STEWARD, STEPHEN MICHAEL, M.A. Zagier's Reduction Theory for Indefinite Binary Quadratic Forms and the Fermat-Pell Equation. (2020)
Directed by Dr. Brett Tangedal. 170 pp.

We give an in depth description of indefinite binary quadratic forms with a particular emphasis on Zagier's reduction theory for such forms. We also connect this theory with the theory of minus continued fractions and as a further application we offer a constructive approach to solving the Fermat-Pell Equation in two different settings.

ZAGIER'S REDUCTION THEORY FOR INDEFINITE BINARY QUADRATIC FORMS AND THE FERMAT-PELL EQUATION

by

Stephen Michael Steward

A Thesis Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Greensboro
2020

Approved by

_____
Committee Chair

APPROVAL PAGE

This thesis written by Stephen Michael Steward has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____
Brett Tangedal

Committee Members _____
Sebastian Pauli

_____
Dan Yasaki

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Brett Tangedal, for instilling in me a true passion for Number Theory, both in MAT 709 and beyond: his enthusiasm and dedication to the subject is truly inspiring. I would also like to thank my committee members, Dr. Pauli and Dr. Yasaki, for their support, encouragement, and feedback towards the completion of this thesis.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

1.1. Notation and Conventions

The classical working environment in which Elementary Number Theory takes place is the set of integers. The integers consist of the counting numbers $1, 2, 3, \ldots$, their negatives, and zero. More succinctly, the symbol $\mathbb{Z}$ is used to denote the set of integers:

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}.$$

The reason that the letter Z is used in this context is that the German word for "number" is Zahl. Two subsets of special interest within the full set of integers are the positive integers (in other words, the counting numbers), which we denote by $\mathbb{Z}^+$, and the set of all integers that are greater than or equal to 0, which we denote by $\mathbb{Z}^{\geq 0}$.

We will make frequent use of lower case letters to represent integers or functions; the distinction will be made the first time a new letter is introduced. Matrices shall be represented using bold upper case letters. Our work deals primarily with integers, but we will have occasion to work within the larger sets of rational and real numbers, denoted by $\mathbb{Q}$ and $\mathbb{R}$, respectively. We state several theorems in this first section, some of which are of great importance, such as Theorem 1.1.11. Proofs of the majority of these theorems may be found in the book by Landau [5].

Due to our emphasis on working within the set of integers, it is essential to discuss the notion of "divisibility."

Definition 1.1.1. We say that a nonzero integer a *divides* another integer b, denoted

by $a \mid b$, if there exists an integer c such that $ac = b$. In this case, we say that a is a *divisor* or *factor* of b. If such an integer c does not exist, then a does not divide b, which we denote by $a \nmid b$.

For example, $2 \mid 10$ since $2 \cdot 5 = 10$. However, $2 \nmid 11$ since there is no integer c such that $2 \cdot c = 11$. Note that $1 \mid b$ for *every* integer b. Also, we have $a \mid a$ for every nonzero integer a.

No discussion of divisibility would be complete without defining the *greatest common divisor*.

Definition 1.1.2. Let $a, b, c \in \mathbb{Z}$ with $a \neq 0$. The *greatest common divisor* of a and b, which we denote by $\gcd(a, b)$, is the largest positive integer that simultaneously divides both a and b. Similarly, the largest positive integer that simultaneously divides a, b, and c is called the *greatest common divisor* of a, b, and c, and is denoted by $\gcd(a, b, c)$.

Definition 1.1.3. Let $a, b, c \in \mathbb{Z}$ with $a \neq 0$. If $\gcd(a, b) = 1$, then a and b are said to be *relatively prime*. Similarly, if $\gcd(a, b, c) = 1$, then the triple of integers a, b, and c is said to be *relatively prime*.

A nice result that establishes a connection between the greatest common divisor of a pair of integers with the greatest common divisor of a triple of integers is the following.

Theorem 1.1.4. *If $a, b, c \in \mathbb{Z}$ with $a \neq 0$, then* $\gcd(a, b, c) = \gcd(\gcd(a, b), c)$.

An extremely useful result about the greatest common divisor of two integers is that it may be expressed as a linear combination of these two integers. More precisely, we have

Theorem 1.1.5. *Given $a, b \in \mathbb{Z}$ with $a \neq 0$, there exist integers $s, t \in \mathbb{Z}$ such that $sa + tb = \gcd(a, b)$.*

A special case of this theorem deserves a separate mention of its own.

Corollary 1.1.6. *Given $a, b \in \mathbb{Z}$ with $a \neq 0$, we have $\gcd(a, b) = 1$ if and only if there exist integers $s, t \in \mathbb{Z}$ such that $sa + tb = 1$.*

Using Theorems 1.1.4 and 1.1.5 in conjunction, we may prove a natural generalization of Corollary 1.1.6 which applies to a triple of relatively prime integers. We only state and prove the one direction of this generalization that will be called upon later in Section 2.3.

Corollary 1.1.7. *If $a \neq 0$, b, and c form a relatively prime triple of integers, then there exist integers $s, t, u \in \mathbb{Z}$ such that $sa + tb + uc = 1$.*

Proof. Let $d = \gcd(a, b)$, and use Theorem 1.1.5 to choose integers x and y such that $xa + yb = d$. If we set $e = 1 = \gcd(a, b, c)$, then Theorems 1.1.4 and 1.1.5 together tell us that there exist integers $u, v \in \mathbb{Z}$ such that $e = vd + uc$. Straightforward substitution yields the following:

$$1 = e = v(xa + yb) + uc = vxa + vyb + uc,$$

and setting $s = vx$ and $t = vy$ completes the proof. □

Another basic result that we will need later (which also provides a crucial step in the proof of the Fundamental Theorem of Arithmetic stated just below) is the following:

Theorem 1.1.8. *Let a, b, c be nonzero integers and assume that $\gcd(a, b) = 1$. If $a \mid bc$, then $a \mid c$.*

We noted above that if $a \geq 2$ is a positive integer, then it has at least two distinct positive divisors, namely, 1 and a itself. Certain special integers, such as $2, 3, 5$, and 7, have no other positive divisors aside from the two just mentioned. This observation is critical to the whole subject of Elementary Number Theory and leads to the following crucial definition.

Definition 1.1.9. A positive integer $p \geq 2$ is said to be a *prime number* if the only two positive integer divisors of p are 1 and p itself. A positive integer $m \geq 2$ that is not prime is said to be a *composite number*.

For example, $4, 6, 8, 9, 10$, and 12 are all composite numbers. We listed the first 4 prime numbers above, in increasing order, and it is easy to continue this ordered list: $11, 13, 17, 19, 23, 29, 31, \ldots$. Indeed, the list of prime numbers is never-ending, which is the content of the following important theorem dating back to the mathematics of ancient Greece.

Theorem 1.1.10. *There are infinitely many prime numbers.*

From a multiplicative standpoint, the prime numbers are best viewed as the "building blocks" from which all positive integers arise. It is very important to give a precise formulation of this, and the resulting statement is—with no exaggeration—known as the Fundamental Theorem of Arithmetic. The statement that we give here is perhaps slightly unorthodox, but useful for our purposes.

Theorem 1.1.11. *For every integer $n \in \mathbb{Z}^+$, there is a prime factorization*

$$n = \prod_p p^{e_p(n)},$$

*where the product is taken over all prime numbers, only finitely many of the exponents $e_p(n)$ are positive, and these exponents are all **uniquely** determined by n.*

For example, if $n = 1$ we have $e_p(1) = 0$ for all primes p. If $n = 25$, we have $e_5(25) = 2$ and $e_p(25) = 0$ for all other primes p. The primes form one important infinite sequence of positive integers and the "perfect squares" form another such sequence.

Definition 1.1.12. A positive integer n is said to be a *perfect square* if it has the form $n = b^2$, for some nonzero integer b.

The list of perfect squares is easy to generate; in ascending order the list begins as follows: $1, 4, 9, 16, 25, 36, 49, 64, 81, 100, \ldots$. A useful characterization of the perfect squares may be given in terms of the exponents appearing in their prime factorization.

Theorem 1.1.13. *An integer $n \in \mathbb{Z}^+$ is a perfect square if and only if every exponent $e_p(n)$ appearing in the prime factorization of n in Theorem 1.1.11 is even.*

The next result, also related to perfect squares, will have important consequences for this thesis as well.

Theorem 1.1.14. *Given $n \in \mathbb{Z}^+$, the square root $\sqrt{n}$ is an irrational number if and only if n is not a perfect square.*

We will have occasion to make use of the notion of "congruence" with respect to a given modulus m. The notation used for this was introduced by Gauss in his epoch-making book entitled *Disquisitiones Arithmeticae* [3] (this title translates from Latin into English as "Arithmetical Investigations").

Definition 1.1.15. Let $a, b, m \in \mathbb{Z}$ with $m \geq 2$. We say that "*a is congruent to b modulo m*" if $m \mid (a - b)$, which we denote by $a \equiv b \pmod{m}$.

