Daniel Shanks' Square Forms Factorization

Stephen McMath

November 24, 2004

This is a detailed expository account of Shanks' Square Forms Factorization (SQUFOF) method, including proofs or detailed references of all results, in light of an understanding of continued fractions, binary quadratic forms, lattices, and ideals.

Contents

> The problem of distinguishing prime numbers from composite numbers and of resolving the latter into their prime factors is known to be one of the most important and useful in arithmetic. It has engaged the industry and wisdom of ancient and modern geometers ... the dignity of the science itself seems to require that every possible means be explored for the solution of a problem so elegant and so celebrated [6].

There is a significant body of knowledge concerning quadratic forms, continued fractions, lattices, and ideals. However, much of this information is very spread out, especially that dealing with the class group "infrastructure". Therefore, one major purpose here is to organize this information into a usable form, along with providing some of the connecting historical information and providing sufficient examples so that the average reader may be able to understand the main concepts, if perhaps not all of the minutia of the proofs.

These are extremely rich fields, and there are problems and ideas that have yet to be addressed concerning binary quadratic forms and even concerning variations of SQUFOF that Shanks considered. This paper is a complete proof of the simplest case only.

Of these objects, continued fractions, described in §1, appear the most concrete and are the easiest to examine examples of, especially with respect to distance, and thus are useful for a conceptual understanding. §2 describes some of the direct applications of continued fractions and the problems that caused Shanks to develop the theory further. Quadratic forms, described in §3, are computationally the simplest format to implement and have given rise to composition, an extremely useful tool. Ideals, in §4 are valuable because they provide an alternate interpretation of composition and provide a link to lattices, described in §5, from which "distance" is derived. §6 uses lattices and ideals to prove Theorem 10, a powerful formula concerning infrastructure distance. §7 puts it all together to analyze Shanks' factorization algorithm.

This investigation will also define mappings between these different objects. These maps will be represented by the letter Φ, with subscripts indicating the two sets being considered. For example, $\Phi_{\mathbb{T},\mathbb{F}}$ would be a map from terms in the continued fraction to quadratic forms and $\Phi_{\mathbb{F},\mathbb{T}} = \Phi_{\mathbb{T},\mathbb{F}}^{-1}$ would be its inverse. Note that the order of the subscripts matters. Each of these maps will be addressed individually.

1 Continued Fractions

One tool used by many different algorithms is the continued fraction expression for $\sqrt{N}$, where N is the number to be factored. This expression is calculated recursively[7]:

$$x_0 = \sqrt{N},\ b_0 = \lfloor x_0 \rfloor \tag{1}$$

$$\forall i \geq 1\ x_i = \frac{1}{x_{i-1} - b_{i-1}},\ b_i = \lfloor x_i \rfloor \tag{2}$$

$$\sqrt{N} = b_0 + \frac{1}{b_1 + \frac{1}{b_2 + \ldots}} \tag{3}$$

Observe that solving equation (2) for x_{i-1} gives $x_{i-1} = b_{i-1} + \frac{1}{x_i}$. Repeatedly substituting this into itself gives equation (3).

Before developing the theory too much further, allow me to offer one example of how this works, just so that the pieces make sense to you. To simplify the expansion some, the integers taken out in the third step of each line are the b_i:

Example 1

$$\begin{array}{cc}
x_0 = \sqrt{41},\ b_0 = 6 & \sqrt{41} = 6 + \frac{1}{x_1} \\
x_1 = \frac{1}{\sqrt{41}-6} = \frac{\sqrt{41}+6}{5} = 2 + \frac{\sqrt{41}-4}{5} & \sqrt{41} = 6 + \frac{1}{2+\frac{1}{x_2}} \\
x_2 = \frac{5}{\sqrt{41}-4} = \frac{\sqrt{41}+4}{5} = 2 + \frac{\sqrt{41}-6}{5} & \sqrt{41} = 6 + \frac{1}{2+\frac{1}{2+\frac{1}{x_3}}} \\
x_3 = \frac{5}{\sqrt{41}-6} = \frac{\sqrt{41}+6}{1} = 12 + \frac{\sqrt{41}-6}{1} & \sqrt{41} = 6 + \frac{1}{2+\frac{1}{2+\frac{1}{12+\frac{1}{x_4}}}}
\end{array}$$

From this step, it is evident that $x_4 = x_1$ and the cycle repeats from here. Observe that if the sequence starts with $x_0 = \sqrt{41} + 6$, then also $x_3 = x_0$. Regardless, the expression on the right, if truncated at any point provides a rational approximation to $\sqrt{41}$. Often this will be written as merely $[6, 2, 2, 12, ...]$ to save space or $[6, \overline{2, 2, 12}]$ to indicate that this part repeats. The various numbers on the left have some important properties that we will now analyze in some depth.

Throughout, assume that N is an odd positive integer and is not a perfect square. For number theory purposes, let

$$x_0 = \frac{\sqrt{N} + P_{-1}}{Q_0}$$

where P_{-1}, Q_0 are integers chosen such that

$$P_{-1}^2 \equiv N \pmod{Q_0}, \quad 0 < P_{-1} < \sqrt{N}, \quad \text{and } |\sqrt{N} - Q_0| < P_{-1}. \tag{4}$$

There are many ways of doing this[1]. The recursive formulas are:

$$x_{i+1} = \frac{1}{x_i - b_i} \quad b_i = \lfloor x_i \rfloor, \ i \geq 0$$

Formally, the assumed equation is:

$$x_{i+1} = \frac{Q_i}{\sqrt{N} - P_i} = \frac{\sqrt{N} + P_i}{Q_{i+1}} = b_{i+1} + \frac{\sqrt{N} - P_{i+1}}{Q_{i+1}}, \ i \geq 0 \tag{5}$$

Note that this equation serves as a definition of $Q_i, P_i, Q_{i+1}, P_{i+1} \in \mathbb{Q}$, so that these equations are true regardless of the conditions on these variables. Theorem 1 provides some well-known fundamental properties and identities of continued fractions. In [13], Hans Riesel provides very clear proofs of most of this.

Theorem 1 *[13] In the continued fraction expansion of x_0 satisfying (4), each x_i reduces to the form $\frac{\sqrt{N}+P_{i-1}}{Q_i}$, with (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, and (g) $Q_{i+1} = Q_{i-1} + b_i(P_{i-1} - P_i)$. Furthermore, (h) this sequence is eventually periodic.*

Proof:

(a) From (5), the equation $\frac{Q_i}{\sqrt{N}-P_i} = \frac{\sqrt{N}+P_i}{Q_{i+1}}$ requires that $N = P_i^2 + Q_i Q_{i+1}$.

(b) It is evident from simplifying the expression on the far right of (5) that

$$\frac{\sqrt{N} + P_i}{Q_{i+1}} = \frac{\sqrt{N} + b_{i+1}Q_{i+1} - P_{i+1}}{Q_{i+1}}.$$

Therefore, $P_{i+1} = b_{i+1}Q_{i+1} - P_i$.

(c) For $i = 0$, by the assumption $|\sqrt{N} - Q_0| < P_{-1}$

$$Q_0 < \sqrt{N} + P_{-1}$$

Therefore,

$$b_0 = \left\lfloor \frac{\sqrt{N} + P_{-1}}{Q_0} \right\rfloor \geq 1$$

For $i > 0$, $b_{i-1} = \lfloor x_{i-1} \rfloor$. By the definition of floor, $x_{i-1} - 1 < b_{i-1} \leq x_{i-1}$. If $b_{i-1} = x_{i-1}$, then the continued fraction $[b_0, b_1, ...b_{i-1}]$ is rational and is equal to $x_0 = \frac{\sqrt{N}+P_{-1}}{Q_0}$, which is irrational

[1] Choosing $x_0 = \sqrt{N} + \lfloor \sqrt{N} \rfloor$, so that $P_{-1} = \lfloor \sqrt{N} \rfloor$ and $Q_0 = 1$ is one possibility. Choosing $x_0 = \frac{\sqrt{N}+P_{-1}}{2}$, where $P_{-1} = \lfloor \sqrt{N} \rfloor$ or $\lfloor \sqrt{N} \rfloor - 1$, such that P_{-1} is odd, is another possibility.

since N is not a perfect square. Therefore, $x_{i-1} - 1 < b_{i-1} < x_{i-1}$, so that $0 < x_{i-1} - b_{i-1} < 1$. Therefore, $x_i = \frac{1}{x_{i-1}-b_{i-1}} > 1$, so that $b_i = \lfloor x_i \rfloor \geq 1$.

Note that there is no integer between $\lfloor\sqrt{N}\rfloor + P_{i-1}$ and $\sqrt{N} + P_{i-1}$, so it is trivial that $\left\lfloor \frac{\sqrt{N}+P_{i-1}}{Q_i} \right\rfloor = \left\lfloor \frac{\lfloor\sqrt{N}\rfloor+P_{i-1}}{Q_i} \right\rfloor$.

(d-e) The statements $|\sqrt{N} - Q_i| < P_{i-1}$ and $0 < P_{i-1} < \sqrt{N}$ may be proven inductively.

Base case: $i = 1$

$P_0 = \lfloor\sqrt{N}\rfloor$ or $\lfloor\sqrt{N}\rfloor - 1$, so by definition $0 < P_0 < \sqrt{N}$.

Since x_0 meets (4), $|\sqrt{N} - Q_0| < P_{-1}$.

Induction: Assume $|\sqrt{N} - Q_i| < P_{i-1}$ and $0 < P_{i-1} < \sqrt{N}$.

Note that these assumptions require that $0 < Q_i < 2\sqrt{N}$. From (c), $0 < x_i - b_i < 1$ means $0 < \frac{\sqrt{N}-P_i}{Q_i} < 1$. Since $Q_i > 0$, $0 < \sqrt{N} - P_i < Q_i$. From the left side of this, $P_i < \sqrt{N}$. Now, either $Q_i \leq \sqrt{N}$ or $Q_i > \sqrt{N}$.

Case 1: If $Q_i \leq \sqrt{N}$, then $\sqrt{N} - P_i < Q_i \leq \sqrt{N}$, so that $P_i > 0$.

Case 2: If $Q_i > \sqrt{N}$, then by (b), $P_i = b_iQ_i - P_{i-1} > b_i\sqrt{N} - \sqrt{N} = (b_i - 1)\sqrt{N} \geq 0$.

Therefore, $0 < P_i < \sqrt{N}$.

Since $x_{i+1} > 1$, it is trivial that $Q_{i+1} < \sqrt{N} + P_i$ so that showing $|\sqrt{N} - Q_{i+1}| < P_i$ reduces to showing $Q_{i+1} > \sqrt{N} - P_i$. Since $1 = \frac{N-P_i^2}{Q_iQ_{i+1}} = \frac{\sqrt{N}+P_i}{Q_i}\frac{\sqrt{N}-P_i}{Q_{i+1}}$, this is equivalent to showing:

$$\frac{\sqrt{N} + P_i}{Q_i} > 1. \tag{6}$$

Assume the contrary, that $Q_i \geq \sqrt{N} + P_i$. Then,

$$\begin{gathered} b_i(\sqrt{N} + P_i) - P_i \leq b_iQ_i - P_i = P_{i-1} < \sqrt{N}, \\ b_i\sqrt{N} + P_i(b_i - 1) < \sqrt{N}, \\ \sqrt{N}(b_i - 1) + P_i(b_i - 1) < 0, \\ (b_i - 1)(\sqrt{N} + P_i) < 0. \end{gathered}$$

But $\sqrt{N}$ and P_i are positive, so this implies $b_i < 1$, contradicting Theorem 1 (c). Therefore, (6) holds.

(f) The fact that $N = P_i^2 + Q_iQ_{i+1}$ requires that $Q_{i+1} = \frac{N-P_i^2}{Q_i}$. In order to show that $\forall i$ Q_i is an integer, the statements that Q_i is an integer and $Q_i \mid N - P_i^2$ may be proven inductively.

Base case: $i = 0$

By definition, Q_0 is an integer and $P_{-1}^2 \equiv N \pmod{Q_0}$. But $P_0 \equiv P_{-1} \pmod{Q_0}$, so $P_0^2 \equiv N \pmod{Q_0}$. Therefore, $Q_0 \mid N - P_0^2$.

Induction: Assume for some i, Q_i is an integer and $Q_i \mid (N - P_i^2)$. Then, since $N = P_i^2 + Q_iQ_{i+1}$, $Q_{i+1} = \frac{N-P_i^2}{Q_i}$, so that since $Q_i \mid (N - P_i^2)$, Q_{i+1} is an integer. Also, $Q_i = \frac{N-P_i^2}{Q_{i+1}}$, so that since Q_i is an integer, $Q_{i+1} \mid (N - P_i^2)$, so that $P_i^2 \equiv N \pmod{Q_{i+1}}$. Since $P_{i+1} = b_{i+1}Q_{i+1} - P_i$, $P_{i+1} \equiv P_i \pmod{Q_{i+1}}$. Therefore, $P_{i+1}^2 \equiv N \pmod{Q_{i+1}}$, so that $Q_{i+1} \mid (N - P_{i+1}^2)$ and the induction is complete.

