms**, Springer-Verlag, 1989.

[Coh] H. Cohen, **Advanced topics in computational number theory**, New York: Springer-Verlag, 2000.

[Cox] D. Cox, **Primes of the form** $x^2 + ny^2$, Wiley Inter-Science, 1989.

[G] C. F. Gauss, **Disquisitiones Arithmeticae**, (1801), republished by Springer-Verlag, 1985.

[Go] J. Gower, *Square forms factorization*, PhD thesis, 2004, Purdue Univ. (advisor S. Wagstaff).

[HW] G. H. Hardy and E. M. Wright, **An introduction to the theory of numbers**, Oxford: Clarendon Press, 1979.

[L1] J. Lagarias, *Worst-case complexity bounds for algorithms in the theory of integral quadratic forms,* J. Algorithms $\underline{1}$ (1980) 142–186.

[L2] ——, "On the computational complexity of determining the solvability or unsolvability of the equation $X^2 - DY^2 = -1$," Trans. Amer. Math. Soc. $\underline{260}$ (1980) 485–508.

[Len] Lenstra, H. W., Jr., *On the calculation of regulators and class numbers of quadratic fields,* **Number theory days, 1980** (Exeter, 1980), 123–150, London Math. Soc. Lecture Note Ser., 56, Cambridge Univ. Press, Cambridge, 1982.

[M] S. McMath, *Parallel integer factorization using quadratic forms*, Trident Report, 2005.
`http://cadigweb.ew.usna.edu/~wdj/mcmath/`

[Ri] H. Riesel, **Prime numbers and computer methods for factorization**, Birkhäuser, Boston, 1985

[Sh1] D. Shanks, *On Gauss and composition II,* in **Number Theory and Applications**, 1989, pp. 179-204.

[Sh2] ——, *Analysis and improvement of the continued fraction method of factorization*, (circa 1975)
`http://cadigweb.ew.usna.edu/~wdj/mcmath/`

[Sh3] ——, *The infrastructure of a real quadratic field and its applications*, August 1972, Number Theory Conference: University of Colorado, Boulder, Colorado.

[Sh4] ——, *SQUFOF notes*, (circa 1975), `http://cadigweb.ew.usna.edu/~wdj/mcmath/`

[W] H. Williams, *Continued fractions and number-theoretic computations,* Rocky Mt. J. Math. 15(1985)621-655.