For example, $10 \equiv 2 \pmod{4}$ since $4 \mid (10 - 2)$. However, $13 \not\equiv 3 \pmod{4}$ since $4 \nmid (13 - 3)$. A list of common properties of congruence is stated below.

Theorem 1.1.16. *Let $a, b, c, m \in \mathbb{Z}$ with $m \geq 2$. Then,*

(i) $a \equiv a \pmod{m}$

(ii) *if* $a \equiv b \pmod{m}$, *then* $b \equiv a \pmod{m}$, *and*

(iii) *if* $a \equiv b \pmod{m}$ *and* $b \equiv c \pmod{m}$, *then* $a \equiv c \pmod{m}$.

There are many other properties of modular congruence, but those listed above are the properties that will be referenced most often throughout this thesis. Theorem 1.1.16 demonstrates that for any fixed $m \geq 2$, "congruence modulo m" defines an equivalence relation on $\mathbb{Z}$; that is, congruence modulo m partitions the set of integers into distinct "congruence classes", and there are exactly m such classes. For example, the class of 0 modulo m is denoted by $\overline{0}$, and consists of those integers congruent to 0 (mod m). It is easy to see that $\overline{0} = \{\ldots, -2m, -m, 0, m, 2m, \ldots\}$. Similarly, $\overline{1} = \{\ldots, 1-2m, 1-m, 1, 1+m, 1+2m, \ldots\}$. Working modulo m, every element in $\mathbb{Z}$ is congruent to exactly one integer in the set $A = \{0, 1, \ldots, m-1\}$, which means that the m congruence classes modulo m may be listed as $\overline{0}, \overline{1}, \ldots, \overline{m-1}$. This particular set A is an example of a "complete set of residues modulo m", but many other subsets of $\mathbb{Z}$ containing exactly m elements may be used instead of A for this purpose. This fact is readily formalized by use of the following definition.

Definition 1.1.17. Given a fixed modulus $m \geq 2$, any set of m distinct integers $B = \{b_1, b_2, \ldots, b_m\}$ with the property that every element in $\mathbb{Z}$ is congruent modulo m to exactly one integer in the set B is called a "*complete set of residues modulo m*".

When working modulo m, the standard choice used is the set A above, but we now give another example of Definition 1.1.17 that will play a helpful role in Section 2.2.

Example 1.1.18. If $m \geq 2$ is a fixed *even* integer, then the set

$$B = \left\{-\frac{m}{2}+1, -\frac{m}{2}+2, \ldots, -\frac{m}{2}+m = \frac{m}{2}\right\}$$

is a complete set of residues modulo m.

For any fixed modulus $m \geq 2$, the congruence classes modulo m respect the operations of addition and multiplication. For example, if $m = 7$, then $\overline{3} + \overline{9} = \overline{5}$. These congruence classes form an "abelian group" of order m under the operation of addition. We will consider other important groups in this thesis that are not abelian and which might contain an infinite number of elements as well. We first give the formal definition of a "group".

Definition 1.1.19. A *group* is a set G with a binary operation $\star$ defined on G such that G is closed under the operation, and such that

(i) $(a \star b) \star c = a \star (b \star c)$ for all $a, b, c \in G$;

(ii) there exists an element $e \in G$, called the *identity* of G, such that for all $a \in G$, we have $e \star a = a \star e = a$;

(iii) for each $a \in G$, there exists an element $a^{-1} \in G$, called the *inverse* of a, such that $a^{-1} \star a = a \star a^{-1} = e$.

If the following additional property holds, then we say that the group is *abelian*:

(iv) We have $a \star b = b \star a$ for all $a, b \in G$.

The *order* of a group G is simply the number of elements contained in the set G. A *subgroup* of a group G is any nonempty subset H of G which forms a group in its own right with respect to the operation $\star$.

The group of congruence classes modulo $m \geq 2$ under addition is usually denoted by $\mathbb{Z}/m\mathbb{Z}$. If a is a nonzero integer such that $\gcd(a, m) = 1$, then we may use Corollary 1.1.6 to prove that for any integer b such that $b \equiv a \pmod{m}$, we have $\gcd(b, m) = 1$ as well. This justifies saying that if $\gcd(a, m) = 1$, then the *class* of a modulo m, $\overline{a}$, is relatively prime to m. It is not difficult to prove that the set of congruence classes modulo m that are relatively prime to m forms an abelian group under the operation of multiplication, usually denoted by $(\mathbb{Z}/m\mathbb{Z})^\times$. The number of elements in this group is designated by $\phi(m)$, which is just the number of integers a in the set $\{1, 2, \ldots, m\}$ for which $\gcd(a, m) = 1$. We will often have occasion to work within the groups $\mathbb{Z}/m\mathbb{Z}$ or $(\mathbb{Z}/m\mathbb{Z})^\times$, usually without formal mention since the context should be clear.

The group denoted by $\mathrm{SL}_2(\mathbb{Z})$, to be defined presently, plays a key role in this thesis, and certain subgroups of $\mathrm{SL}_2(\mathbb{Z})$ will also be of great importance.

Definition 1.1.20. Let $\mathrm{SL}_2(\mathbb{Z})$ denote the set of all 2×2 matrices with determinant 1 and coefficients in $\mathbb{Z}$.

Theorem 1.1.21. *The set* $\mathrm{SL}_2(\mathbb{Z})$ *forms a nonabelian group of infinite order under the operation of matrix multiplication, known as the* ***special linear group****.*

Proof. Let

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

be an arbitrary element of $\mathrm{SL}_2(\mathbb{Z})$. By definition, $\det \mathbf{A} = ad - bc = 1$, and $a, b, c, d \in \mathbb{Z}$. Note that

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

is an element of $\mathrm{SL}_2(\mathbb{Z})$, and $\mathbf{I}$ plays the role of the identity element e in part (ii) of Definition 1.1.19. It is easy to verify that the matrix

$$\mathbf{B} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

is such that $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A} = \mathbf{I}$. Note that $\mathbf{B} \in \mathrm{SL}_2(\mathbb{Z})$, and so part (iii) of Definition 1.1.19 holds. It is well known that matrix multiplication is associative for square matrices of the same size, and thus part (i) of Definition 1.1.19 holds. Finally, from the basic multiplicative property of determinants of square matrices, we have $\det(\mathbf{A} \cdot \mathbf{B}) = \det \mathbf{A} \cdot \det \mathbf{B} = 1 \cdot 1 = 1$, and thus $\mathrm{SL}_2(\mathbb{Z})$ is closed under matrix multiplication. We only need to exhibit a single counter-example of part (iv) of Definition 1.1.19 to prove that $\mathrm{SL}_2(\mathbb{Z})$ is nonabelian. Given the two matrices

$$\mathbf{A} = \begin{pmatrix} 3 & 1 \\ -1 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 5 & 1 \\ -1 & 0 \end{pmatrix}$$

in $\mathrm{SL}_2(\mathbb{Z})$, we note that

$$\mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} 14 & 3 \\ -5 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{B} \cdot \mathbf{A} = \begin{pmatrix} 14 & 5 \\ -3 & -1 \end{pmatrix},$$

in violation of part (iv). To see that there are infinitely many distinct matrices in $\mathrm{SL}_2(\mathbb{Z})$, we note that every matrix of the form

$$\begin{pmatrix} n & 1 \\ -1 & 0 \end{pmatrix}$$

for any $n \in \mathbb{Z}$ is an element of $\mathrm{SL}_2(\mathbb{Z})$. This completes the proof of Theorem 1.1.21. $\square$

We finally introduce a somewhat nonstandard operation on 2×2 matrices that will appear later in this thesis. We will simply make up our own notation here since there is no standard usage.

Definition 1.1.22. Given a 2×2 matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with coefficients in $\mathbb{Z}$, we set

$$\mathbf{A}^w = \begin{pmatrix} d & b \\ c & a \end{pmatrix}.$$

It is easy to see that if $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$, then we also have $\mathbf{A}^w \in \mathrm{SL}_2(\mathbb{Z})$. Proving the following lemma is also a straightforward exercise.

Lemma 1.1.23. *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are* 2×2 *matrices with coefficients in* $\mathbb{Z}$*, then*

$$(\mathbf{A} \cdot \mathbf{B})^w = \mathbf{B}^w \cdot \mathbf{A}^w. \tag{1.1.1}$$

We also have

$$\left(\mathbf{A}^w\right)^w = \mathbf{A}. \tag{1.1.2}$$

1.2. Historical Motivation

In this section, we provide a historical context for the motivation of this thesis. From antiquity, mathematicians and scholars have posed many questions related to the representation of various integers. The Greek Diophantus of Alexandria, in his series of works collectively called *Arithmetica*, challenged the mathematicians of the 3rd Century A.D. to compute solutions to 130 algebraic equations, in both determinate and indeterminate forms. Many of these problems could be reduced

to solving quadratics, and came to be known as "Diophantine equations." While Diophantus did not rule out the possibility of rational number solutions, the modern usage of "Diophantine equations" insists that the solutions be restricted to the integers, and we follow modern usage in this thesis. One of the most familiar Diophantine equations was inspired by the Pythagorean Theorem. When restricted to the integers, the solutions of $a^2 + b^2 = c^2$ are called "Pythagorean triples", for which there are infinitely many, among which are $(a, b, c) = (3, 4, 5)$ and $(5, 12, 13)$.