(g) Solving (b) for b_i gives $\frac{P_{i-1}+P_i}{Q_i} = b_i$. Multiply by $(P_{i-1} - P_i)$ to obtain:

$$\frac{P_{i-1}^2 - P_i^2}{Q_i} = b_i(P_{i-1} - P_i)$$

Rearranging and adding $\frac{N}{Q_i}$ gives:

$$\frac{N-P_i^2}{Q_i} = \frac{N-P_{i-1}^2}{Q_i} + b_i(P_{i-1}-P_i)$$
$$Q_{i+1} = Q_{i-1} + b_i(P_{i-1}-P_i)$$

.

(h) Since each x_i and thus the entire sequence that follows it is defined by the two integers Q_i and P_{i-1}, limited by the bounds $0 < Q_i < 2\sqrt{N}$ and $0 < P_i < \sqrt{N}$, there is only a finite number of distinct x_i's. Therefore, for some π and some k, $\forall i \geq k\ x_i = x_{i+\pi}$. **QED**

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 $< 2\sqrt{N}$.

One application of continued fractions is rational approximations.

$$\sqrt{N} = b_0 + \frac{1}{b_1 + \frac{1}{b_2+...}}$$

If this continued fraction is truncated at any point, the result is an approximation to $\sqrt{N}$. One might imagine that it is necessary to start simplifying at the lower right end of this expression to obtain this approximation. However, Theorem 2, also included in [13], provides a simpler answer.

Theorem 2 *Let:*

$$A_{-1} = 1,\ A_0 = b_0,\ A_i = b_iA_{i-1} + A_{i-2},\ i > 0$$
$$B_{-1} = 0,\ B_0 = 1,\ B_i = b_iB_{i-1} + B_{i-2},\ i > 0$$

Then for $i \geq 0$, $[b_0, b_1, ...b_i] = \frac{A_i}{B_i}$ *and* $A_{i-1}^2 - B_{i-1}^2N = (-1)^iQ_i$.

Note that the last equation gives $A_{i-1}^2 \equiv (-1)^iQ_i \pmod N$. Although this equation will change some when generalized to other continued fractions, these denominators $\{Q_i\}$ will consistently be referred to as *pseudo-squares*. A proof of this theorem is given in [13]. Therefore, instead of reproducing the proof, I will provide an example:

Example 2

$$x_0 = \sqrt{403} + 20 = 40 + \sqrt{403} - 20$$
$$x_1 = \frac{1}{\sqrt{403}-20} = \frac{\sqrt{403}+20}{3} = 13 + \frac{\sqrt{403}-19}{3}$$
$$x_2 = \frac{3}{\sqrt{403}-19} = \frac{\sqrt{403}+19}{14} = 2 + \frac{\sqrt{403}-9}{14}$$
$$x_3 = \frac{14}{\sqrt{403}-9} = \frac{\sqrt{403}+9}{23} = 1 + \frac{\sqrt{403}-14}{23}$$
$$x_4 = \frac{23}{\sqrt{403}-14} = \frac{\sqrt{403}+14}{9} = 3 + \frac{\sqrt{403}-13}{9}$$
$$x_5 = \frac{9}{\sqrt{403}-13} = \frac{\sqrt{403}+13}{26} = 1 + \frac{\sqrt{403}-13}{26}$$
$$x_6 = \frac{26}{\sqrt{403}-13} = \frac{\sqrt{403}+13}{9} = 3 + \frac{\sqrt{403}-14}{9}$$
$$x_7 = \frac{9}{\sqrt{403}-14} = \frac{\sqrt{403}+14}{23} = 1 + \frac{\sqrt{403}-9}{23}$$

From this, a table may be used to recursively calculate the approximation[2]*:*

[2]Actually, in this case $b_0 = 40$, but since that would be an approximation to $x_0 = \sqrt{403} + 20$, subtracting 20 from b_0 yields an approximation to $\sqrt{403}$.

i	-1	0	1	2	3	4	5	6
b_i		20	13	2	1	3	1	3
A_i	1	20	261	542	803	2951	3754	14213
B_i	0	1	13	27	40	147	187	708

filling in A_i and B_i from left to right. From the last column, $\sqrt{403} \approx 14213/708$.

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. Example 2 provides some indication as to how the recursive formulas may be reversed, as $\{Q_i\}$ and $\{b_i\}$ are symmetric about x_5, so that after x_5 these numbers are cycled through in reverse order. Lemma 1 addresses how each b_i is calculated two different ways and Lemma 2 shows that by exchanging these two related expressions, the direction is reversed.

Lemma 1

$$\lfloor \frac{\sqrt{N}+P_i}{Q_i} \rfloor = \lfloor \frac{\sqrt{N}+P_{i-1}}{Q_i} \rfloor = b_i$$

Proof: The second part of this equation, that $\lfloor \frac{\sqrt{N}+P_{i-1}}{Q_i} \rfloor = b_i$ follows from the definition of b_i.

Theorem 1 (e) implies that $Q_i > \sqrt{N} - P_{i-1}$. Therefore,

$$\lfloor \frac{\sqrt{N}+P_i}{Q_i} \rfloor = \lfloor \frac{\sqrt{N}+b_iQ_i-P_{i-1}}{Q_i} \rfloor = b_i + \lfloor \frac{\sqrt{N}-P_{i-1}}{Q_i} \rfloor = b_i. \textbf{ QED}$$

Considering Example 2, it is then natural to suspect that the mechanism for going in the opposite direction will be precisely the same as the standard approach, 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 2 *Let x_i, b_i, P_i, Q_i, and N be as in Theorem 1, $i \geq 0$. Let $y_0 = \frac{\sqrt{N}+P_{i+1}}{Q_{i+1}}$ and let $c_0 = \lfloor y_0 \rfloor$. Define inductively $y_j = \frac{1}{y_{j-1}-c_{j-1}}$. Then $c_0 = b_{i+1}$ and $y_j = \frac{\sqrt{N}+P_{i-j+1}}{Q_{i-j+1}}$, $j \geq 0$.*

Proof: By (6) and Lemma 1, $c_0 = \lfloor y_0 \rfloor = \lfloor \frac{\sqrt{N}+P_{i+1}}{Q_{i+1}} \rfloor = b_{i+1}$. By mathematical induction it suffices to prove the case $j = 1$. Using Theorem 1

$$y_1 = \frac{1}{y_0 - c_0} = \frac{1}{\frac{\sqrt{N}+P_{i+1}}{Q_{i+1}} - b_{i+1}} = \frac{1}{\frac{\sqrt{N}+P_{i+1}-b_{i+1}Q_{i+1}}{Q_{i+1}}}$$

$$= \frac{1}{\frac{\sqrt{N}-P_i}{Q_{i+1}}} = \frac{\sqrt{N}+P_i}{\frac{N-P_i^2}{Q_{i+1}}} = \frac{\sqrt{N}+P_i}{Q_i} \textbf{ QED}$$

This demonstrates an important fact about continued fractions, the fact 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.

Using Lemma 2, x_3 may be used, for example, to find x_2 and x_1. Continuing this process, denote the terms before x_0 as $x_{-1}, x_{-2}, \ldots$. Define Q_{-i} and P_{-i} similarly[3]. Example 3 demonstrates this with the continued fractions from Example 2:

[3]Since y_0 meets (4) and the same recursive formula is applied, it is clear that Theorem 1 still applies to negative indices.

Example 3 $x_3 = \frac{\sqrt{403}+9}{23}$ *and* $P_3 = 14$*, so let* $y_0 = \frac{\sqrt{403}+14}{23}$ *to obtain*

$$\begin{aligned} y_0 &= \tfrac{\sqrt{403}+14}{23} = 1 + \tfrac{\sqrt{403}-9}{23} \\ y_1 &= \tfrac{23}{\sqrt{403}-9} = \tfrac{\sqrt{403}+9}{14} = 2 + \tfrac{\sqrt{403}-19}{14} \\ y_2 &= \tfrac{14}{\sqrt{403}-19} = \tfrac{\sqrt{403}+19}{3} = 13 + \tfrac{\sqrt{403}-20}{3} \\ y_3 &= \tfrac{3}{\sqrt{403}-20} = \tfrac{\sqrt{403}+20}{1} = 40 + \tfrac{\sqrt{403}-20}{1} \\ y_4 &= \tfrac{1}{\sqrt{403}-20} = \tfrac{\sqrt{403}+20}{3} = 13 + \tfrac{\sqrt{403}-19}{3} \\ y_5 &= \tfrac{3}{\sqrt{403}-19} = \tfrac{\sqrt{403}+19}{14} = 2 + \tfrac{\sqrt{403}-9}{14} \end{aligned}$$

Then, just as y_2 *gives* $x_1 = \frac{\sqrt{403}+20}{3}$*,* y_4 *gives* $x_{-1} = \frac{\sqrt{403}+19}{3}$ *and* y_5 *gives* $x_{-2} = \frac{\sqrt{403}+9}{14}$*.*

Combining periodicity with reversibility strengthens Theorem 1 (h).

Lemma 3 *There exists a positive integer* π *such that* $\forall i\ x_i = x_{i+\pi}$*,* i *not necessarily positive.*

Proof: From the proof of Theorem 1 (h) there are k and π such that $\forall i \geq k$, $x_i = x_{i+\pi}$. Essentially,this is equivalent to proving that there is no lower bound for k. Assume the contrary, that there is some lower bound k. Let k and π be the smallest such integers. Then $x_k = x_{k+\pi}$. But by Lemma 2 $x_{k-1} = x_{k+\pi-1}$, so that $k-1$ also meets this criteria, violating the assumption that k is the smallest such integer. Therefore, $\forall i\ x_i = x_{i+\pi}$. **QED**

Throughout, π will consistently denote the period, even when considering this period in the context of quadratic forms or lattices.

Often the continued fraction may have other characteristics that are interesting besides its periodicity. For factorization, continued fractions with symmetries, such as at x_0 and x_5 from Example 2, will be especially important. If the starting condition near some point is the same in both directions, the entire sequence will be symmetric about that point. This is the point of Lemma 4.

Lemma 4 *Let* $x_0 = \frac{\sqrt{N}+P_{-1}}{Q_0}$ *meet (4) such that* $Q_0 \mid 2P_{-1}$*. The sequence of pseudo-squares is symmetric about* Q_0*, so that* $\forall i\ Q_i = Q_{-i}$*.*

Proof:
Observe that $0 < \sqrt{N} - P_0 < Q_0$, with $P_0 = b_0 Q_0 - P_{-1}$, so that

$$0 < \sqrt{N} - b_0 Q_0 + P_{-1} < Q_0.$$

There can only be one possible integer value of b_0 that satisfies this inequality. Since $0 < \sqrt{N} - P_{-1} < Q_0$, $b_0 = 2P_{-1}/Q_0$ satisfies this inequality, so that $P_0 = P_{-1}$.

Let $y_{-1} = \frac{\sqrt{N}+P_1}{Q_1}$. Then, by Lemma 2, $y_0 = \frac{\sqrt{N}+P_0}{Q_0} = \frac{\sqrt{N}+P_{-1}}{Q_0} = x_0$

Therefore, the sequence of pseudo-squares will be symmetric about Q_0, since in either direction the first continued fraction term is the same. Therefore, $Q_i = Q_{-i}$. **QED**

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[4]:

[4]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 [6] and first applied by Shanks [18].

Theorem 3 *Let $s = \lfloor \frac{\pi}{2} \rfloor$, where π is the period from Lemma 3. If π is even, $\forall i\ Q_{s+i} = Q_{s-i}$, but $Q_s \neq Q_0$ and $Q_s \mid 2N$. If π is odd, $\forall i\ Q_{s+i+1} = Q_{s-i}$ and either $\gcd(Q_s, N)$ is a nontrivial factor of N or -1 is a quadratic residue of N.*

Proof:

Case 1: If π is even, $\pi = 2s$. Then, by Lemmas 4 and 3, $Q_{s+i} = Q_{-s-i} = Q_{2s-s-i} = Q_{s-i}$. Since $Q_{s+1} = \frac{N-P_s^2}{Q_s}$ and $Q_{s-1} = \frac{N-P_{s-1}^2}{Q_s}$, this simplifies to $P_s^2 = P_{s-1}^2$, but since $\forall i\ P_i > 0$, this provides $P_s = P_{s-1}$.

Now $Q_s = \frac{P_s+P_{s-1}}{b_s} = \frac{2P_s}{b_s}$, so that $Q_s \mid 2P_s$.

Assume $Q_s = Q_0$. If Q_s is even, then $P_0 \equiv P_s \equiv 1 \pmod 2$ and if Q_s is odd, $Q_s \mid P_s$. Either way, there is then a unique integer in the range $(\sqrt{N} - Q_0, \sqrt{N})$ satisfying these conditions, so that $P_s = P_0$. Therefore, $x_s = \frac{\sqrt{N}+P_0}{Q_0} = x_0$, contradicting the fact that π is the smallest positive integer such that $\forall i\ Q_i = Q_{i+\pi}$. Therefore, $Q_s \neq Q_0$.

Now $N = P_s^2 + Q_s Q_{s+1}$, so it is apparent that if Q_s is odd, then $Q_s \mid P_s$, so that $Q_s \mid N$. Conversely, if Q_s is even, then $(Q_s/2) \mid P_s$, so that $(Q_s/2) \mid N$. Either way, $Q_s \mid 2N$.

Case 2: If π is odd, $\pi = 2s+1$. Then, by Lemma 4 and 3, $Q_{s+i+1} = Q_{-s-i-1} = Q_{2s+1-s-i-1} = Q_{s-i}$.