Yet another Diophantine equation, again with a long history stretching back to antiquity, is the Fermat-Pell Equation, often referred to simply as Pell's Equation. There are many variants of this equation, depending upon the situation and field of study, but we are mainly interested in finding all integer pair solutions (x, y) to the equation $x^2 - Dy^2 = 4$, where D is a positive integer satisfying certain conditions which are spelled out in detail in Assumption 2.2.1.

There are many mathematicians who have contributed to the modern development of the subject of Number Theory. The individual chiefly credited with reigniting a new interest in the subject during The Renaissance, after a long period of abeyance, in so much as to be appropriately called the "Father of Modern Number Theory", is the French mathematician Pierre de Fermat (1601 - 1665). Although Fermat published very few works, much of his personal and professional correspondence has survived, and from these letters, many of Fermat's achievements and advances in Number Theory have come to light. Fermat studied Diophantine equations, and he is credited with proving various theorems, for example, related to the integer solutions of equations of the form $p = x^2 + my^2$, where p is an odd prime and $m \in \mathbb{Z}$. The expression on the right hand side, $x^2 + my^2$, is a special type of "binary quadratic form".

In 1657, Fermat issued a challenge to the prominent British mathematicians of

the era to compute *integer* solutions to Pell's Equation for a fixed discriminant [2]. The standard at the time was to allow rational solutions, but Fermat insisted on integer solutions only. While some British mathematicians were able to answer specific cases of Fermat's challenge successfully, Fermat was not completely satisfied. What Fermat was truly after was a general proof that Pell's Equation always possessed nontrivial solutions, as opposed to a method that would successfully produce such solutions in special cases. William Brouncker (1620 - 1684), a well-known Irish mathematician, was one of the individuals who responded to Fermat's challenge. Brouncker had previously developed the theory of continued fractions and had given a remarkable formula for the number $4/\pi$ in terms of such specialized fractions which was published by the English mathematician John Wallis (1616 - 1703) in his famous book *Arithmetica Infinitorum*. Brouncker found that a variation on his method of continued fractions could produce nontrivial solutions to Pell's Equation in his response to Fermat's challenge, even if he could not offer Fermat general proofs. Continued fractions had earlier been discovered by the Italian mathematician Pietro Cataldi (1548 - 1626), but his use of these specialized fractions was not as sophisticated as Brouncker's, and Brouncker apparently had no knowledge of Cataldi's work in this area. Euler and Lagrange would later expand upon the notation and theory of continued fractions, and would take the subject far beyond where Brouncker had left it.

Continuing the work of Fermat, Leonhard Euler (1707 - 1783) also studied the representation of prime numbers by means of various binary quadratic forms (see Section 2.1 for the definition of such forms). Euler was led by these studies to be the first to formulate and state the Law of Quadratic Reciprocity. After proving certain special cases of this famous Law, he was able to rigorously demonstrate several statements left unproven by Fermat. Euler was the first to seriously take up the many

challenges in Number Theory that Fermat had left to posterity. He spent over 50 years of his life making steady progress towards proving Fermat's statements, at the same time laying the proper foundations that now form the basis of modern texts on Number Theory.

Crucial to this thesis is the work of Italian mathematician Joseph-Louis Lagrange (1736 - 1813), a keen student of Euler's work who would later succeed Euler as the director of the mathematics section at the Prussian Academy of Sciences in Berlin. Lagrange was the first to publish a proof that the Fermat-Pell 1-Equation $x^2 - dy^2 = 1$ always has a nontrivial solution pair (x, y) (meaning that x and y are both positive integers) for any given $d \in \mathbb{Z}^+$ that is not a perfect square (see Section 5.2 for an approach to this Diophantine equation using minus continued fractions). Lagrange was also the first to develop what we now call "reduction theory" in his elaboration of the theory of binary quadratic forms. Reduction theory has been studied intensively since its inception, and it forms the core of this thesis. We focus specifically on *indefinite* binary quadratic forms in this thesis, and in this context there is no universally accepted definition for a "reduced form". The definition we prefer in this thesis is due to Don Zagier (born 1951), and the main reference we use in following his approach is his book [8].

The "classic" positive continued fractions first described by Cataldi and Brouncker, and later used by Euler and Lagrange as well, allow any real number β to be expressed as a cascading infinite fraction

$$\beta = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots}}}.$$

In this thesis, however, we will instead employ the theory of "minus" continued fractions, described in [4] and [8]. Using minus continued fractions, we may express any real number β in the form

$$\beta = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots}}},$$

where $n_0 \in \mathbb{Z}$ and $n_j \in \mathbb{Z}^{\geq 2}$ for $j = 1, 2, 3, \ldots$. We discuss minus continued fractions in greater detail in Section 4.2.

Carl Friedrich Gauss (1777 - 1855), widely considered to be one of the greatest mathematicians of all time, made significant contributions to Number Theory in his work entitled *Disquisitiones Arithmeticae*, published in 1801 when he was 24 years old. In the years preceding the *Disquisitiones*, most of the major theorems and contributions to Number Theory were disjointed, isolated, and full of gaps in their logic and proofs. Gauss took it upon himself to collect these scattered works, refine and fill in the holes in the proofs, and to publish the works under a single title. Alongside the contributions of other mathematicians, Gauss also published his own extensive and oftentimes revolutionary results in the same text. The publication of the *Disquisitiones*, coupled with Gauss' renowned status as a highly respected and authoritative mathematician, ignited interest in Number Theory, and set the stage for the development of the subject as we know it today.

CHAPTER II
INDEFINITE BINARY QUADRATIC FORMS AND REAL QUADRATIC IRRATIONALS

2.1. Discriminants

Definition 2.1.1. An *integral binary quadratic form* $f(x, y)$ is a homogeneous polynomial expression of degree two in two variables: $f(x, y) = ax^2 + bxy + cy^2$, where a, b, and c are fixed integer coefficients, not all equal to zero, and the variables x and y are restricted to taking on only integer values.

For brevity, we often use the term "form" or just say "binary quadratic form" instead of "integral binary quadratic form" since this is the only type of form we consider in this thesis. As an abbreviation, we sometimes denote the form $f(x, y)$ by its ordered list of coefficients $[a, b, c]$, contained within square brackets, or simply by f. The fundamental quantity associated to a form is its *discriminant*.

Definition 2.1.2. The *discriminant* D of a binary quadratic form $[a, b, c]$ is defined to be the integer $D = b^2 - 4ac$.

For example, the form $x^2 + y^2$ has discriminant -4, the form $x^2 + 2xy + y^2$ has discriminant 0, and the form $x^2 + 3xy + y^2$ has discriminant 5.

Given a form $f(x, y)$, the simplest way to classify it initially depends upon the set of integers that it represents. We say that a form $f(x, y)$ "represents the integer n" if there exist integers x_1 and y_1 such that $f(x_1, y_1) = n$. There are some forms that represent only negative integers and 0 and these are of no interest to us. Excluding

these, there are only two types of forms that arise and the following two definitions delineate these two types.

Definition 2.1.3. A form f is said to be *nonnegative* if the only integers it represents are greater than or equal to zero.

It is easy to see that each of the forms $x^2 + y^2$ and $x^2 + 2xy + y^2 = (x + y)^2$ is nonnegative. On the other hand, since $1 \cdot (-1)^2 + 3 \cdot (-1)(1) + 1 \cdot (1)^2 = -1$, the form $x^2 + 3xy + y^2$ is not of this type.

Definition 2.1.4. An *indefinite* form is one which, for suitable values of (x, y), can represent both positive and negative integers.