Specifically, $Q_s = Q_{s+1}$, so that $N = P_s^2 + Q_s Q_{s+1} = P_s^2 + Q_s^2$, so that $P_{s+1}^2 \equiv -Q_s^2 \pmod N$. If $\gcd(Q_s, N) > 1$, this is a nontrivial factor of N, and the proof is done. Therefore, assume that Q_s and N are relatively prime, so that $Q_s^{-1} \pmod N$ exists. Then $(Q_s^{-1})^2 P_{s+1}^2 \equiv -1 \pmod N$. Then $Q_s^{-1} P_{s+1}$ is a square root of -1 modulo N. **QED**

One final concept that will appear much more important in later sections is equivalence. Define the set $\mathbb{T}$ to be set of all numbers of the form $\frac{\sqrt{N}+P}{Q}$ such that:

$$P^2 \equiv N \pmod Q, \tag{7}$$

Then define

$$\mathbb{T}^* = \{x \in \mathbb{T} : 0 < P < \sqrt{N},\ |\sqrt{N} - Q| < P\}.$$

An element $x \in \mathbb{T}$ is *reduced* if $x \in \mathbb{T}^*$. For $x, y \in \mathbb{T}^*$, x is *equivalent* to y if x appears in the same continued fraction expansion as y and it is trivial that this is an equivalence relation on $\mathbb{T}^*$. Extending this to all of $\mathbb{T}$ requires a lemma relating elements of $\mathbb{T} - \mathbb{T}^*$ with elements of $\mathbb{T}^*$.

Lemma 5 *Let $x = x_0 \in \mathbb{T} - \mathbb{T}^*$. x_0 may be reduced by applying*

$$x_{i+1} = \frac{1}{x_i - b_i},\ b_i = \lfloor x_i - 1/2 \rfloor,\ i \geq 0 \tag{8}$$

until $|Q_i| < 2\sqrt{N}$ for some i and then applying equation (2) normally until $x_k \in \mathbb{T}^*$ for some $k > 0$.

Proof: The choice of b_i yields that $\left|\frac{\sqrt{N}-P_i}{Q_i}\right| < \frac{1}{2}$ after the first step, so that $|P_i| < \frac{1}{2}|Q_i| + \sqrt{N}$. Therefore,

$$\begin{aligned} |Q_{i+1}| &= \left|\tfrac{N-P_i^2}{Q_i}\right| \\ &= \left|\tfrac{\sqrt{N}-P_i}{Q_i}\right| |\sqrt{N} + P_i| \\ &< \tfrac{1}{2}(2\sqrt{N} + \tfrac{1}{2}|Q_i| \\ &= \sqrt{N} + \tfrac{1}{4}|Q_i| \end{aligned}$$

so that $|Q_i|$ will decrease as long as $|Q_i| > \frac{4}{3}\sqrt{N}$. When $|Q_r| < 2\sqrt{N}$, revert back to the standard formula for b_r. There are three cases for what Q_r is:

Case 1: $0 < Q_r < \sqrt{N}$. In this case, it is clear that x_{r+1} will be reduced.

Case 2: $\sqrt{N} < Q_r < 2\sqrt{N}$. In this case, if $P_r > 0$, then x_{r+1} will be reduced. Otherwise, $|P_r| < \sqrt{N}$, so that x_{r+1} will be in Case 1.

Case 3: $-2\sqrt{N} < Q_r < 0$. Then $\sqrt{N} < P_r < \sqrt{N} + |Q_r|$, yielding $Q_{r+1} < 2\sqrt{N} + |Q_r|$. If $Q_{r+1} < 2\sqrt{N}$, it is in Case 1 or Case 2. If $2\sqrt{N} < Q_{r+1} < 2\sqrt{N} + |Q_r|$, the choice of b_r provides $\sqrt{N} + P_r > Q_{r+1}$, so that $0 < P_{r+1} < \sqrt{N}$, so that x_{r+2} will be in Case 1. **QED**

I will provide one example of reduction:

Example 4

$$x_0 = \frac{\sqrt{403}+267}{-134} = -2 + \frac{\sqrt{403}-1}{-134}$$
$$x_1 = \frac{-134}{\sqrt{403}-1} = \frac{\sqrt{403}+1}{-3} = -8 + \frac{\sqrt{403}-23}{-3}$$
$$x_2 = \frac{-3}{\sqrt{403}-23} = \frac{\sqrt{403}+23}{42} = 1 + \frac{\sqrt{403}-19}{42}$$
$$x_3 = \frac{42}{\sqrt{403}-19} = \frac{\sqrt{403}+19}{1} = 39 + \frac{\sqrt{403}-20}{1}$$

Lemma 5 defines a map from $\mathbb{T}$ to $\mathbb{T}^*$ (Elements of $\mathbb{T}^*$ are mapped to themselves). Then two elements are equivalent if their corresponding elements of $\mathbb{T}^*$ are equivalent and it is clear that this is still an equivalence relation. Essentially, this equates to saying that two numbers x and y are equivalent if their continued fraction expansions have the same "tail", so that after a certain number of terms of each they have identical cycles. §3 will define a different equivalence relation on binary quadratic forms and then prove that it corresponds to this.

2 From Morrison-Brillhart to Shanks

Morrison and Brillhart developed one simple and fairly intuitive algorithm for using continued fractions for factorization [11]. The entire algorithm is a bit more complicated, but here is a description sufficient for our purposes.

If the equation

$$x^2 \equiv y^2 \pmod{N} \tag{9}$$

can be solved such that

$$x \not\equiv \pm y \pmod{N}, \tag{10}$$

then it is evident that $\gcd(x - y, N)$ provides a nontrivial factor of N. Choosing a value for x and then looking for a value that works for y is not computationally effective for large N. Fermat, the first to employ this concept, tested values of x greater than $\sqrt{N}$ to find a value such that $x^2 \pmod{N}$ was already a perfect square. However, these numbers get large quickly. Continued fractions provide a better approach to achieving this and since $0 < Q_i < 2\sqrt{N}$, the chances of finding a perfect square are greatly improved. The second part of Theorem 2 states that $A_{i-1}^2 - B_{i-1}^2 N = (-1)^i Q_i$. Therefore $A_{i-1}^2 \equiv (-1)^i Q_i \pmod{N}$. If for some i, Q_{2i} is a perfect square then this provides a solution to (9) and it only remains to check whether or not $\gcd(x+y, N)$ or $\gcd(x-y, N)$ provide a nontrivial factor of N. However, there are not very many perfect squares in the continued fraction expansion so Morrison and Brillhart [11] used products to obtain squares. For example, for $N = 1333$, the continued fraction expansion provides

$73^2 \equiv -3 \pmod{N}$ and $1789^2 \equiv -12 \pmod{N}$. From this, $(73 \cdot 1789)^2 \equiv (-3)(-12) = 6^2 \pmod{N}$, quickly yielding $1333 = 31 \cdot 43$.

There are a couple of problems with this algorithm. First, it requires the calculation of the A_i's, which are of the same size as N, after reduction modulo N, while the rest of the algorithm only requires arithmetic on numbers of size $\sqrt{N}$. Second, after going through a nontrivial amount of computation to find a relation that solves (9), not all of these result in a factorization. I provide one example:

Example 5 *In the continued fraction for* $\sqrt{1333}$*,* $Q_6 = 9 = 3^2$ *and* $A_5 = 10661$*, so that* $10661^2 \equiv 3^2 \pmod{1333}$*. Unfortunately* $10661 \equiv -3 \pmod{1333}$*, so this does not result in a nontrivial factor of* 1333*.*

Although it is well enough on a computer to run the algorithm a few times and have it fail a few times, Daniel Shanks decided he needed to understand it a little better. Based on an understanding of quadratic forms and the class group infrastructure, Daniel Shanks developed several very interesting algorithms for factorization ([18],[15]). First he developed an improvement to the Morrison-Brillhart algorithm. Roughly speaking, rather then saving the A_i's, he was able to use composition of quadratic forms to combine numbers to produce squares and then use the "infrastructure" to use those squares to find a factorization. In addition, he developed from the concept of infrastructure a system of predicting whether or not any given square would provide a nontrivial factor. Unfortunately, this didn't save very much time and was a much more complicated algorithm than the Morrison-Brillhart algorithm.

From here the development of the algorithm was prompted by the number $2^{60} + 2^{30} - 1$. It failed a Fermat primality test[5], but when Morrison and Brillhart tried to factor it, it failed 114 times. Therefore, they stopped, multiplied it by some small constant and tried again. This time it worked on the first try, but they wanted to know why it had failed so many times. So they asked Shanks to analyze it. Unfortunately (or fortunately in hindsight), Shanks only had an HP-65 available and he couldn't fit his entire algorithm into it. Therefore, he discarded all the work of combining numbers to form squares and just cycled through until he found one already there. The code for this was much shorter, and as it turned out the algorithm was actually significantly faster.

3 Quadratic Forms

A fuller account of binary quadratic forms can be found in Gauss's [6] and in Buell's [1]. However, here are the necessary fundamental ideas.

A binary quadratic form is a polynomial of the form $F(x, y) = ax^2 + bxy + cy^2, x, y, a, b, c \in \mathbb{Z}$ (Often this is abbreviated as (a, b, c)). In some sense then, a quadratic form may be considered to be the set of all the numbers it can represent for various values of x and y. Thus, two quadratic forms are *equivalent* if they represent the same set of integers. It is evident that if one form is transformed into another by the substitution

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix}, \; ad - bc = \pm 1, \tag{11}$$

[5]The Fermat primality test is based on Fermat's classical result that for p prime, $a^p \equiv a \pmod{p}$ [7]. Equivalently, if $a^N \not\equiv a \pmod{N}$ for some integer a, then N isn't prime.

then, since this matrix is invertible, the two forms are equivalent. As one further useful distinction, (11) is *proper* if its determinant is $+1$ and *improper* if its determinant is -1. The symbol ($\sim$) will only apply to proper equivalence. For the purpose of factorization, the interesting forms are those that can be improperly transformed into themselves, referred to as *ambiguous forms.*

Example 6 *$F(x, y) = -14x^2 + 10xy + 5y^2$ is transformed into itself by the substitution:*

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix},$$

so $(-14, 10, 5)$ is ambiguous.

Denote the set of all quadratic forms with discriminant Δ by $\mathbb{F}_\Delta$, or often just $\mathbb{F}$. The next obvious question is the organization of all of the quadratic forms equivalent to some given form. Since there are an infinite number of forms equivalent to any form, the search must be narrowed some by first defining reduced forms.

Definition 1 A quadratic form $ax^2 + bxy + cy^2$, with positive discriminant $\Delta = b^2 - 4ac$ is *reduced* if*:*

$$0 < b < \sqrt{\Delta} \tag{12}$$

$$|\sqrt{\Delta} - 2|a|| < b \tag{13}$$

Note that $\Delta = b^2 - 4ac$ and (12) require that $ac < 0$, so that a and c must have opposite signs.

Making Gauss's description of the organization make some sense will require one more of his definitions:

Definition 2 Two forms $F(x, y) = ax^2+bxy+cy^2$ and $F'(x, y) = a'x^2+b'xy+c'y^2$ are *adjacent* if $c = a'$.

To each quadratic form, there is a unique reduced equivalent form adjacent to it on each side[6], and since (12-13) imply a finite number of possible coefficients, this process eventually repeats, forming a cycle. The important aspect of this is that the cycle is actually all of the reduced forms equivalent to the first form:

Theorem 4 *[6] If the reduced forms F,F′ are properly equivalent, each of them will be contained in the period of the other.*

Gauss proves this in Article 193 of [6], Lenstra proves this in [9], and it is a corollary of Lemma 13 in §5. Therefore the proof is omitted.

For now, the important detail is that the quadratic forms correspond directly to the elements of $\mathbb{T}$, and that the reduced quadratic forms correspond to elements of $\mathbb{T}^*$. Note that the elements of $\mathbb{T}$ have attached indices, where the important trait of the indice is whether it is odd or even. Define a map from $\mathbb{T}$ to $\mathbb{F}$ by

$$\begin{gathered} \Phi_{\mathbb{T},\mathbb{F}} : \mathbb{T} \to \mathbb{F} \\ \tfrac{\sqrt{N}-P_i}{Q_i} \to F_i(x, y) = Q_i(-1)^i x^2 + 2P_i xy + Q_{i+1}(-1)^{i+1} y^2 \end{gathered} \tag{14}$$

[6]Although Gauss had a recursive mechanism for finding these, continued fractions provide a sufficient mechanism for this that will be defined momentarily. Note that reversal suddenly becomes trivial.

The inverse map is

$$\begin{gathered}\Phi_{\mathbb{F},\mathbb{T}} : \mathbb{F} \to \mathbb{T} \\ ax^2 + bxy + cy^2 \to x_i = \frac{\sqrt{\Delta/4} - b/2}{|a|} \in \mathbb{T}\end{gathered} \tag{14'}$$

where the discriminant of the quadratic form is Δ is the element of $\mathbb{T}$ is given either an even or odd indice as a is positive or negative, respectively. Note that $\Delta = b^2 - 4ac$ gives $4a \mid \Delta - b^2$, so that x_i really is in $\mathbb{T}$.

§1 defined an equivalence on $\mathbb{T}$ and suggested that it corresponded with an equivalence of binary quadratic forms. Theorem 5 formalizes this:

Theorem 5 *Under the mapping $\Phi_{\mathbb{T},\mathbb{F}}$, the equivalence classes of $\mathbb{T}$ correspond to the equivalence classes of $\mathbb{F}$. That is, for $x_i, x_j \in \mathbb{T}$ corresponding to $F_i = \Phi_{\mathbb{T},\mathbb{F}}(x_i), F_j = \Phi_{\mathbb{T},\mathbb{F}}(x_j) \in \mathbb{F}$, respectively, $x_i \sim x_j$ if and only if $F_i \sim F_j$.*

Proof:

Let $x_i \sim x_j$. Since x_i and x_j must be in the same continued fraction expansion, assume without loss of generality that $j = i + 1$. The other cases may be easily derived from this case. Then the quadratic form related to x_i is given in (14). The substitution

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & (-1)^i b_{i+1} \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix}$$

transforms F_i into

$$Q_{i+1}(-1)^{i+1}x^2 + 2P_{i+1}xy + Q_{i+2}(-1)^{i+2}y^2$$

Observe that this matrix has determinant 1, so that this equivalence is proper.

In order to prove the converse, that the x_i's related to equivalent quadratic forms are equivalent, by Theorem 4, observe that the last coefficient of the quadratic form related to x_i, $Q_{i+1}(-1)^{i+1}$, is then the first coefficient of the quadratic form related to x_{i+1}. Therefore, these two forms are adjacent and thus equivalent. **QED**

The real value of quadratic forms is the composition of quadratic forms. In Article 236 of [6], Gauss provides a very flexible definition of composition. Gauss defines composition as multiplying two quadratic forms together and then making a substitution to simplify this into another binary quadratic form. The algorithm he provides is very complicated, allowing for choices of variables along the way that permit the result to be any quadratic form in the resulting equivalence class. The result of composition should be predictable, so definition needs to be limited some. Shanks and Buell both provide a significant simplifaction of this algorithm. The symbol $(*)$ will consistently be used for composition.

Proposition 1 *[1] Let $F_1 = (a_1, b_1, c_1)$ and $F_2 = (a_2, b_2, c_2)$ be primitive forms of discriminants d_1 and d_2, respectively, such that $d_1 = \Delta n_1^2$ and $d_2 = \Delta n_2^2$ for integers n_1 and n_2 and Δ, with $\Delta = \gcd(d_1, d_2)$. Let*

$$m = \gcd(a_1 n_2, a_2 n_1, \frac{b_1 n_2 + b_2 n_1}{2}).$$

Then the congruences

$$mn_1B \equiv mb_1 \pmod{2a_1}$$
$$mn_2B \equiv mb_2 \pmod{2a_2}$$
$$m(b_1n_2 + b_2n_1)B \equiv m(b_1b_2 + \Delta n_1n_2) \pmod{4a_1a_2}$$

are simultaneously solvable[7] *for an integer* B*, and the composition of* F_1 *and* F_2 *is:*

$$F_1 * F_2 = (\frac{a_1a_2}{m^2}, B, \frac{(B^2 - \Delta)m^2}{4a_1a_2})$$

of discriminant Δ*.*

See [17] for a derivation of this in the case where the discriminants are equal or [1] for a proof of this case. Buell [1] also provides the substitutions that would be needed for Gauss's definition of composition.

The next question is how this operation is related to equivalence.

Theorem 6 *If* $F_1 \sim F_2$*, then* $F * F_1 \sim F * F_2$*.*

Gauss proves this in Article 237-239 of [6].

Therefore, composition treats the equivalence classes in the convenient manner. These equivalence classes are then the elements of the class group, with composition as the group operation. The application of Theorem 6 is that it doesn't matter which form is used to represent an equivalence class.

The significance of ambiguous forms for factorization has been mentioned some above. It is evident that if one form is ambiguous, then its entire equivalence class is also ambiguous. Lemmas 6 generalizes the reasons to be interested in these classes.

Lemma 6 *An ambiguous equivalence class contains two points of symmetry, that is, pairs of reduced adjacent forms,* (c, b, a) *and* (a, b, c) *in the cycle that are the symmetric reverse of each other. Let* a *be the connecting term of either symmetry point. Either* a *divides the determinant, or* $a/2$ *divides the determinant.*

Proof:

Let A be an ambiguous equivalence class and let $F = ax^2 + bxy + cy^2 \in A$. Let $F' = cx^2 + bxy + ay^2$. Then since $F \in A$, there is a substitution of determinant -1 that maps F into itself. Since the obvious substitution to exchange x and y in F has determinant -1, the product of these two is a proper substitution that transforms F into F'. Therefore, $F' \in A$, so that if F is F_0, F' is F_j for some j. Then that F_1 must be the reverse of F_{j-1}, and so forth. Now, if j is even, then by this process $F_{j/2}$ is its own reverse. However, by the definition of being reduced, the end coefficients of each form must have opposite sign, so this is impossible. Therefore, j must be odd, and then $F_{\frac{j-1}{2}}$ is the reverse of $F_{\frac{j+1}{2}}$.

At this point, observe that since the end-coefficients alternate signs, the entire period must be even. By the same arguments as Theorem 3, one could show that there must be another point of symmetry with the property that $\forall i\ Q_{s+i} = Q_{s-i}$, but such that Q_s is not the same as the connecting term at the first symmetry point. The two quadratic forms containing Q_s as an end coefficient then meet the criteria.

The fact that either a divides the determinant, or $a/2$ divides the determinant was proven in Theorem 3, since the determinant is $4N$. **QED**.

[7] By convention, choose the answer with the smallest absolute value.

Note that Theorem 3 described two different types of points of symmetry. With the quadratic form cycle, the second case can be ignored because of the alternating signs. However, it is quite possible for the term at one symmetry point to be merely the negative of the term at the other symmetry point. This would correspond to the continued fraction having an odd period and there would be a symmetry point of the second type in the continued fraction at half-way. However, this type of symmetry does not generally provide a factorization for N.

Lastly, it is important how these ambiguous forms fit into the rest of the class group. First, addressing the class group structure requires inverses. Lemma 7 is fairly elementary and is probably stated somewhere else. Let 1 represent the form in the principal cycle whose first coefficient is 1. Let F^{-1} indicate the symmetric reverse of F, $(a,b,c)^{-1} = (c,b,a)$. Lemma 7 justifies this notation:

Lemma 7 $F * F^{-1} \sim 1$

Proof:

Let $F = ax^2 + bxy + cy^2$. Then $F^{-1} = cx^2 + bxy + ay^2$. Let G be the next form adjacent to F^{-1}, that is $G = ax^2 + b'xy + c'y^2$, with $a \mid (b + b')$ from the correspondance with continued fractions. Composing $F * G$, $n_1 = n_2 = 1$ and $m = a$, so that the first coefficient of $F * G$ is 1. Therefore, $F * G \sim 1$, but $F^{-1} \sim G$, so $F * F^{-1} \sim 1$. **QED**.

Note that this implies that the square of a symmetry point is 1.

Theorem 7 was probably known by Shanks, since SQUFOF depends highly on it, but it does not seem that he states this explicitly anywhere.

Theorem 7 *An equivalence class has order* 2 *or* 1 *in the class group if and only if it is ambiguous.*

Proof:

Let A be an ambiguous class. Let $F \in A$. Then $F \sim F^{-1}$, so that $F * F \sim F * F^{-1} \sim 1$. Therefore $F * F$ is in the principal cycle, so that A has order 2 or 1 in the class group.

Conversely, assume that an equivalence class A has order 2 or 1 in the class group. Let $F \in A$. Then $F * F$ is in the principal ideal, so that $F * F \sim (F * F)^{-1}$. But from composition, it is clear that $(F*F)^{-1} \sim F^{-1}*F^{-1}$. So $F*F \sim F^{-1}*F^{-1}$. Since the class group is associative, composing on the right with F maintains equivalence. Therefore:

$$\begin{gathered}(F * F) * F \sim (F^{-1} * F^{-1}) * F \\ 1 * F \sim F^{-1} * (F^{-1} * F) \\ F \sim F^{-1}\end{gathered}$$

Therefore, A is ambiguous. **QED**.

Certainly the class group structure is interesting, but it is now possible to return to the problem from the Morrison-Brillhart algorithm of Example 5 with $Q_3 = 3$, so $Q_6 = 9$ doesn't provide a nontrivial factor of N. The quick explanation is that if you square the quadratic form with first coefficient Q_3, you obtain the quadratic form with first coefficient Q_6. Since the principal cycle is closed under composition, it seems as though, and perhaps would be convenient if, the forms in the principal cycle formed a group. However, the problem of reduction prevents this:

Example 7 *Consider the quadratic form* $F = (36, 70, -3)$*, with determinant* $4 \cdot 1333$*. Compare* $(F * F) * F$*, with* $F * F * F$*, where the difference is that in the first the result is reduced after*

the first composition. $F * F = (324, -38, -3)$ *and the very next adjacent form* $(-3, 68, 59)$*, is reduced.* $F * (-3, 68, 59) = (-12, 70, 9)$*, which is already reduced. However, without reduction* $F * F * F = F * (324, -38, -3) = (729, 448, -348)$*. When this is reduced, the first reduced form found, after* 2 *steps, is* $(9, 56, -61)$*.*

Therefore, the principal cycle, with the operation being composition followed by reduction, doesn't even meet the requirements for being power associative. However, the observation that the two results are adjacent forms, and that the second reduction took one step longer, prompts us to dig a little deeper.

Understanding this requires what Shanks referred to as infrastructure distance. For $m < n$, and for $x_i \in \mathbb{T}$, the terms in the continued fraction in (5), define

$$D_{\mathbb{F}}(x_m, x_n) = \log(\prod_{k=m+1}^{n} x_k) \tag{15}$$

Lenstra [9] adds a term of $\frac{1}{2}\log(Q_n/Q_m)$ to this, with the effect that the resulting formulas are slightly simplified but the proofs are more complicated and less intuitive. This definition is used by Williams in [20].

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 3 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,

$$R = D_{\mathbb{F}}(F_0, F_\pi) = D(1, F_\pi)$$

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

Further analysis of this distance will require two more tools: ideals and lattices. In order to relate to continued fractions, the ideals will be in $\mathbb{Z}[\sqrt{N}] = \{a + b\sqrt{N} : \ a, b \in \mathbb{Z}\}$ and the lattices will be in $\mathbb{Q}(\sqrt{N}) = \{a + b\sqrt{N} : \ a, b \in \mathbb{Q}\}$where N is a non-square positive integer.

Remark: The ideals in $\mathbb{Z}[\sqrt{N}]$ typically correspond only to quadratic forms of discriminant $4N$. Note that if $N \equiv 1 \pmod 4$, then $\mathbb{Z}[\sqrt{N}]$ is not the ring of integers for $\mathbb{Q}(\sqrt{N})$. For $N \equiv 1 \pmod 4$, an analysis of ideals in $\mathbb{Z}[\frac{\sqrt{N}+1}{2}]$ is also interesting, but will be avoided in the interest of simplicity. Quadratic forms of discriminant $N \equiv 1 \pmod 4$ may be related to ideals in $\mathbb{Z}[\sqrt{N}]$ via first multiplying by 2 to obtain quadratic forms of discriminant $4N$.

4 Ideals

For $\xi \in \mathbb{Q}(\sqrt{N})$, let $\overline{\xi}$ refer to the *conjugate* of ξ (i.e. $\overline{1 + \sqrt{3}} = 1 - \sqrt{3}$).

The *norm* of a number in $\mathbb{Q}(\sqrt{N})$ is $\mathcal{N}(\xi) = \xi\overline{\xi} \in \mathbb{Q}$.

To simplify notation, the symbols H, I, J, and K will consistently be ideals, u and v will be elements of ideals, α and β will be elements of $\mathbb{Z}[\sqrt{N}]$, ξ and ζ will be elements of $\mathbb{Q}(\sqrt{N})$, and $\mathcal{L}$ will be a lattice.

Our definition of an ideal is the same as in any other commutative ring with identity:

Definition 4 A subset I of a ring R is an *ideal* if for $u, v \in I$, $u \pm v \in I$ and for $\alpha \in R$, $u \cdot \alpha \in I$, that is I is closed under addition and multiplication by an element of R. Define $L(I)$ to be the least positive rational integer in I.

Describing ideals will require the notation for the lattice generated by a set. If

$$\alpha_1, \alpha_2, ..., \alpha_k \in \mathbb{Z}[\sqrt{N}],$$

denote[8] the lattice generated by these as

$$[\alpha_1, \alpha_2, ..., \alpha_k] = \{\sum_{i=1}^{k} n_i \alpha_i : n_i \in \mathbb{Z}\} \tag{16}$$

Lemma 8 identifies necessary and sufficient conditions for a set in $\mathbb{Z}[\sqrt{N}]$ to be an ideal.

Lemma 8 *For $Q, s, N, P \in \mathbb{Z}$, N non-square and positive, $[Q, s\sqrt{N}+P]$ is an ideal of the ring $\mathbb{Z}[\sqrt{N}]$ if and only if $sQ \mid \mathcal{N}(s\sqrt{N}+P)$, $s \mid Q$, and $s \mid P$.*

Proof: Assume that $I = [Q, s\sqrt{N}+P]$ is an ideal of $\mathbb{Z}[\sqrt{N}]$. Then, choosing $\alpha = P - s\sqrt{N}$, $\mathcal{N}(s\sqrt{N}+P) \in I$. Since this is an integer, $Q \mid \mathcal{N}(s\sqrt{N}+P)$. Choosing $\alpha = \sqrt{N}$, $Q\sqrt{N} \in I$. Therefore, $s \mid Q$. Since $Q \mid \mathcal{N}(s\sqrt{N}+P)$, this also implies that $s \mid P$. Therefore, α could have been chosen $\alpha = P/s - \sqrt{N}$ so that $Q \mid \mathcal{N}(s\sqrt{N}+P)/s$, so that $sQ \mid \mathcal{N}(s\sqrt{N}+P)$.