One of the reasons that the discriminant of a form is so important is that it allows us to instantaneously categorize the form with respect to Definitions 2.1.3 and 2.1.4. If f is a form having negative discriminant, then we know that $a \neq 0$ since otherwise $D = b^2 \geq 0$. If $D < 0$ and $a < 0$, then f represents only negative integers and zero. If $D < 0$ and $a > 0$, then f is a nonnegative form; even stronger, it is "positive definite" (see [5], p. 172). Since so much is already known about forms of negative discriminant, we have decided in this thesis to focus exclusively on forms of nonnegative discriminant. Note that for any discriminant D, we have $D = b^2 - 4ac \equiv b^2 \pmod 4$. Since for any integer b, we have either $b^2 \equiv 0 \pmod 4$ or $b^2 \equiv 1 \pmod 4$, we see that in general we have either $D \equiv 0 \pmod 4$ or $D \equiv 1 \pmod 4$. This shows that the list of all possible nonnegative discriminants, in ascending order, starts out as follows: $0, 1, 4, 5, 8, 9, 12, 13, \ldots$. Every integer on this list arises as the discriminant of some binary quadratic form. If $D \equiv 0 \pmod 4$, then the form

$$x^2 - \frac{D}{4}y^2$$

has discriminant D. If $D \equiv 1 \pmod 4$, then the form

$$x^2 + xy + \frac{1-D}{4}y^2$$

has discriminant D. Even if all of the discriminants in the list above are possible, not all of them are equally interesting. If $D = 0$, or if D is a perfect square (see Definition 1.1.12), then any form $f(x, y)$ having such a value of D as its discriminant can be factored into a product of linear forms as follows ([5], p. 171):

$$f(x, y) = (rx + sy)(tx + uy), \tag{2.1.1}$$

where r, s, t, and u are all integers. Conversely, if $D \in \mathbb{Z}^+$ is a discriminant that is not a perfect square, then *no* form of that discriminant may be factored into a product of linear forms. If a binary quadratic form can be factored as in (2.1.1), then it becomes significantly easier to work with and loses its second degree quality. For all of the reasons just given, we henceforth only consider in this thesis binary quadratic forms having a positive integer discriminant D which is not a perfect square. Such forms are *always* indefinite, as we presently show. It is worth noting that since we only consider those $D \in \mathbb{Z}^+$ such that D is not a perfect square, any form $f = [a, b, c]$ of such discriminant must have $a \neq 0$ and $c \neq 0$ since otherwise $D = b^2$. The fact that we always have $a \neq 0$ and $c \neq 0$ for every form $[a, b, c]$ under consideration from this point onwards will prove to be very important on many occasions. We will give regular reminders of this fact, but there shall be instances where this is tacitly assumed to be known and where no explicit mention of this fact will be made.

Theorem 2.1.5. *Every binary quadratic form $f = [a, b, c]$ having a discriminant $D = b^2 - 4ac > 0$ that is not a perfect square is necessarily indefinite.*

Proof. We first note that $f(1,0) = a$, and that

$$f(b, -2a) = a \cdot b^2 - b \cdot 2ba + c \cdot 4a^2 = a\left(4ac - b^2\right) = -Da.$$

Since $a \neq 0$ and $D > 0$, one of the two integers just computed will always be positive and the other one will be negative. By Definition 2.1.4, our proof is complete. □

Remark. As we saw above, if $f = [a, b, c]$ is a form of discriminant D, then $b^2 \equiv D \pmod 4$. It is well-known that if $b \in \mathbb{Z}$ is even, then b^2 is even, and if b is odd, then b^2 is odd. Thus, if b is even, then $D \equiv 0 \pmod 4$, which means that D is even, and we have $b \equiv D \pmod 2$. Likewise, if b is odd, then $D \equiv 1 \pmod 4$, which means that D is odd, and we have $b \equiv D \pmod 2$. In general, we conclude that if $f = [a, b, c]$ is a form of discriminant D, then $b \equiv D \pmod 2$. In other words, the middle coefficient b and the discriminant D always have the same parity.

2.2. Classes of Forms

Based upon the considerations in Section 2.1, we restrict ourselves to the study of only those binary quadratic forms having a discriminant satisfying the following conditions.

Assumption 2.2.1. A *discriminant* (throughout the remainder of this thesis) is a positive integer $D \in \mathbb{Z}^+$ that is not a perfect square, and for which we have either $D \equiv 0 \pmod 4$ or $D \equiv 1 \pmod 4$.

The list of such D values starts in ascending order as follows:

$$5, 8, 12, 13, 17, 20, 21, 24, 28, 29, 32, 33, 37, 40, 41, 44, 45, 48, 52, 53, 56, 57, 60, 61, 65, \ldots.$$

By Theorem 2.1.5, all of the binary quadratic forms that we consider from this point onwards are indefinite forms. For a fixed discriminant $D \in \mathbb{Z}^+$, we show later

in this section that there are infinitely many distinct binary quadratic forms whose discriminant is equal to D. There is one particular form of discriminant $D \in \mathbb{Z}^+$ which has a special name because of its great importance.

Definition 2.2.2. If $D \equiv 0 \pmod 4$, then the indefinite form

$$x^2 - \frac{D}{4}y^2 \tag{2.2.1}$$

is called the *principal form of discriminant* D. If $D \equiv 1 \pmod 4$, then the indefinite form

$$x^2 + xy + \frac{1-D}{4}y^2 \tag{2.2.2}$$

is similarly called the *principal form of discriminant* D.

Definition 2.2.3. For a fixed discriminant $D \in \mathbb{Z}^+$, we let $Q(D)$ denote the set of all binary quadratic forms of discriminant D. By Definition 2.2.2, we know that $Q(D)$ is a nonempty set.

Given a form $f = [a, b, c]$ of discriminant D, it is of interest to know exactly which integers f represents. Questions of this type date all the way back to the work of Fermat during the 1630's. A famous result of Fermat, for example, states that *no* prime number that is congruent to 3 modulo 4 (it is not difficult to prove that the list of such primes is infinitely long, starting with $3, 7, 11, 19, \dots$) is representable by the form $x^2 + y^2$. On the other hand, *every* odd prime number $p \equiv 1 \pmod 4$ (again, there are infinitely many such primes starting with $5, 13, 17, 29, \dots$) *is* representable by the form $x^2 + y^2$. The form $x^2 + y^2$ is admittedly positive definite, but similar restrictions hold with respect to indefinite forms as well. As the originators of this subject discovered, it often happens that two distinct forms f and g represent exactly the same integers, whereas one of these two forms could be much easier to work

with because it might have significantly smaller coefficients than the other. These observations have led to an extensive theory where a given form f is transformed to another form g which represents the exact same integers as f. We only consider transformations of a very special type, commonly known as being "unimodular". Given a form $f(x, y) = ax^2 + bxy + cy^2$, we may replace the variable x by $rX + sY$ and the variable y by $tX + uY$ to obtain by substitution the following:

$$\begin{aligned} f(x, y) &= f(rX + sY, tX + uY) \\ &= a(rX + sY)^2 + b(rX + sY)(tX + uY) + c(tX + uY)^2 \\ &= a_1X^2 + b_1XY + c_1Y^2 = g(X, Y), \end{aligned} \tag{2.2.3}$$

where the coefficients a_1, b_1, and c_1 are given below by the expressions in (2.2.7), (2.2.8), and (2.2.9), respectively.

Definition 2.2.4. The binary quadratic form $f = [a, b, c]$ is said to be *equivalent* to the form $g = [a_1, b_1, c_1]$ if there exist four integers r, s, t, and u, for which $ru - st = 1$, and such that the substitutions

$$x = rX + sY, \quad y = tX + uY \tag{2.2.4}$$

transform $f(x, y)$ into $g(X, Y)$ as in (2.2.3) above. The equations in (2.2.4) give a *unimodular* transformation from the form f to the form g which may be represented by the transformation matrix

$$\begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}).$$

In this case, we use the shorthand notation $f \sim g$.

We now show that the discriminant of a form f is preserved under unimodular transformations, which is another way of saying that equivalent forms always have the same discriminant.

Theorem 2.2.5. *If the form f has discriminant D and $f \sim g$, then g also has discriminant D.*

Proof. Let $f = [a, b, c] = ax^2 + bxy + cy^2$ be such that $b^2 - 4ac = D$. Since $f \sim g$, there exist four integers r, s, t, and u, with $ru - st = 1$, such that the transformation in (2.2.4) allows us to express g in the form

$$g(X, Y) = a(rX + sY)^2 + b(rX + sY)(tX + uY) + c(tX + uY)^2. \tag{2.2.5}$$

It is easy to verify that, after expansion and collection of terms, we may rewrite (2.2.5) as

$$g(X, Y) = a_1X^2 + b_1XY + c_1Y^2, \tag{2.2.6}$$

where

$$a_1 = ar^2 + brt + ct^2, \tag{2.2.7}$$

$$b_1 = 2ars + b(ru + st) + 2ctu, \quad \text{and} \tag{2.2.8}$$

$$c_1 = as^2 + bsu + cu^2. \tag{2.2.9}$$

A straightforward, if somewhat tedious, calculation reveals that we have $b_1^2 - 4a_1c_1 = b^2 - 4ac = D$, and thus $g = [a_1, b_1, c_1]$ also has discriminant D. □

Theorem 2.2.6. *The relationship $\sim$ given between forms in Definition 2.2.4 is reflexive, symmetric, and transitive, and therefore establishes an equivalence relation with respect to all forms of the same discriminant D.*

Before we delve into the proof of Theorem 2.2.6, it is helpful to first introduce some convenient notation. The two linear equations in (2.2.4) may be rewritten in matrix form as

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}. \tag{2.2.10}$$

Assume that a second unimodular transformation from the variables X, Y to the variables x', y' is defined by

$$X = r_1 x' + s_1 y', \quad Y = t_1 x' + u_1 y', \tag{2.2.11}$$

which in matrix form looks like:

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} r_1 & s_1 \\ t_1 & u_1 \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}. \tag{2.2.12}$$

Whether we use (2.2.11) or (2.2.12), the form $g(X, Y)$ is taken to a new form $h(x', y')$, which also has discriminant D by Theorem 2.2.5. Switching from the variables x, y to X, Y, and then from X, Y to x', y', is carried out by plugging the equations (2.2.11) into (2.2.4) to obtain

$$x = r(r_1 x' + s_1 y') + s(t_1 x' + u_1 y') = (rr_1 + st_1)\, x' + (rs_1 + su_1)\, y',$$

$$y = t(r_1 x' + s_1 y') + u(t_1 x' + u_1 y') = (tr_1 + ut_1)\, x' + (ts_1 + uu_1)\, y'.$$

Equivalently, we may substitute (2.2.12) into (2.2.10) and apply the associative law of

matrix multiplication to obtain

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \cdot \left[\begin{pmatrix} r_1 & s_1 \\ t_1 & u_1 \end{pmatrix} \cdot \begin{pmatrix} x' \\ y' \end{pmatrix} \right] = \left[\begin{pmatrix} r & s \\ t & u \end{pmatrix} \cdot \begin{pmatrix} r_1 & s_1 \\ t_1 & u_1 \end{pmatrix} \right] \cdot \begin{pmatrix} x' \\ y' \end{pmatrix}$$
$$= \begin{pmatrix} rr_1 + st_1 & rs_1 + su_1 \\ tr_1 + ut_1 & ts_1 + uu_1 \end{pmatrix} \cdot \begin{pmatrix} x' \\ y' \end{pmatrix}. \tag{2.2.13}$$

Notation. If the form f is taken to the form g by use of the equations in (2.2.4) and

$$\mathbf{A} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

then $f \sim g$, and we employ the following shorthand notation:

$$f \cdot \mathbf{A} = g. \tag{2.2.14}$$

For example, referring back to (2.2.12), if

$$\mathbf{B} = \begin{pmatrix} r_1 & s_1 \\ t_1 & u_1 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

then $g \sim h$, and we also have

$$g \cdot \mathbf{B} = h. \tag{2.2.15}$$

Proof of Theorem 2.2.6. The fact that $\mathrm{SL}_2(\mathbb{Z})$ is a group, proved as Theorem 1.1.21, is critical to the present proof. Assume that $f \sim g$ with $f \cdot \mathbf{A} = g$, and that $g \sim h$ with $g \cdot \mathbf{B} = h$. By (2.2.13), we see that

$$f \cdot (\mathbf{A} \cdot \mathbf{B}) = h. \tag{2.2.16}$$

Since $\mathbf{A}$ and $\mathbf{B}$ are both elements in the *group* $\mathrm{SL}_2(\mathbb{Z})$, we have $(\mathbf{A} \cdot \mathbf{B}) \in \mathrm{SL}_2(\mathbb{Z})$,

and (2.2.16) thus states that $f \sim h$, which establishes transitivity. If we replace g in (2.2.15) by use of (2.2.14), we are able to append two more equalities to (2.2.16) to obtain

$$f \cdot (\mathbf{A} \cdot \mathbf{B}) = h = g \cdot \mathbf{B} = (f \cdot \mathbf{A}) \cdot \mathbf{B}, \tag{2.2.17}$$

and we may therefore conclude that

$$(f \cdot \mathbf{A}) \cdot \mathbf{B} = f \cdot (\mathbf{A} \cdot \mathbf{B}). \tag{2.2.18}$$

The order in which matrices are multiplied is crucial, and the fact that (2.2.18) holds means that we have a "right group action" (see [1], §1.7) of the group $\mathrm{SL}_2(\mathbb{Z})$ on the set $Q(D)$. Note that $f \sim f$ since f is transformed into itself by the identity transformation

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

which establishes reflexivity. The symmetric property is obtained using the following steps. Assuming that $f \sim g$ just means that

$$f \cdot \mathbf{A} = g \tag{2.2.19}$$

for some matrix $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$. Since $\mathrm{SL}_2(\mathbb{Z})$ is a *group*, $\mathbf{A}^{-1} \in \mathrm{SL}_2(\mathbb{Z})$, and application of $\mathbf{A}^{-1}$ coming in from the right hand side in (2.2.19) gives

$$(f \cdot \mathbf{A}) \cdot \mathbf{A}^{-1} = g \cdot \mathbf{A}^{-1}. \tag{2.2.20}$$

By (2.2.18), the left hand side of (2.2.20) may be re-expressed as

$$(f \cdot \mathbf{A}) \cdot \mathbf{A}^{-1} = f \cdot \left(\mathbf{A} \cdot \mathbf{A}^{-1}\right) = f \cdot \mathbf{I} = f, \tag{2.2.21}$$

and combining this with (2.2.20) we conclude that $g \cdot \mathbf{A}^{-1} = f$, which implies that $g \sim f$. $\square$

The equivalence relation established in Theorem 2.2.6 allows us to partition the forms in $Q(D)$ into distinct *classes of forms*. There will always be at least one class of forms of discriminant $D \in \mathbb{Z}^+$, namely the class to which the principal form of discriminant D belongs.

Definition 2.2.7. Given a fixed discriminant $D \in \mathbb{Z}^+$, the set of all forms in $Q(D)$ that are equivalent to the principal form of discriminant D makes up the *principal class of discriminant* D.

For some discriminants, there is only one class of forms, which just means that *every* form of that discriminant is equivalent to the principal form. Examples of such "one class" discriminants are $D = 5, 8, 13, 17, 29, 37, 41, 53$, and 61, as we verify in Section 3.1.

Given a fixed discriminant $D \in \mathbb{Z}^+$, we mentioned earlier in this section that there are infinitely many distinct forms in $Q(D)$. This is an immediate consequence of the following more refined result.

Theorem 2.2.8. *For a fixed discriminant $D \in \mathbb{Z}^+$, every class of forms within $Q(D)$ contains infinitely many distinct individual forms.*

Proof. Let $\mathcal{C}$ denote a fixed class of forms within $Q(D)$ and let $f_1 = [a_1, b_1, c_1]$ be an arbitrarily given member of $\mathcal{C}$. Given any fixed integer $s \in \mathbb{Z}$, the matrix

$$\mathbf{E}(s) = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \tag{2.2.22}$$

is clearly an element of $\mathrm{SL}_2(\mathbb{Z})$, and we have $f_1 \cdot \mathbf{E}(s) = f_2 = [a_2, b_2, c_2]$, where $a_2 = a_1$ by (2.2.7), and $b_2 = b_1 + s \cdot (2a_1)$ by (2.2.8). Since $a_1 \neq 0$, infinitely many distinct values will arise for the middle coefficient b_2 as s runs through all integers in $\mathbb{Z}$. $\square$

We now show that if two forms f and g are in the same class of forms $\mathcal{C} \subseteq Q(D)$, then they represent exactly the same integers. This shows that just having our hands on *one* form in a given class is sufficient with respect to questions regarding the representation of various subsets of integers.

Theorem 2.2.9. *Equivalent forms represent exactly the same integers.*

Proof. If $f \sim g$, then by definition there is a matrix

$$\mathbf{A} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

such that $f \cdot \mathbf{A} = g$. If the integer $k_1 \in \mathbb{Z}$ is represented by the form g, then there exist integers X_1 and Y_1 such that $g(X_1, Y_1) = k_1$. Setting $x_1 = rX_1 + sY_1$ and $y_1 = tX_1 + uY_1$, we see by (2.2.3) that we have $f(x_1, y_1) = k_1$, and so k_1 is represented by the form f as well. Conversely, assume that the integer $l_2 \in \mathbb{Z}$ is represented by the form f, which means there exist integers x_2 and y_2 such that $f(x_2, y_2) = l_2$. Since $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$, the linear equations in (2.2.4) are invertible, which implies that there exist uniquely determined integers X_2 and Y_2 such that $x_2 = rX_2 + sY_2$ and $y_2 = tX_2 + uY_2$. Again, by (2.2.3), we have $f(x_2, y_2) = g(X_2, Y_2) = l_2$, which shows that the form g represents l_2 as well. $\square$

It was Lagrange who first introduced the notion of classes of binary quadratic forms of a fixed discriminant. It was also he who first proved the remarkable theorem that for any given discriminant $D \in \mathbb{Z}^+$, the number of classes of forms that the

equivalence relation in Theorem 2.2.6 sets up is always *finite*. We give two independent proofs of this famous theorem, one in this section and one in Section 3.1. The proof in this section follows the one that is given starting on page 175 of Landau's book [5]. Landau's presentation is very concise and we aim to give a more leisurely and detailed presentation here. The following lemma is the key to proving Theorem 2.2.11 below.