Conversely, let $I = [Q, s\sqrt{N}+P]$ and assume that $sQ \mid \mathcal{N}(s\sqrt{N}+P)$, $s \mid Q$, and $s \mid P$. Closure under addition is trivial. To see that I is closed under multiplication by an element of $\mathbb{Z}[\sqrt{N}]$, one need only consider multiplication by 1 and $\sqrt{N}$, since they form a basis for $\mathbb{Z}[\sqrt{N}]$. Multiplication by 1 is trivial. For $\sqrt{N}$,

$$Q\sqrt{N} = \frac{Q}{s}(s\sqrt{N}+P) - \frac{P}{s}Q$$

and Q/s and $-P/s$ are integers. Also,

$$(s\sqrt{N}+P)\sqrt{N} = sN + P\sqrt{N} = \frac{P}{s}(s\sqrt{N}+P) + (\frac{-P^2+s^2N}{sQ})Q$$

and $\frac{P}{s}$ and $(\frac{-P^2+s^2N}{sQ})$ are integers. **QED**

If $s = 1$, an ideal is *primitive*. Since $s \mid P$ and $s \mid Q$, ideals that are not primitive will often be written $(s)[Q, \sqrt{N}+P]$. Let $\mathbb{I}$ be the set of all primitive ideals.

Represented in the form $I = [Q, \sqrt{N}+P]$, it is clear that $|Q|$ is the smallest positive rational integer in I. Define

$$L(I) = \min\{I \cap \mathbb{Z}^+\} \tag{17}$$

At this point, it is possible to define a correspondance between quadratic forms (of discriminant $\Delta \equiv 0 \pmod 4$) and ideals by:

$$\Phi_{\mathbb{F},\mathbb{I}}(F(x,y) = Ax^2 + Bxy + Cy^2) = [A, \sqrt{(\frac{B}{2})^2 - AC} + \frac{B}{2}] \tag{18}$$

$$\Phi_{\mathbb{I},\mathbb{F}}([Q, \sqrt{N}+P]) = F(x,y) = Qx^2 + 2Pxy + (\frac{P^2-N}{Q})y^2 \tag{18'}$$

and define a *reduced* ideal as an ideal corresponding to a reduced quadratic form. Note that $\Delta = 4N$.

[8]Observe the difference between the use of [...] here and in §1. This expression is completely unrelated to rational approximations.

For example, the quadratic form $(15, 2 \cdot 12, -1)$ corresponds to the ideal $[15, \sqrt{159} + 12]$. The one potential problem that immediately becomes apparent is that while $[15, \sqrt{159}+12]$ and $[-15, \sqrt{159} + 12]$ are the same ideal, $(15, 2 \cdot 12, -1)$ and $(-15, 2 \cdot 12, 1)$ are different quadratic forms. However, it is apparent that the negative sign is merely carried through composition without affecting the computations. Since each of these forms is in the same location within its respective cycle, this difference will not be important to this investigation of composition and distance.

$\Phi_{\mathbb{T},\mathbb{F}}$ and $\Phi_{\mathbb{F},\mathbb{I}}$ may be combined to obtain

$$\Phi_{\mathbb{T},\mathbb{I}}(\frac{Q}{\sqrt{N} - P}) = [Q, \sqrt{N} - P]$$

and $\Phi_{\mathbb{I},\mathbb{T}}$ is defined in the related obvious way.

If $A = [\alpha_i]$ and $B = [\beta_i]$, $i = 1, 2, 3..., d$, then it is clear that $A = B$ if and only if there exists a $d \times d$ matrix M with determinant ± 1 such that:

$$\langle \alpha_i \rangle = M \langle \beta_i \rangle$$

where $\langle \alpha_i \rangle$ and $\langle \beta_i \rangle$ are vectors.

For these purposes, the most important operation with ideals is their multiplication. Multiplication is defined by

$$[\alpha_i] * [\beta_j] = [\alpha_i \beta_j]$$

For example,

$$I = [15, \sqrt{159} + 12] * [10, \sqrt{159} + 13] = [150, 10\sqrt{159} + 120, 15\sqrt{159} + 195, 315 + 25\sqrt{159}]$$

The 4th component is the sum of the 2nd and 3rd, so it is unnecessary for describing the ideal.

Applying the matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

which has determinant 1, subtracts the 2nd component from the third to obtain

$$I = [150, 10\sqrt{159} + 120, 5\sqrt{159} + 75]$$

The matrix, with determinant 1,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix}$$

will subtract twice the 3rd component from the 2nd to obtain

$$I = [150, -30, 5\sqrt{159} + 75]$$

Here the 1st component is a multiple of the 2nd and is thus unnecessary. The answer is simplified to obtain

$$I = [30, 5\sqrt{159} + 75] = 5[6, \sqrt{159} + 15]$$

The process of multiplying ideals can be greatly simplified by several well-known formulae[9] [10].

Theorem 8 *Let* $I = [Q, \sqrt{N} + P]$ *and* $J = [Q', \sqrt{N} + P']$ *be ideals of* $Q\sqrt{N}$. *Let* $C = \frac{N-P^2}{Q}$, $C' = \frac{N-(P')^2}{Q'}$. *If* $\gcd(Q, P, C) = \gcd(Q', P', C') = 1$, *then* $I * J = s[q, \sqrt{N} + p]$, *where*

$$s = \gcd(Q, Q', P + P') \tag{19}$$
$$h = \gcd(Q, Q', C, C', 2) \tag{20}$$
$$q = hQQ'/s^2 \tag{21}$$
$$p \equiv P \pmod{Q/s} \tag{22}$$
$$p \equiv P' \pmod{Q'/s} \tag{23}$$
$$(P + P')p \equiv N + PP' \pmod{QQ'/s} \tag{24}$$

Proof[10]**:**
Consider the product:

$$I * J = [QQ', Q\sqrt{N} + QP', Q'\sqrt{N} + Q'P, N + PP' + (P + P')\sqrt{N}] \tag{25}$$

The smallest integer in $I * J$ may be found by considering the smallest integers that may be produced taking these elements pair-wise. Let $\{..\}$ represent the least common multiple,

$$L(I * J) = \gcd(QQ', \{Q, Q'\}(P - P'), \frac{\{Q, P + P'\}Q'C'}{P + P'}, \frac{\{Q', P + P'\}QC}{P + P'})$$

Let $s = \gcd(Q, Q', P + P')$, $h = \gcd(Q, Q', C, C', 2)$, $w = hQQ'/s$. Let $f \neq 2$ be a prime. Let a, b, c, d, e, k be the largest possible integers such that $f^a \mid Q$, $f^b \mid Q'$, $f^c \mid (P+P')$, $f^d \mid C$, $f^e \mid C'$, $f^k \mid (P - P')$. Then $f^{a+b} \parallel QQ'$, $f^{\max(a,b)+k} \parallel \{Q, Q'\}(P - P')$, $f^{\max(a,c)+b+e-c} \parallel \frac{\{Q,P+P'\}Q'C'}{P+P'}$, and $f^{\max(b,c)+a+d-c} \parallel \frac{\{Q',P+P'\}QC}{P+P'}$).

The following analysis proves that if $f \neq 2$, the maximum exponent of f in $L(I*J)$ is $a+b-\min(a, b, c)$ while if $h = 2$, then the maximum exponent of 2 in $L(I*J)$ is $a+b+1-\min(a, b, c)$, while if $h = 1$, then the maximum exponent of f in $L(I * J)$ is $a + b - \min(a, b, c)$. As this is broken in several different cases, an outline of the proof is helpful:

1) $a = 0$ or $b = 0$ or $c = 0$
2) $a \neq 0$,$b \neq 0$, and $c \neq 0$
  2.1) $f \neq 2$
    2.1.1) $a + d \neq b + e$
      $f \mid (P - P')$
      $f \nmid (P - P')$
    2.1.2) $a + d = b + e$
      $f \mid (P - P')$
      $f \nmid (P - P')$

[9] In [10], (24) is stated as $(P - p)(P' - p) \equiv n + tp + p^2 \pmod{QQ'/s}$, but in this case $t = 0$ and $n = -N$.
[10] Some of the arguments were taken from Buell's proof in [1] concerning composition of quadratic forms.

2.2) $f = 2$

2.2.1) $a + d \neq b + e$

$c > 1,\ k > 1$

$c > 1,\ k \leq 1$

$c = 1$

2.2.2) $a + d = b + e$

$c > 1,\ k > 1$

$c = 1$

$k = 1$

Case 1) If $a = 0$, then $\max(a, c) + b + e - c = b + e \geq b$, $\max(b, c) + a + d - c \geq b$ and $a + b = b$, so the maximum exponent for f in $L(I * J)$ is $b = a + b - \min(a, b, c)$. Similarly, if $b = 0$, then the maximum exponent is $a = a + b - min(a, b, c)$.

Assume $c = 0$. $f^{a+d} \parallel QC = N - P^2$ and $f^{b+e} \parallel Q'C' = N - (P')^2$, subtracting, $f^{\min(a+d,b+e)} \mid (P^2 - (P')^2) = (P + P')(P - P')$. Since $c = 0$, then $f^{\min(a+d,b+e)} \mid (P - P')$. Therefore, $f^{\max(a,b)+\min(a+d,b+e)} \mid \{Q, Q'\}(P - P')$. However, $\max(a, b) + \min(a + d, b + e) \geq \max(a, b) + \min(a, b) = a + b$. Therefore, the maximum exponent for f in $L(I * J)$ is

$$\begin{aligned}\min(a + b, \max(a, c) + b + e - c, \max(b, c) + a + d - c) &= \min(a + b, a + b + e, a + b + d)\\ &= a + b = a + b - \min(a, b, c)\end{aligned}$$

Case 2.1.1) Assume $c \neq 0$, $f \neq 2$, and $a + d \neq b + e$. Then, $f^{\min(a+d,b+e)} \parallel (P^2 - (P')^2) = (P + P')(P - P')$. If $f \mid (P - P')$, then $f \mid 2P$ and $f \mid 2P'$. Since $f \neq 2$, this gives $f \mid P$, $f \mid P'$. Then, $d = e = 0$. Also, $c \leq \min(a, b)$. Then, the maximum exponent for f in $L(I * J)$ is

$$\min(a+b, \max(a, b)+\min(a, b)-c, \max(a, c)+b-c, \max(b, c)+a-c) = a+b-c = a+b-min(a, b, c)$$

If $f \nmid (P - P')$ then $c = \min(a + d, b + d)$ and the maximum exponent for f in $L(I * J)$ is

$$\min(a + b, \max(a, b), \max(a, c) + b + e - c, \max(b, c) + a + d - c)$$

If $a = \min(a, b, c)$, then this is $\min(b, c+b+e-c, \max(b, c)+a+d-c) = b = a+b-\min(a, b, c)$. The case is similar if $b = \min(a, b, c)$. If $c = \min(a, b, c)$, then since $c = \min(a + d, b + e)$, this gives $c = \min(a, b)$. Then the maximum exponent is

$$\min(\max(a, b), a + b + e - c, b + a + d - c) = \max(a, b) = a + b - \min(a, b) = a + b - \min(a, b, c)$$

Case 2.1.2) Assume $c \neq 0$, $f \neq 2$, but $a+d = b+e$. As before, if $f \mid (P-P')$, then $d = e = 0$. In this case also $a = b$. Assume $c \leq a$. Note that $f^{a-c} \mid (P - P')$ and $\max(a, b) + \min(a + d, b + e) - c = a + b - c$. Then the maximum exponent is

$$\min(a + b, \max(a, c) + b - c, \max(b, c) + a - c) = a + b - c = a + b - \min(a, b, c).$$

Alternately, assume $c > a$. Then for some k, $f^k \parallel (P - P')$. The maximum exponent is

$$\min(a + b, \max(a, b) + k, b, \max(b, c) + a - c) = b = a + b - a = a + b - \min(a, b, c)$$

Conversely, assume $f \nmid (P - P')$. Then $c \geq a + d = b + e$ and the maximum exponent is

$$\min(a+b, \max(a,b), \max(a,c)+b+e-c, \max(b,c)+a+d-c) = \min(\max(a,b), a+d)$$
$$= \max(a,b) = a+b-\min(a,b) = a+b-\min(a,b,c)$$

Case 2.2.1) Let $f = 2$. Assume $a + d \neq b + e$. Then $2^{\min(c,k)} \parallel 2P$ and $2^{\min(c,k)} \parallel 2P'$, so that $2^{\min(c,k)-1} \parallel P, P'$. If $c > 1$ and $k > 1$, then $d = e = 0$ and as before the largest exponent is $a + b - \min(a, b, c)$. Assume $c > 1$, $k \leq 1$. Then $k = 1$ and $c = \min(a + d, b + e) - 1$. The largest exponent is then

$$\min(a+b, \max(a,b)+1, \max(a,c)+b+e-c, \max(b,c)+a+d-c)$$

If $a \leq \min(b, c)$ this reduces to $b+\min(e, 1) = a+b+\min(e, 1)-\min(a, b, c)$. $c+1 \leq a+d \leq c+d$, so $d \geq 1$. Note that if $e \geq 1$ then this is a special case and $h = 2$. If $e = 0$, it is the same as before. The cases when $b \leq \min(a, c)$ are similar.

If $c \leq \min(a, b)$, this exponent is $\min(\max(a, b)+1, a+b+e-c, b+a+d-c)$. Without loss of generality, assume $a + d > b + e$ so that $c = a + d - 1 \geq c + d - 1$, so that $d = 1$ and $a = c$. Then the exponent is $\min(b+1, b+e, b+d) = b+\min(1, e, d) = a+b+\min(1, e, d)-\min(a, b, c)$. Note again that if $e \geq 1$ and $d \geq 1$, then this is the special case where $h = 2$. Otherwise, it is the same as before.

Assume $c = 1$. Then $k \geq 1$. Then the exponent is

$$\min(a+b, \max(a,b)+\min(a+d, b+e)-1, a+b+e-1, a+b+d-1).$$

If $d = 0$ or $e = 0$, $h = 1$ and this is $a + b - 1 = a + b - \min(a, b, c)$. Otherwise, $h = 2$ and the exponent is $a + b = a + b + 1 - \min(a, b, c)$.

Case 2.2.2) Lastly, assume that $c \neq 0$ but $a+d = b+e$. For some k, $2^k \parallel (P-P')$. $c+k \geq a+d$. If $c > 1$ and $k > 1$, then $d = e = 0$, $a = b$. The exponent is then $\min(2a, a+k, \max(a, c)+b-c)$. If $c > a$, this is $\min(2a, a + k, a) = a = a + b - \min(a, b, c)$. If $c \leq a$, the exponent is $\min(2a, a + k, 2a - c) = 2a - c = a + b - \min(a, b, c)$.

Alternately, if $c = 1$, then $k \geq a + d - 1$ and the exponent is

$$\min(a+b, \max(a,b)+k, a+b+e-1, b+a+d-1) = \min(a+b, a+b+e-1, a+b+d-1).$$

If $e > 0$ and $d > 0$, $h = 2$ and this exponent is $a + b = a + b + 1 - \min(a, b, c)$. If $e = 0$ or $d = 0$, $h = 1$ and this is $a + b - 1 = a + b - \min(a, b, c)$.

If $k = 1$ then $c \geq 1$ and specifically $c \geq a+d-1 = b+e-1$. If $c \leq \min(a, b)$, then $d = e = 1$ so that $h = 2$, $c = a = b$ and the exponent is

$$\min(2a, a+1) = a+1 = a+b+1-\min(a,b,c)$$

If $a \leq \min(b, c)$, then $d \geq e$ and the exponent is

$$\min(a+b, b+1, b+e, \max(b,c)+a+d-c) = \min(b+1, b+e)$$

If $e \geq 1$, then $d \geq e \geq 1$, so $h = 2$ and in this case the exponent is $b+1 = a+b+1-\min(a, b, c)$. If $e = 0$, $h = 1$ and in this case the exponent is $b = a + b - \min(a, b, c)$.

Therefore,

$$L(I * J) = hQQ'/s$$

so that for $f \neq 2$, the highest exponent is $a+b-\min(a,b,c)$ and for $f=2$, the highest exponent is $a+b-\min(a,b,c)$ if $h=1$ and $a+b+1-\min(a,b,c)$ if $h=2$. Note that this is still divisible by s. After that s is factored out, the result is $q = hQQ'/s^2$.

There are integers t, u, and v such that $tQ+uQ'+v(P+P') = s$. First, consider divisability by s. This is trivial for every term except $N+PP'$. By the definition of s,

$$\begin{gathered} P+P' \equiv 0 \pmod{s} \\ P' \equiv -P \pmod{s} \\ PP' \equiv -P^2 \pmod{s} \\ N+PP' \equiv N-P^2 \pmod{s} \end{gathered}$$

and since $s \mid Q$ and $Q \mid (N-P^2)$, $s \mid N+PP'$.

The linear combination of the last three elements with coefficients t, u, and v respectively is:

$$s\sqrt{N} + tQP' + uQ'P + v(N+PP')$$

so that it is evident that after s is factored out, the remaining ideal is primitive. Since this is the element of $I*J$ with the smallest coefficient of $\sqrt{N}$, clearly $p = t(Q/s)P' + u(Q'/s)P + v(N+PP')/s$, modulo $L(I*J)$. Then,

$$\begin{gathered} p = t(Q/s)P' + (s-tQ-v(P+P'))P/s + v(N+PP')/s \\ \equiv P + v(N-P^2)/s \pmod{Q/s} \\ \equiv P \pmod{Q/s} \end{gathered}$$

since $Q \mid (N-P^2)$. By symmetric arguments, $p \equiv P' \pmod{Q'/s}$.

To prove (24), consider:

$$\begin{gathered} (P+P')sp = (P+P')(tQP' + uQ'P + v(N+PP') \\ = (P+P')(tQP' + uQ'P) + (P+P')(N+PP')v \\ = (P+P')(tQP' + uQ'P) + (s-tQ-uQ')(N+PP') \\ = s(N+PP') + tQ((P')^2 - D) + uQ'(P^2 - N) \\ \equiv s(N+PP') \pmod{QQ'} \end{gathered}$$

Therefore, $(P+P')p \equiv N+PP' \pmod{QQ'/s}$. **QED**

Observe that when $h=1$ (21) could be restated as

$$L(I*J) = L(I)L(J)/s^2 \tag{26}$$

remembering that $L(I)$ is defined as the smallest positive rational integer in I. This equation is proven in [20] and will be useful later.

Also observe that for $h=1$, the equations describing the product of two ideals correspond exactly to the composition of two quadratic forms. Shanks notes this in [17]. Therefore, the equations concerning distance and multiplication of ideals will correspond to distance and composition of quadratic forms.

The case when $h=2$ connects composition of quadratic forms of discriminant $\equiv 1 \pmod 4$ to multiplication of ideals. If F and G are two quadratic forms with discriminant $N \equiv 1 \pmod 4$, then $2F$ and $2G$ have discriminat $4N$ and correspond to ideals I_{2F} and I_{2G} in $\mathbb{Z}[\sqrt{N}]$. Multiplying, $h=2$ and $I_{2F} * I_{2G} = I_{2(F*G)}$. Therefore, although this case will not be considered further, it is readily seen that the distance formulas derived from ideals in $\mathbb{Z}[\sqrt{N}]$ will still correspond to composition of quadratic forms of discriminant $\equiv 1 \pmod 4$.

5 Lattices

Consider lattices in $\mathbb{Q}(\sqrt{N}) \times \mathbb{Q}(\sqrt{N})$. Define $\mathbb{L}$ as the set of all lattices in $\mathbb{Q}(\sqrt{N}) \times \mathbb{Q}(\sqrt{N})$. Define the map $M : \mathbb{Q}(\sqrt{N}) \to \mathbb{Q}(\sqrt{N}) \times \mathbb{Q}(\sqrt{N})$ by

$$M(\xi) = \langle \xi, \overline{\xi} \rangle$$

Multiplication in $\mathbb{Q}(\sqrt{N}) \times \mathbb{Q}(\sqrt{N})$ is defined componentwise, that is, $\langle \xi, \xi' \rangle \cdot \langle \zeta, \zeta' \rangle = \langle \xi\zeta, \xi'\zeta' \rangle$, so that it is clear that M is homomorphic and one-to-one.

Distance will relate to a concept called a minimum:

Definition 5 For a vector $v = \langle v_1, v_2, ...v_d \rangle$, the *normed body* of v, $\mathcal{R}(v)$ is the set

$$\mathcal{R}(v) = \{\langle x_1, x_2, ...x_d \rangle : x_i \in \mathbb{R}, |x_i| < |v_i|, i = 1, 2, ...d\}$$

Abusing notation, denote $\mathcal{R}(\xi) = \mathcal{R}(\langle \xi, \overline{\xi} \rangle)$.

A number ξ (or actually the corresponding vector) is a *minimum* of $\mathcal{L}$ if $\mathcal{R}(\xi) \cap \mathcal{L} = \{0\}$, where 0 is the vector $\langle 0, 0 \rangle$.

A lattice $\mathcal{L}$ is *reduced* if $1 \in \mathcal{L}$ and 1 is a minimum.

For this case with $d = 2$, the normed body is a rectangle in $\mathbb{R}^2$. Note that for $\xi \in \mathbb{Q}(\sqrt{N})$ the normed body $\mathcal{R}(\xi)$ has area equal to four times the absolute value of the norm $|\mathcal{N}(\xi)|$.

To avoid unnecessary generality, this investigation will focus specifically on the lattices corresponding to ideals. Specifically, for the primitive ideal $I = [Q, \sqrt{N} + P]$, define the associated lattice containing 1 in $\mathbb{Q}(\sqrt{N})$ as $\mathcal{L}_I = [1, (\sqrt{N} + P)/Q]$.

Conversely, to each lattice containing 1 in $\mathbb{Q}(\sqrt{N})$ there is an associated primitive lattice (which may or may not be an ideal) in $\mathbb{Z}[\sqrt{N}]$. Equation (17) defined the function L. In a similar fashion, for a lattice $\mathcal{L}$, define

$$L(\mathcal{L}) = \min\{n \in \mathbb{Z}^+ : n\mathcal{L} \subset \mathbb{Z}[\sqrt{N}]\} \tag{27}$$

Then if $L(\mathcal{L})\mathcal{L}$ is an ideal of $\mathbb{Z}[\sqrt{N}]$ it is the primitive ideal associated to a lattice $\mathcal{L}$. Note that if an ideal I is associated to a lattice $\mathcal{L}_I$, then $L(I) = L(\mathcal{L}_I)$. Define

$$\Phi_{\mathbb{I},\mathbb{L}}([Q, \sqrt{N} + P]) = [1, (\sqrt{N} + P)/Q]$$

and

$$\Phi_{\mathbb{L},\mathbb{I}}(\mathcal{L}) = L(\mathcal{L})\mathcal{L}$$

Note that for some lattices $\mathcal{L}$, $\Phi_{\mathbb{L},\mathbb{I}}(\mathcal{L})$ may not actually be an ideal. Lemma 9 provides conditions for it to be an ideal sufficient for this analysis:

Lemma 9 *Let I be a primitive ideal and let $\mathcal{L} = \Phi_{\mathbb{I},\mathbb{L}}(I)$. If $\mathcal{L}'$ is a lattice with basis $\{1, \xi\}$ and for some θ, $\theta\mathcal{L}' = \mathcal{L}$, then $J = \Phi_{\mathbb{L},\mathbb{I}}(\mathcal{L})$ is a primitive ideal and*

$$(L(I)\theta)J = (L(J))I$$

Proof: Let $I = [Q, \sqrt{N} + P]$. Then $\mathcal{L} = [1, (\sqrt{N} + P)/Q]$. The statement that $\theta\mathcal{L}' = \mathcal{L}$ requires that

$$\theta \begin{bmatrix} 1 \\ \xi \end{bmatrix} = T \begin{bmatrix} 1 \\ (\sqrt{N} + P)/Q \end{bmatrix}$$

where T is a 2×2 matrix with determinant ± 1. Multiplying by $L(I) = L(\mathcal{L}) = Q$ and $L(J) = L(\mathcal{L}')$:

$$Q\theta \begin{bmatrix} L(\mathcal{L}') \\ L(\mathcal{L}')\xi \end{bmatrix} = L(\mathcal{L}')T \begin{bmatrix} Q \\ (\sqrt{N} + P) \end{bmatrix}$$

so that $(L(I)\theta)J = (L(J))I$. Therefore, J is an ideal. It is primitive by the definition of $\Phi_{\mathbb{L},\mathbb{I}}$. **QED**

For an example of minima, consider the lattice $[1, \sqrt{159} - 12]$. $\mathcal{R}(1)$ is a square with sides of length 2 centered at the origin and a simple graph demonstrates that 0 is the only point in the lattice and contained in this square. Therefore 1 is a minimum. $\sqrt{159} - 12$ is also a minimum. $\mathcal{R}(\sqrt{159} + 12)$ is a narrower and taller rectangle also centered at the origin.

Given two minima, it is important to be able to determine whether or not there is another minimum between them. In vector format, if $\langle x_1, y_1 \rangle$ and $\langle x_2, y_2 \rangle$ are minima with $|x_1| > |x_2|$ and $|y_1| < |y_2|$, these two minima are *adjacent* if there does not exist another minima $\langle x_3, y_3 \rangle$ such that $|x_2| < |x_3| < |x_1|$ and $|y_1| < |y_3| < |y_2|$.

Voronoi developed a method (and a theorem) concerning adjacent minima ([3], [20]).

Theorem 9 *Let $\mathcal{L}$ be a lattice with $\{\xi, \zeta\}$ as a basis, where $\xi, \zeta \in \mathbb{Q}(\sqrt{N})$ and suppose that $\zeta > \xi > 0$. Then ζ and ξ are adjacent minima of $\mathcal{L}$ if and only if $|\overline{\xi}| > |\overline{\zeta}|$ and $\overline{\zeta}\overline{\xi} < 0$.*

Proof: Assume ξ and ζ are adjacent minima. Since they are both minima, $|\overline{\xi}| > |\overline{\zeta}|$, or else ζ would not be a minima. Also $0 < \zeta - \xi < \zeta$. Since ζ is a minima, this requires that $|\overline{\zeta} - \overline{\xi}| > |\overline{\zeta}|$. If $\overline{\zeta}$ and $\overline{\xi}$ had the same sign, this would not be possible. Therefore, $\overline{\zeta}\overline{\xi} < 0$.