Lemma 2.2.10. *Every class of forms $\mathcal{C} \subseteq Q(D)$ contains an individual binary quadratic form $f = [a, b, c]$ for which*

$$|b| \leq |a| \leq |c| \,. \tag{2.2.23}$$

Proof. Let $f_0(x, y) = a_0x^2 + b_0xy + c_0y^2$ be an arbitrarily given fixed form of discriminant $D \in \mathbb{Z}^+$ lying in the class $\mathcal{C}$. Recall from Theorem 2.1.5 that f_0 represents both positive and negative integers. Let V denote the nonempty set of all *nonzero* integers that are represented by f_0, and taking absolute values of these nonzero integers allows us to form the following set:

$$T = \left\{ |n| \in \mathbb{Z}^+ : n \in V \right\}.$$

Since T is a nonempty subset of $\mathbb{Z}^+$, the Well-Ordering Principle guarantees the existence of a least element in T, namely, there is an integer $a \in \mathbb{Z} \setminus \{0\}$ such that $f_0(r, t) = a$ for a pair of integers r and t, and $|a| \leq |n|$ for every $n \in V$. The integers r and t are not uniquely determined, so we just fix one choice for this pair such that

$$a = a_0r^2 + b_0rt + c_0t^2. \tag{2.2.24}$$

Since $a \neq 0$ by construction, the integers r and t in (2.2.24) can not both be equal to zero. We claim that $\gcd(r, t) = 1$. To demonstrate this, suppose for the sake of

contradiction that we have $\gcd(r,t) = d > 1$. Since d is a divisor of both r and t, there exist integers k and ℓ such that $dk = r$ and $d\ell = t$. Thus, equation (2.2.24) may be rewritten as follows:

$$\begin{aligned} a &= a_0(dk)^2 + b_0(dk)(d\ell) + c_0(d\ell)^2 \\ &= a_0 d^2 k^2 + b_0 d^2 k\ell + c_0 d^2 \ell^2 \\ &= d^2\left[a_0 k^2 + b_0 k\ell + c_0 \ell^2\right]. \end{aligned} \tag{2.2.25}$$

By assumption, we have $1 < d$, and therefore $1 < d^2$. Since $0 < |a|$, we have $|a| < |a|d^2$, and so

$$\frac{|a|}{d^2} < |a|. \tag{2.2.26}$$

By (2.2.25), we note that a/d^2 is an integer (it is nonzero as well) and this integer is represented by f_0 since $f_0(k,\ell) = a/d^2$ by (2.2.25). We conclude that $a/d^2 \in V$ and therefore

$$\left|\frac{a}{d^2}\right| = \frac{|a|}{d^2} \in T. \tag{2.2.27}$$

On the other hand, $|a|$ is the least element in T, and (2.2.26) and (2.2.27) stand in contradiction to $|a|$ being this least element. This establishes our claim that $\gcd(r,t) = 1$. Since the integers r and t are relatively prime, Corollary 1.1.6 guarantees the existence of two integers s and u such that $ru - st = 1$, and this gives us in turn a matrix

$$\mathbf{A} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}).$$

We now set $f_1 = [a_1, b_1, c_1] = f_0 \cdot \mathbf{A}$, and note that $f_0 \sim f_1$, and

$$a_1 = a = a_0 r^2 + b_0 rt + c_0 t^2$$

by (2.2.7) and (2.2.24). The "a–coefficient" we are seeking for the inequalities in (2.2.23) has now been found. We still need one further transformation to obtain the sought-after b–coefficient. Recall from the proof of Theorem 2.2.8 that the transformation matrix $\mathbf{E}(s)$ [see (2.2.22)] leaves the a–coefficient unchanged so that $a_2 = a_1$. On the other hand, the new b–coefficient is given by

$$b_2 = b_1 + s \cdot (2a), \tag{2.2.28}$$

since $a_1 = a$. We set $f_2 = [a_2, b_2, c_2] = f_1 \cdot \mathbf{E}(s)$, and our claim is that an integer s may be chosen in (2.2.28) such that $|b_2| \leq |a|$. Note that (2.2.28) tells us that all of the potential b–coefficients are separated from each other by a distance measured in multiples of $|2a|$. Let $m = |2a|$, which is an even integer greater than or equal to 2. Working modulo m, we may choose $s \in \mathbb{Z}$ such that the integer b_2 in (2.2.28) falls into the set B defined in Example 1.1.18:

$$B = \{-|a| + 1, -|a| + 2, \ldots, -|a| + |2a| = |a|\}.$$

This implies that $-|a| < b_2 \leq |a|$, which verifies our claim. A more geometric way to visualize this is to note that there are $m + 1$ integers that are greater than or equal to $-|a|$ and less than or equal to $|a|$. In some cases, two distinct values of s in (2.2.28) will lead to $|b_2| \leq |a|$. For example, if $a = -4$ and $b_1 = -20$, then $s = -2$ leads to $b_2 = -4$, and $s = -3$ leads to $b_2 = 4$. This is illustrated by the red dots in Figure 2.1 below, which are each separated by a distance of $8 = |2a|$. On the other hand, if $b_1 = -19$ instead, then only the value $s = -2$ gives an answer of $b_2 = -3$ in the proper range. This is illustrated by the blue dots in Figure 2.1. We have removed certain portions of the real line in Figure 2.1 to preserve the scale.

Figure 2.1. Picture proof of the existence of a suitable integer s

We now have $f_2 = [a, b, c_2] = f_1 \cdot \mathbf{E}(s)$, with $|b| \leq |a|$ for an appropriate choice of $s \in \mathbb{Z}$. The b–coefficient we are seeking for the inequalities in (2.2.23) has now been found. We have $f_1 \sim f_2$, and since $f_0 \sim f_1$, we have $f_2 \in \mathcal{C}$ by transitivity. We now claim that $c = c_2$ satisfies $|a| \leq |c|$ automatically because of the way in which a was chosen, and thus the form $f = f_2 = [a, b, c]$ has coefficients that satisfy (2.2.23), which then completes the proof of Lemma 2.2.10. To verify this last claim, we recall from Theorem 2.2.9 that the two forms f_0 and f represent exactly the same integers since they both lie in the same class $\mathcal{C}$. Since $f(0, 1) = c$, the form f represents c and therefore f_0 represents c as well. Because $c \neq 0$, we have $c \in V$, and so $|a| \leq |c|$ by our minimality choice of a. □

The proof just given of Lemma 2.2.10 hinges upon the existence of a smallest nonzero integer a (in terms of absolute value), representable by the form f_0. From an algorithmic standpoint, there is not a straightforward process to find the integer a when working with an *indefinite* form f_0. Therefore, even if the proof given of Lemma 2.2.10 is completely rigorous, it is not completely satisfactory from an algorithmic point of view. The presentation of algorithmically satisfactory proofs and methods for indefinite forms is the theme of Section 3.1.

Theorem 2.2.11. *For a fixed discriminant $D \in \mathbb{Z}^+$, the number of equivalence classes of forms of discriminant D is finite.*

Proof. We consider the set $S(D)$ of all forms $f = [a, b, c]$ of discriminant D that satisfy the conditions in (2.2.23): $|b| \leq |a| \leq |c|$. If $D \equiv 0 \pmod 4$, then the principal

form $[1, 0, -D/4]$ satisfies these conditions; if $D \equiv 1 \pmod 4$, then the principal form $[1, 1, (1-D)/4]$ also satisfies these conditions, which shows that the set $S(D)$ is nonempty for every choice of $D \in \mathbb{Z}^+$ satisfying Assumption 2.2.1. We remind the reader that for *any* form having such a discriminant, both the a–coefficient and the c–coefficient are necessarily nonzero.

Our goal is to show that the set $S(D)$ has only *finitely* many elements. Lemma 2.2.10 tells us that every form in $Q(D)$ is equivalent to a form in $S(D)$, and once it is known that $S(D)$ is a finite set, it follows immediately from Lemma 2.2.10 that there are only a finite number of classes of forms of discriminant D.

In order to show that $S(D)$ is a finite set, we need to establish several inequalities that follow from (2.2.23). Let $f = [a, b, c]$ be any given form in the set $S(D)$. From $|b| \leq |a|$ and $|b| \leq |c|$, we conclude that $b^2 = |b|^2 \leq |a||c| = |ac|$. Since $b^2 = D + 4ac$ and $0 < D$, we find that

$$4ac < D + 4ac = b^2 \leq |ac|. \tag{2.2.29}$$

The inequalities in (2.2.29) imply that $ac < 0$. To see this, we first note that $ac \neq 0$. If we had $0 < ac$, it would follow that $|ac| = ac < 4ac$, in contradiction to (2.2.29). We conclude that $ac < 0$, and so

$$-ac = |ac|. \tag{2.2.30}$$

From $|a| \leq |c|$, we see that $a^2 = |a|^2 \leq |a||c| = |ac|$. Combining with (2.2.30), we obtain

$$4a^2 \leq 4|ac| = -4ac = D - b^2 \leq D. \tag{2.2.31}$$

From (2.2.31), we conclude that $0 < |a|^2 \leq D/4$, and taking square roots leads to the

inequality

$$|a| \leq \frac{\sqrt{D}}{2} \tag{2.2.32}$$

for the a–coefficient of f. Since $|b| \leq |a|$, we also obtain

$$|b| \leq \frac{\sqrt{D}}{2} \tag{2.2.33}$$

for the b–coefficient of f. Of course, $\sqrt{D}/2$ is a fixed positive constant and there are only finitely many pairs of integer-valued choices for a and b that simultaneously satisfy (2.2.32) and (2.2.33). For each such appropriate pair of integers a and b, the c–coefficient of f is uniquely determined by the equation

$$c = \frac{b^2 - D}{4a}. \tag{2.2.34}$$

We conclude that $S(D)$ is a finite set. Examples of the explicit construction of all forms in $S(D)$ for various values of $D \in \mathbb{Z}^+$ are given below in Example 2.2.13. $\square$

Given Theorem 2.2.11, we are now in a position to define the following crucial invariant associated to any given discriminant $D \in \mathbb{Z}^+$.

Definition 2.2.12. For a fixed discriminant $D \in \mathbb{Z}^+$, we let the positive integer $t(D)$ denote the number of equivalence classes of forms of discriminant D.

There is a straightforward algorithm to find all forms in the set $S(D)$ for any given discriminant $D \in \mathbb{Z}^+$. We illustrate this algorithm for a few small values of $D \in \mathbb{Z}^+$ in Example 2.2.13 below. Knowing how many forms are in the set $S(D)$ gives us immediately an upper bound on the size of the invariant $t(D)$. An exact determination of $t(D)$, however, requires a more sophisticated method, and such an algorithmically effective method giving the precise determination of the invariant $t(D)$ is presented in Section 3.1.

Example 2.2.13. We give here several examples of how to find all forms in the set $S(D)$ for a few small values of $D \in \mathbb{Z}^+$. If $D = 5$, then $\sqrt{D}/2 = 1.118\ldots$, and we have either $a = 1$ or $a = -1$ by (2.2.32), since $a \neq 0$. Similarly, the only choices for the b–coefficient are -1, 0, and 1 by (2.2.33), but $b = 0$ is ruled out since the b–coefficient and D must have the same parity by the Remark at the end of Section 2.1. Using (2.2.34), if $a = 1$ and $b = 1$, then $c = -1$, so that $[1, 1, -1] \in S(5)$. Continuing in this way, we find that $S(5)$ consists precisely of the four forms $[1, 1, -1], [1, -1, -1], [-1, 1, 1]$, and $[-1, -1, 1]$. By Lemma 2.2.10, we have $t(5) \leq 4$, and we have $t(5) = 4$ only if all four forms in $S(5)$ lie in different equivalence classes. In Section 3.1, we will find that all four forms in $S(5)$ lie in the same class (all four are in the principal class!), so that $t(5) = 1$. If $D = 8$, then $\sqrt{D}/2 = 1.414\ldots$, and again we must have either $a = 1$ or $a = -1$. This time both $b = 1$ and $b = -1$ are ruled out since their parity does not match that of D, so only $b = 0$ is allowed. We find that $S(8)$ consists of only the two forms $[1, 0, -2]$ and $[-1, 0, 2]$, and so $t(8) \leq 2$. We will find in Section 3.1 that $t(8) = 1$. A similar analysis shows that $S(12)$ consists of only the two forms $[1, 0, -3]$ and $[-1, 0, 3]$, but in this case we will find in Section 3.1 that $t(12) = 2$. If $D = 17$, then $\sqrt{D}/2 = 2.061\ldots$, and a must be chosen among the four possibilities: $2, 1, -1$, and -2. By parity considerations, we are only allowed the two b–values of 1 and -1. Therefore, $S(17)$ consists of the following eight forms: $[2, 1, -2], [2, -1, -2], [1, 1, -4], [1, -1, -4], [-1, 1, 4], [-1, -1, 4], [-2, 1, 2], [-2, -1, 2]$, and so $t(17) \leq 8$. We will find in Section 3.1 that $t(17) = 1$, and this already begins to show that the number of forms in the set $S(D)$ only gives a fairly crude upper bound on the size of $t(D)$.

Given a fixed discriminant $D \in \mathbb{Z}^+$, there is an important distinction to be

made among the forms in the set $Q(D)$, which is formalized in the following definition. We will see the relevance of this distinction in Section 2.3.

Definition 2.2.14. A form $[a, b, c] \in Q(D)$ is said to be *primitive* if $\gcd(a, b, c) = 1$, and it is said to be *imprimitive* if $\gcd(a, b, c) > 1$.

For example, the form $[2, 9, 5]$ of discriminant 41 is primitive, whereas the form $[3, 12, 6]$ of discriminant 72 is imprimitive since $\gcd(3, 12, 6) = 3$. Note that the principal form of discriminant D is always primitive since the a–coefficient is equal to 1 by definition, which forces the value of $\gcd(a, b, c)$ to be 1. This shows that we always have at least one form in $Q(D)$ that is primitive. Primitive forms have certain properties that make them more desirable to work with than imprimitive forms, but we treat all forms in $Q(D)$ on an equal footing whenever we can.

The following theorem shows that any two given equivalent forms in $Q(D)$ are either both primitive or they are both imprimitive. This implies that we can designate each *class* $\mathcal{C}$ of forms in $Q(D)$ as being either primitive or imprimitive. Since the principal form of discriminant D is primitive, we note that the principal *class* of discriminant D is primitive as well, which shows that at least one class of forms of discriminant D is primitive.

Theorem 2.2.15. *If $f = [a, b, c]$ and $g = [a_1, b_1, c_1]$ are any two given forms in $Q(D)$ with $f \sim g$, then $\gcd(a, b, c) = \gcd(a_1, b_1, c_1)$.*

Proof. First, we show that $\gcd(a, b, c) \leq \gcd(a_1, b_1, c_1)$. Set $d = \gcd(a, b, c)$ and $e = \gcd(a_1, b_1, c_1)$. By definition, d is a positive integer and $d \mid a$, $d \mid b$, and $d \mid c$, so there exist integers $k, \ell, m \in \mathbb{Z}$ such that $a = dk$, $b = d\ell$, and $c = dm$. Since $f \sim g$ by assumption, there exist four integers r, s, t, and u, with $ru - st = 1$, such that the coefficients of the form g are given in terms of the coefficients of the form f by the

three equations (2.2.7), (2.2.8), and (2.2.9). If we replace a, b, and c in these three equations by the expressions above, we obtain

$$\begin{aligned}
a_1 &= ar^2 + brt + ct^2 = dkr^2 + d\ell rt + dmt^2 = d\left(kr^2 + \ell rt + mt^2\right), \\
b_1 &= 2ars + b(ru + st) + 2ctu = 2dkrs + (d\ell)(ru + st) + 2dmtu \\
&= d\left(2krs + \ell(ru + st) + 2mtu\right), \quad \text{and} \\
c_1 &= as^2 + bsu + cu^2 = dks^2 + d\ell su + dmu^2 = d\left(ks^2 + \ell su + mu^2\right).
\end{aligned}$$

Thus, we have that $d \mid a_1$, $d \mid b_1$, and $d \mid c_1$, which implies that the positive integer d is a common divisor of a_1, b_1, and c_1. By definition, e is the largest positive integer that simultaneously divides a_1, b_1, and c_1, and thus $d \leq e$.

In order to complete the proof, we now show that $e = \gcd(a_1, b_1, c_1) \leq d$. By definition, e is a positive integer and $e \mid a_1$, $e \mid b_1$, and $e \mid c_1$, so there exist integers $h, i, j \in \mathbb{Z}$ such that $a_1 = eh$, $b_1 = ei$, and $c_1 = ej$. We are assuming that $f \cdot \mathbf{A} = g$, where

$$\mathbf{A} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}).$$

At the end of the proof of Theorem 2.2.6, we found that $g \cdot \mathbf{A}^{-1} = f$. If we set

$$\mathbf{A}^{-1} = \begin{pmatrix} r_1 & s_1 \\ t_1 & u_1 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

then as we saw in the proof of Theorem 1.1.21, we have $r_1 = u$, $s_1 = -s$, $t_1 = -t$, and

$u_1 = r$. The equations analogous to (2.2.7), (2.2.8), and (2.2.9) are

$$\begin{aligned} a &= a_1 r_1^2 + b_1 r_1 t_1 + c_1 t_1^2, \\ b &= 2a_1 r_1 s_1 + b_1(r_1 u_1 + s_1 t_1) + 2c_1 t_1 u_1, \quad \text{and} \\ c &= a_1 s_1^2 + b_1 s_1 u_1 + c_1 u_1^2. \end{aligned}$$

Using the same argument as above, we see that $e \mid a$, $e \mid b$, and $e \mid c$, which implies that the positive integer e is a common divisor of a, b, and c. By definition, d is the largest positive integer that simultaneously divides a, b, and c, and thus $e \leq d$, which in conjunction with the first half of the proof establishes the equality $d = e$, completing the proof of Theorem 2.2.15. □

Now that we know that the forms in a given class of forms of discriminant $D \in \mathbb{Z}^+$ are either all primitive or all imprimitive, the following definition makes perfect sense.