Conversely, assume that $|\overline{\xi}| > |\overline{\zeta}|$ and $\overline{\zeta}\overline{\xi} < 0$. Assume that ξ is not a minimum of $\mathcal{L}$. Then there exists some $\omega \in \mathbb{Q}(\sqrt{N})$ such that $|\omega| < \xi$ and $|\overline{\omega}| < |\overline{\xi}|$. Since $\omega = a\xi + b\zeta$ for some $a, b \in \mathbb{Z}$, $|a\xi + b\zeta| < \xi$ and $|a\overline{\xi} + b\overline{\zeta}| < |\overline{\xi}|$. If $ab = 0$, then either $a = 0$ or $b = 0$. If $a = 0$, then the second statement contradicts the hypothesis. If $b = 0$, then the first statement gives $\xi < \xi$, clearly false. However, if $ab > 0$ then $|a\xi + b\zeta| > \xi$ and if $ab < 0$, then since $\overline{\zeta}\overline{\xi} < 0$, $|a\overline{\xi} + b\overline{\zeta}| > |\overline{\xi}|$. Therefore, ξ must be a minima. By similar reasoning, ζ must be a minima.

Concerning adjacency, assume that there is another minima ω between ξ and ζ. Since $\omega = a\xi + b\zeta$ for some $a, b \in \mathbb{Z}$, $\xi < |a\xi + b\zeta| < \zeta$ and $|\overline{\zeta}| < |a\overline{\xi} + b\overline{\zeta}| < |\overline{\xi}|$. Since $\zeta > \xi > 0$, the first statement requires that $b = 0$ and then the second statement simplifies to $|a| < 1$, requiring that $a = 0$ and providing a contradiction. Therefore, ξ and ζ are adjacent minima. **QED**

From the previous example, it is now possible to check that $\xi = 1$ and $\zeta = \sqrt{159} + 12$ are indeed adjacent minima.

The idea that will actually connect to continued fractions (and distance) is the search for a sequence of adjacent minima. This sequence is formed by relating different lattices. The following Lemmas are due to Williams [20].

Lemma 10 *Let $\mathcal{L}$ and $\mathcal{L}'$ be reduced lattices. If $\xi\mathcal{L}' = \mathcal{L}$, then ξ is a minimum of $\mathcal{L}$.*

Proof: Since $1 \in \mathcal{L}'$, $\xi \in \mathcal{L}$. If ξ is not a minimum of $\mathcal{L}$, then there exists a $\zeta \in \mathcal{L}$ suh that $\zeta \neq 0$ and $|\zeta| < |\xi|$ and $|\overline{\zeta}| < |\overline{\xi}|$. Let $\beta = \zeta/\xi$, so that $\beta \in \mathcal{L}'$. $|\beta| = |\zeta/\xi| < 1$ and $|\overline{\beta}| = |\overline{\zeta}/\overline{\xi}| < 1$, contradicting the fact that $\mathcal{L}'$ is reduced. Therefore, ξ is a minimum of $\mathcal{L}$. **QED**

Now consider the converse of this statement. Note that $\lfloor x \rfloor$ denotes the floor of x.

Lemma 11 *Let $\mathcal{L} = [1, \xi]$, where 1 and ξ are adjacent minima of $\mathcal{L}$ with $1 > \xi > 0$. Let $\mathcal{L}' = (1/\xi)\mathcal{L}$. Then $\mathcal{L}'$ is a reduced lattice.*

Proof: $\mathcal{L}' = (1/\xi)[1,\xi] = [1/\xi, 1] = [1/\xi - \lfloor 1/\xi \rfloor, 1]$, so that $1 \in \mathcal{L}'$. It is sufficient to show that 1 and $\xi' = 1/\xi - \lfloor 1/\xi \rfloor$ are adjacent minima. First, 1 and ξ' are a basis for $\mathcal{L}'$ and $1 > \xi' > 0$. Since $0 < \xi < 1$, $\lfloor 1/\xi \rfloor > 1$. Since $\overline{\xi} < 0$, $\overline{\xi'} = 1/\overline{\xi} - \lfloor 1/\xi \rfloor < 0 - 1 = -1$. Thereby satisfying both the requirement that $\overline{\xi}' \cdot 1 < 0$ and the requirement that $|\overline{\xi}'| > 1$. Therefore, by Theorem 9, 1 and ξ' are adjacent minima of $\mathcal{L}'$ and thus $\mathcal{L}'$ is a reduced lattice. **QED**

Actually, these proofs provide a bit more by actually finding the minimum adjacent to 1 in the new lattice. The next Lemma makes use of this minimum [20]:

Lemma 12 *Let $\mathcal{L}$, $\mathcal{L}'$, ξ, and ξ' be as above. Let ζ be the minimum adjacent to ξ other than* 1 *in $\mathcal{L}$. Then $\zeta = \xi\xi'$.*

Proof: $\xi\xi' = \xi(1/\xi - \lfloor 1/\xi \rfloor) = 1 - \xi\lfloor 1/\xi \rfloor$, so that $[\xi, \xi\xi']$ is a basis for $\mathcal{L}$. Since $1 > \xi' > 0$, $\xi > \xi\xi' > 0$. Since $|\overline{\xi}'| > 1$, $|\overline{\xi\xi}'| > |\overline{\xi}|$. Since $\overline{\xi}' < 0$, $\overline{\xi \cdot \xi\xi'} = \overline{(\xi)^2\xi'} < 0$. Therefore, by Theorem 9, ξ and $\xi\xi'$ are adjacent minima. Since $\xi\xi' \neq 1$, $\zeta = \xi\xi'$. **QED**

Observe that by a similar process, one could find a reduced lattice $\mathcal{L}'' = 1/\xi'\mathcal{L}'$, etc. Then $\mathcal{L}'' = 1/(\xi\xi')\mathcal{L}$. To generalize, define $\xi = \xi_1$ and $\mathcal{L} = \mathcal{L}_1$ and this is a sequence of reduced lattices and their minima, A chain of adjacent minima of $\mathcal{L}_1$ may be defined by

$$\theta_n = \prod_{i=1}^{n-1} \xi_i \tag{28}$$

and then

$$\theta_n \mathcal{L}_n = \mathcal{L}_1 \tag{29}$$

Since each $\mathcal{L}_n$ is a reduced lattice, by Lemma 10 each θ_n is a minimum of $\mathcal{L}_1$.

Although it is not true in higher dimensions, it is fairly trivial in 2-d that this chain of adjacent minima provides a complete (although infinite) list of the minima with x-coordinate between 0 and 1.

Lemma 13 *Let $\langle \phi, \overline{\phi} \rangle$ be a minimum of a lattice $\mathcal{L}$, with $0 < \phi < 1$. Then for some n, $\phi = \theta_n$, where θ_n is defined by equation (28)*

Define *distance* in terms of this chain of minima by

$$D_{\mathbb{L}}(\mathcal{L}_n, \mathcal{L}_m) = \log(\theta_n/\theta_m) \tag{30}$$

It will become readily apparent that the subscript $\mathbb{L}$ is unnecessary, but it provides clarity for now. Before continuing it is appropriate to provide an example of these concepts. First, as a reference, consider the steps for the continued fraction expansion of $\sqrt{159} - 12$ and the quadratic form distances $D_{\mathbb{F}}$ covered to the end of each step:

$$\begin{aligned}
x_1 &= \tfrac{1}{\sqrt{159}-12} = \tfrac{\sqrt{159}+12}{15} = 1 + \tfrac{\sqrt{159}-3}{15}, & D_{\mathbb{F}}(F_0, F_1) &= \log(\tfrac{\sqrt{159}+12}{15}) \\
x_2 &= \tfrac{15}{\sqrt{159}-3} = \tfrac{\sqrt{159}+3}{10} = 1 + \tfrac{\sqrt{159}-7}{10}, & D_{\mathbb{F}}(F_0, F_2) &= \log(\tfrac{\sqrt{159}+13}{10}) \\
x_3 &= \tfrac{10}{\sqrt{159}-7} = \tfrac{\sqrt{159}+7}{11} = 1 + \tfrac{\sqrt{159}-4}{11}, & D_{\mathbb{F}}(F_0, F_3) &= \log(\tfrac{2\sqrt{159}+25}{11}) \\
x_4 &= \tfrac{11}{\sqrt{159}-4} = \tfrac{\sqrt{159}+4}{13} = 1 + \tfrac{\sqrt{159}-9}{13}, & D_{\mathbb{F}}(F_0, F_4) &= \log(\tfrac{3\sqrt{159}+38}{13}) \\
x_5 &= \tfrac{13}{\sqrt{159}-9} = \tfrac{\sqrt{159}+9}{6} = 3 + \tfrac{\sqrt{159}-9}{6}, & D_{\mathbb{F}}(F_0, F_5) &= \log(\tfrac{5\sqrt{159}+63}{6}).
\end{aligned}$$

The continued fraction corresponds to quadratic forms which correspond to ideals, which are associated with lattices that contain 1. In this case, the lattice associated with x_1 is $\mathcal{L}_1 = [1, 1/x_1] = [1, \sqrt{159} - 12]$ and $1/x_1$ is a minimum adjacent to 1 in $\mathcal{L}_1$. From here:

$$\mathcal{L}_2 = \tfrac{1}{\sqrt{159}-12}\mathcal{L}_1 = [\tfrac{1}{\sqrt{159}-12}, 1] = [\tfrac{\sqrt{159}+12}{15}, 1] = [1, \tfrac{\sqrt{159}-3}{15}] = [1, 1/x_2]$$
$$\mathcal{L}_3 = \tfrac{15}{\sqrt{159}-3}\mathcal{L}_2 = [\tfrac{15}{\sqrt{159}-3}, 1] = [\tfrac{\sqrt{159}+3}{10}, 1] = [1, \tfrac{\sqrt{159}-7}{10}] = [1, 1/x_3]$$
$$\ldots$$

and it is apparent that this same pattern of correspondance will continue, that is

$$\Phi_{\mathbb{T},\mathbb{L}}(x_n) = \Phi_{\mathbb{T},\mathbb{I}}(\Phi_{\mathbb{I},\mathbb{L}}(x_n)) = [1, 1/x_n] = \mathcal{L}_n \tag{31}$$

from which it is also apparent that the sequences of lattices will be periodic.

Computing equation (28), for example, $\theta_3 = (\sqrt{159}-12)(\frac{\sqrt{159}-3}{15}) = 13-\sqrt{159}$. With $\theta_1 = 1$,

$$D(\mathcal{L}_1, \mathcal{L}_3) = \log(1/(13 - \sqrt{159})) = \log((\sqrt{159} + 13)/10)$$

It is readily apparent that the definition of distances in lattices corresponds to the definition given for quadratic forms. Note that these distances must still be considered modulo R, the regulator, since the sequence of lattices is still cyclic.

6 The Generalized Distance Formula

Going back to ideals, note that if $I_1 = L(\mathcal{L}_1)\mathcal{L}_1$ and $I_n = L(\mathcal{L}_n)\mathcal{L}_n$ is another ideal corresponding to a lattice later in the same sequence, then

$$\begin{aligned} \theta_n\mathcal{L}_n &= \mathcal{L}_1 \\ L(\mathcal{L}_1)L(\mathcal{L}_n)\theta_n\mathcal{L}_n &= L(\mathcal{L}_1)L(\mathcal{L}_n)\mathcal{L}_1 \\ (L(\mathcal{L}_1)\theta_n)I_n &= (L(\mathcal{L}_n)I_1 \end{aligned} \tag{32}$$

where once again, the distance (this time between ideals) is given by $D(I_1, I_n) = -\log(\theta_n)$. Now, this definition of distance is well and good for reduced ideals, but as of yet, it hasn't been applied it to non-reduced ideals. To relate the definitions of reduced lattices and continued fractions observe that the definition of a reduced continued fraction implies that for a term $x_i = \frac{\sqrt{N}+P_{i-1}}{Q_i}$, being reduced equates to

$$\frac{\sqrt{N} + P_{i-1}}{Q_i} > 1$$
$$0 < \frac{\sqrt{N} - P_{i-1}}{Q_i} < 1$$

so that it is clear that if $\mathcal{L}_x = [1, 1/x]$, then 1 and x are adjacent minima and the lattice is reduced. The process of dealing with a non-reduced lattice correlates to the process of reducing a continued fraction as demonstrated in the proof of Lemma 5. See [20] for a more general Lemma.

Lemma 14 *Let I be any primitive ideal in $\mathbb{Z}[\sqrt{N}]$. There exists a reduced ideal I_n and a $\theta_n \in I$ such that*

$$(L(I)\theta_n)I_n = (L(I_n))I \tag{33}$$

Proof:

Let $I = [Q, \sqrt{N}+P]$. Then the associated lattice is $\mathcal{L}_I = [1, \frac{\sqrt{N}+P}{Q}] = [1, \xi_1]$. If I is reduced, $I_n = I$, $u = L(I)$, and the proof is done. If I is not reduced, then $\mathcal{L}_I$ is not reduced. Without loss of generality, assume that $0 < \xi_1 < 1$ (since otherwise it would just have to be reduced by an integer.). Let $\mathcal{L}_2 = 1/\xi_1\mathcal{L}_I = [1/\xi, 1] = [1, 1/\xi - \lfloor 1/\xi - 1/2 \rfloor]$. Then $\xi_1\mathcal{L}_2 = \mathcal{L}_I$. Continuing in similar manner[11], by Lemma 5 and the correspondance between lattices and continued fractions for some n,ξ_n reduced, and thus $\mathcal{L}_n$ reduced. As in (28), set

$$\theta_n = \prod_{i=1}^{n-1} \xi_i$$

so that

$$\theta_n\mathcal{L}_n = \mathcal{L}_I$$

Then $(L(I)\theta_n)I_n = (L(I_n)I$. **QED**

Let I_1, J_1 be reduced primitive ideals. Let K_1 be the primitive ideal found by multiplying I_1 and J_1 and removing a factor and let s be the factor removed, so that $(s)K_1 = I_1J_1$, $s \in \mathbb{Z}$. By Lemma 14 there exists a reduced ideal K_j and a $\lambda_j \in K_1$ such that

$$(L(K_1)\lambda_j)K_j = (L(K_j))K_1 \tag{34}$$

corresponding to$D(K_1, K_j) = -\log(\lambda_j)$.

Let $I_n \sim I_1$ and $J_m \sim J_1$ and let H_1 be the primitive ideal found by multiplying I_n and J_m and removing a factor and let t be the factor removed, so that $(t)H_1 = I_nJ_m$, $t \in \mathbb{Z}$. By Lemma 14 there exists a reduced ideal H_k and a $\eta_k \in K_1$ such that

$$(L(H_1)\eta_k)H_k = (L(H_k))H_1 \tag{35}$$

corresponding to $D(H_1, H_k) = -\log(\eta_k)$.

Also, there exist minima μ_n and ϕ_m in the lattices corresponding to I_1 and J_1, respectively, such that

$$(L(I_1)\mu_n)I_n = (L(I_n))I_1 \tag{36}$$

and

$$(L(J_1)\phi_m)J_m = (L(J_m))J_1 \tag{37}$$

corresponding to $D(I_1, I_n) = -\log(\mu_n)$ and $D(J_1, J_m) = -\log(\phi_m)$.

By combining (26) and (34)-(37):

[11]Note that it is irrelevent whether or not the second components of the intermediate lattices are either minima or adjacent to 1. Also note that, as in Lemma 5, the formula would change slightly when the denominators get small.

$$
\begin{aligned}
(L(H_k))K_j &= \left(\frac{L(H_k)L(K_j)}{L(K_1)\lambda_j}\right) K_1 \\
&= \left(\frac{L(H_k)L(K_j)}{L(K_1)\lambda_j s}\right) I_1 J_1 \\
&= \left(\frac{L(H_k)L(K_j)L(I_1)L(J_1)\mu_n\phi_m}{L(K_1)\lambda_j sL(I_n)L(J_m)}\right) I_n J_m \\
&= \left(\frac{L(H_k)L(K_j)s\mu_n\phi_m}{\lambda_j L(I_n)L(J_m)}\right) I_n J_m \\
&= \left(\frac{L(H_k)L(K_j)s\mu_n\phi_m t}{\lambda_j L(I_n)L(J_m)}\right) H_1 \\
&= \left(\frac{L(H_k)L(K_j)s\mu_n\phi_m}{\lambda_j tL(H_1)}\right) H_1 \\
&= \left(\frac{L(K_j)s\mu_n\phi_m\eta_k}{\lambda_j t}\right) H_k
\end{aligned}
$$

Set

$$\psi = \frac{s\mu_n\phi_m\eta_k}{t\lambda_j}$$

and then

$$(L(K_j)\psi)H_k = (L(H_k))K_j \tag{38}$$

Since K_j and H_k are reduced, by Lemma 10 ψ is a minimum of the lattice $\mathcal{L}_{K_j}$, so that for some n, $\psi = \theta_n$. Therefore,

$$
\begin{aligned}
D(K_j, H_k) = -\log(\psi) &= -\log(\mu_n) - \log(\phi_m) - \log(\eta_k) + \log(\lambda_j) - log(s/t) \\
&= D(I_1, I_n) + D(J_1, J_m) + \zeta
\end{aligned}
$$

where $\zeta = D(H_1, H_j) - D(K_1, K_j) + log(t/s)$ will be small compared to $D(K_j, H_k)$ for m, n large.

By the correspondance between multiplication of ideals and composition of quadratic forms, this result may be restated in terms of forms:

Theorem 10 *If $F_1 \sim F_n$ are equivalent forms and $G_1 \sim G_m$ 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 $F_n * G_m$ and s and t are the factors cancelled in each respective composition, then*

$$D(F_1 * G_1, F_n * G_m) = D(F_1, F_n) + D(G_1, G_n) + \zeta$$

where $\zeta = D_{\rho,2} - D_{\rho,1} + \log(t/s)$.

Example 5 from the Morrison-Brillhart algorithm is in the principal cycle. By Theorem 10 when a form F is composed with itself, the distance from 1 to F is roughly doublied, $d(1, F^2) = 2d(1, F) + \zeta$. Therefore, the index is roughly doubled, since distance is roughly proportional to the difference in indices, so that $F_3 * F_3 = F_6$, and $Q_6 = 9$ is the square of $Q_3 = 3$.

Since the square of any symmetry point has first coefficient 1, observe that if the distance around some cycle were unrelated to the distance around the principal cycle, then this result would be affected by which symmetry point this distance was referenced from. From Definition 3 $R = D(F_0, F_\pi)$ in the principal cycle. At this point, it is clear that the distance in other cycles must be the same.

Lemma 15 *Let A be a primitive amibiguous cycle with a period π. Then,*

$$R = D(F_0, F_\pi)$$

Proof: Let $\{F_i\}$ have period π and let F_0 and $F_{\pi/2}$ be the two symmetry points of A. Then $F_0 * F_0 = 1 = F_{\pi/2} * F_{\pi/2}$, with $D_{\rho,1} = D_{\rho,2} = 0$, s and t the respective first coefficients. Therefore,

$$0 = D(F_0 * F_0, F_{\pi/2} * F_{\pi/2}) = 2D(F_0, F_{\pi/2}) + \log(t/s) = D(F_0, F_\pi)$$

where the 3rd step is obtained from the 2nd by the fact that the product in $D(F_0, F_{\pi/2})$ includes the last denominator t and not the first denominator s.

Therefore, $D(F_0, F_\pi) = nR$. Considering composition of F_0 with forms in the principal cycle, clearly $D(F_0, F_\pi) \leq R$, so that $D(F_0, F_\pi) = R$. **QED**

7 Square Forms Factorization (SQUFOF)

It's not certain how much Shanks may have rigorously proven concerning distances, but based on the understanding he had of distance and infrastructure, he was able to develop Square Forms Factorization. A short example will demonstrate and explain the algorithm: let $N = 3193$. Expanding the continued fraction (principal cycle), $Q_{10} = 49$. The quadratic form for this is $F = 49x^2 + 58xy - 48y^2$. Since 49 is a perfect square, $7x^2 + 58xy - 336y^2$, which reduces with $D_\rho = 0$ to $G = 7x^2 + 100xy - 99y^2$ is a quadratic form whose square is F. Therefore, by Theorem 7, G is in a class of order 2 or 1, so that G is an ambiguous form, so that there are two points of symmetry in its cycle. Since by Theorem 10, $2D(G_s, G) = D(1, F) \pmod R$. So $D(G_s, G) = D(1, F)/2 \pmod{R/2}$. Since the two points of symmetry are $R/2$ away from each other, this means that there is a symmetry point at distance $D(1, F)/2$ 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 7 somewhere before F in the continued fraction expansion. If the coefficient is 2, then this symmetry point could be composed with G to find 14 at an earlier point in the principle cycle. Therefore, the symmetry point provides a nontrivial factor for N. In this case, after 6 steps it provides 31 as a factor of 3193.

The second phase of this algorithm can be made significantly (at least for larger numbers) faster if the quadratic forms from the continued fraction expansion with indices that are powers of 2 are saved. In this example, $F = F_{10}$, so that G is about the 5^{th} form in its cycle[12]. The composition of G^{-1} with F_4 and F_1 is close and a simultaneous search in both direction from there quickly finds the symmetry point. In this case, it is only necessary to store $\log_2 k$ forms for k steps, so that it is more efficient to check each square to see if it works than to check each square root against the previous pseudo-squares to predict whether it will work.

Formally, here is the algorithm for factoring N:

[12]Roughly, since in this case $5 \approx 6$.

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

In [18], Shanks states that for N having $k+1$ distinct prime factors, the average distance between two square quadratic forms that provide a factorization of N is

$$\Delta n = \ln(8)\frac{2+\sqrt{2}}{4}\frac{\sqrt[4]{N}}{2^k - 1}.$$

Since finding a square form is the slowest portion of the algorithm, this quickly would have proven that SQUFOF is an $O(\sqrt[4]{N})$ algorithm. Unfortunately, Shanks did not provide a proof for this statement. Lacking this, a proof of the runtime for SQUFOF has not been found, but the runtime may be estimated fairly well [1]. Since for a reduced form with first coefficient a, $0 < a < 2\sqrt{N}$, there are $O(\sqrt{N})$ integers that could potentially be the first coefficient of a quadratic form (where the constant is affected primarily by the number of factors of N). In order for a square of a form to be reduced, its first coefficient must be less than $\sqrt{2}N^{1/4}$, so that there are $O(N^{1/4})$ of these. At worst case, N has only 1 ambiguous cycle other than the principal cycle, so that only roughly half of these square forms are in the non-principal ambiguous cycle, but this only introduces a constant to the calculations. Assuming even distribution, these two estimates may be divided to estimate the number of forms between each square form, to get $O(\sqrt{N}/N^{1/4}) = O(N^{1/4})$. Compared to finding a square form, the other parts of the algorithm are negligibly fast, so that the expected runtime is $O(N^{1/4})$.

8 Acknowledgements

I must thank H. Williams and D. Buell for providing me with whatever of Shanks notes they could find, as well as providing their own research on the subject.

References

[1] Buell, D. A. *Binary Quadratic Forms: Classical Theory and Modern Computations*. New York: Springer-Verlag, 1989.

[2] Crandall, Richard and Carl Pomerance. *Prime Numbers: a Computational Perspective*. New York: Springer. 2001

[3] Delone, B. N. and D. K. Faddeev. *The Theory of Irrationalites of the Third Degree*. American Mathematical Society. Providence, RI. 1964.

[4] Ellis, J. H. "The history of non-secret encryption". *Cryptologia*. Number 23, July 1999, p 267-273.

[5] Davenport, Harold. *Multiplicative Number Theory*. Springer-Verlag. New York. 1980.

[6] Gauss, Carl Friedrich. *Disquisitiones Arithmeticae*. Trans. by Arthur A. Clarke. New Haven, Yale University Press, 1966.

[7] Hardy, G. H. and Wright, E. M. *An Introduction to the Theory of Numbers*. Oxford: Clarendon Press. 1979.

[8] L. K. Hua, *Introduction to Number Theory*. Springer-Verlag. New York. 1982.

[9] Lenstra, H. W. Jr. "A New Method for Factoring Integers". Lecture Notes, 2001, University of California, Berkeley. p 5.

[10] Mollin, R. A. and A. J. van der Poorten. "A Note on Symmetry and Ambiguity". *Bull. Austral. Math. Soc.* Austral. Math. Publ. Assoc. 1995

[11] Morrison, Michael A. and John Brillhart "A Method of Factoring and the Factorization of F_7." Mathematics of Computation, Vol. 29, No. 129 (Jan, 1975), 183-205.

[12] Poorten, Alfred J. "A Note on NUCOMP". *Mathematics of Computation*. Volume 72, Number 244, April 2003, p 1935-1946.

[13] Riesel, Hans. *Prime Numbers and Computer Methods for Factorization*. Boston : Birkhuser, 1985. p 191-195, 300-317.

[14] Rivest, R., A. Shamir and L. Adleman. "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems". *Communications of the ACM*, 21 (2), p 120-126, February 1978

[15] Shanks, Daniel. "Analysis and Improvement of the Continued Fraction Method of Factorization." Unpub. circa 1975. Latexed by Stephen McMath March 2004. <http://cadigweb.ew.usna.edu/∼wdj/mcmath/>

[16] Shanks, Daniel. "The Infrastructure of a Real Quadratic Field and its Applications." Proceedings of the 1972 Number Theory Conference : University of Colorado, Boulder, Colorado, August 14-18, 1972.

[17] Shanks, Daniel. "On Gauss and Composition II". Ed. R. A. Mollin. *Number Theory and Applications*. Kluwer Academic Publishers. 1989. p 179-204.

[18] Shanks, Daniel. "SQUFOF Notes." circa 1975. Unpub. Latexed by Stephen McMath March 2004. <http://cadigweb.ew.usna.edu/∼wdj/mcmath/>

[19] Trappe, Wade and Lawrence C. Washington. *Introduction to Cryptography with Coding Theory*. New Jersey: Prentice-Hall, Inc. 2002. p 137-142.

[20] Williams, Hugh C. "Continued Fractions and Number-Theoretic Computations." *Rocky Mountain Journal of Mathemtaics*. Volume 15, no. 2, Spring 1985.