Definition 2.2.16. For a fixed discriminant $D \in \mathbb{Z}^+$, we let the positive integer $h(D)$ denote the number of equivalence classes of forms of discriminant D which contain only primitive forms.

There are only a total number of $t(D) \in \mathbb{Z}^+$ classes of forms of discriminant D, so it is clear that $h(D) \leq t(D)$. We noted above that the principal class of discriminant D contains only primitive forms, and so $1 \leq h(D)$. A given discriminant $D \in \mathbb{Z}^+$ is said to be *fundamental* if $h(D) = t(D)$. We listed all of the 25 positive discriminants from 5 to 65 inclusive at the beginning of this section, and an examination of Table 3.1.3 at the end of Section 3.1 shows that all of these discriminants are fundamental except

for $D = 20, 32, 45, 48$, and 52. Landau states (see [5], p. 179) that

$$t(D) = \sum h\left(\frac{D}{g^2}\right),$$

where the sum runs over all $g \in \mathbb{Z}^+$ such that $g^2 \mid D$, and with the quantity D/g^2 being itself a discriminant. This shows that if we know $h(D)$ for each positive discriminant $D \in \mathbb{Z}^+$, then we can easily recover the value of $t(D)$ for all such discriminants as well. Given this, a generally usable formula for $h(D)$ would be of prime importance! In Section 2.5, we present a famous analytic formula for $h(D)$ that was originally derived by Peter Gustav Dirichlet in 1839.

2.3. Automorphs and the Fermat-Pell 4-Equation

Let $D \in \mathbb{Z}^+$ be a fixed discriminant, and choose any form $f \in Q(D)$. There are two specific matrices $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$ such that $f \cdot \mathbf{A} = f$, namely

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

This shows that the set $\mathrm{Aut}\,(f)$, to be defined presently, is always nonempty.

Definition 2.3.1. Given $f \in Q(D)$, we define the set $\mathrm{Aut}\,(f)$, known as the set of *automorphs of* f, to be the collection of all $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$ such that $f \cdot \mathbf{A} = f$.

Given any $f \in Q(D)$, we already know that $\mathrm{Aut}\,(f)$ contains at least two elements. An extremely important theorem, proved in Section 3.2, states that if f is a primitive form, then $\mathrm{Aut}\,(f)$ is a set with infinitely many distinct elements. This theorem holds specifically for indefinite forms of discriminant D satisfying Assumption 2.2.1, and it most certainly does not hold, for example, with respect to positive definite forms!

Theorem 2.3.2. *For a fixed discriminant $D \in \mathbb{Z}^+$, and any given form $f \in Q(D)$, the set* $\mathrm{Aut}\,(f)$ *forms a subgroup of* $\mathrm{SL}_2(\mathbb{Z})$.

Proof. We already noted above that $\mathrm{Aut}\,(f)$ is nonempty. If both $\mathbf{A}$ and $\mathbf{B}$ are in $\mathrm{Aut}\,(f)$, we first wish to prove that the product $\mathbf{A} \cdot \mathbf{B}$ is in $\mathrm{Aut}\,(f)$ as well. By assumption, we have $f \cdot \mathbf{A} = f$ and $f \cdot \mathbf{B} = f$. Using (2.2.18), we obtain

$$f \cdot (\mathbf{A} \cdot \mathbf{B}) = (f \cdot \mathbf{A}) \cdot \mathbf{B} = f \cdot \mathbf{B} = f,$$

which confirms that $\mathbf{A} \cdot \mathbf{B} \in \mathrm{Aut}\,(f)$. We finally need to verify that $\mathbf{A}^{-1} \in \mathrm{Aut}\,(f)$, assuming that $\mathbf{A} \in \mathrm{Aut}\,(f)$. Towards the end of the proof of Theorem 2.2.6, we showed that if $f \cdot \mathbf{A} = g$, then $g \cdot \mathbf{A}^{-1} = f$. Setting $g = f$ here allows us to see that if $\mathbf{A} \in \mathrm{Aut}\,(f)$, then $\mathbf{A}^{-1} \in \mathrm{Aut}\,(f)$. □

Our main goal in this section is to develop an important connection between the automorphs of an indefinite form f and the solutions of a famous Diophantine equation typically known as "Pell's Equation", which we instead refer to as the "Fermat-Pell Equation". We prefer this terminology as a way to give the proper credit to Fermat for his impressive contributions to our understanding of the solutions of this equation. There are several variants of the Fermat-Pell Equation, but the version most directly connected to the automorphs of an indefinite form f of discriminant $D \in \mathbb{Z}^+$ is the following:

$$t^2 - Du^2 = 4. \tag{2.3.1}$$

The goal, dating back to Fermat, is to find all *integer pair* solutions $(t, u) \in \mathbb{Z}^2$ to (2.3.1). We are immediately able to find two "trivial" such solutions to (2.3.1), namely, $(2, 0)$ and $(-2, 0)$. The crux of the matter, due to Fermat, is that there are *always* nontrivial solutions to (2.3.1), corresponding to any given fixed discriminant $D \in \mathbb{Z}^+$.

For the sake of clarity, we refer to (2.3.1) as the "Fermat-Pell 4-Equation". In Section 5.2, we study the integer pair solutions to the "Fermat-Pell 1-Equation"

$$t^2 - du^2 = 1. \tag{2.3.2}$$

The following theorem is the main result of this section. As usual, $D \in \mathbb{Z}^+$ is a fixed discriminant. A new twist here is that the form $f \in Q(D)$ chosen in this theorem must be *primitive* (from our results in Section 2.2, we know that there are infinitely many primitive forms in $Q(D)$). In Example 2.3.4 below, we look at a few specific examples that illustrate what can go wrong assuming f is not primitive.

Theorem 2.3.3. *Every automorph* $\mathbf{A} \in \mathrm{Aut}\,(f)$ *associated to a given primitive form* $f = [a, b, c] \in Q(D)$ *satisfies the formula*

$$\mathbf{A} = \begin{pmatrix} \frac{t-bu}{2} & -cu \\ au & \frac{t+bu}{2} \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})\,, \tag{2.3.3}$$

where (t, u) *is an arbitrary integer pair solution of the Fermat-Pell 4-Equation* (2.3.1).

Remarks.

1) Assuming that the integer pair (t, u) is a solution of (2.3.1), we claim that both $(t + bu)/2$ and $(t - bu)/2$ are integers. By the Remark at the end of Section 2.1, we have $b \equiv D \pmod 2$. By this same Remark, any integer and its square have the same parity, and it is easy to see that any integer and its additive inverse also have the same parity. Using this information, and the fact that t and u are both integers, we note that

$$t + bu \equiv t + Du \equiv t^2 - Du^2 = 4 \equiv 0 \,(\mathrm{mod}\ 2).$$

Thus, $t + bu$ is evenly divisible by 2, and so the quantity $(t + bu)/2$ is an integer. The same argument may be used to demonstrate that the quantity $(t - bu)/2$ is an integer as well, proving our claim.

2) Computing the determinant of the transformation matrix in (2.3.3), we find that

$$\frac{(t-bu)}{2}\cdot\frac{(t+bu)}{2}+acu^2 = \frac{t^2-b^2u^2}{4}+acu^2 = \frac{t^2-(b^2-4ac)\,u^2}{4} = \frac{t^2-Du^2}{4} = 1,$$

with the last equality holding since the integer pair (t, u) is a solution of (2.3.1). Taking this together with Remark 1) above shows that the matrix in (2.3.3) is an element of $\mathrm{SL}_2(\mathbb{Z})$.

3) Throughout the following proof, it is important to keep in mind that $a \neq 0$. This follows from the fact that the discriminant D satisfies Assumption 2.2.1.

Proof. Assume throughout that $f = [a, b, c] \in Q(D)$ is a fixed primitive form, and that (t, u) is any given integer pair solution to (2.3.1). We first wish to show that the matrix $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$ in (2.3.3) takes f into itself. In order to do this, it suffices to show that the coefficients a and b are left unchanged by the action of $\mathbf{A}$, since c is uniquely determined by D, a, and b, and will thus be left unchanged as well. Substituting the transformation values into (2.2.7), the new coefficient a_1 is given by

$$\begin{aligned} a_1 &= a\left(\frac{t-bu}{2}\right)^2 + b\left(\frac{t-bu}{2}\right)au + ca^2u^2 \\ &= a\frac{t^2}{4} - ab\frac{tu}{2} + ab^2\frac{u^2}{4} + ab\frac{tu}{2} - ab^2\frac{u^2}{2} + a^2cu^2 \\ &= \frac{a}{4}\left[t^2 - \left(b^2-4ac\right)u^2\right] \\ &= \frac{a}{4}\left(t^2 - Du^2\right) = a, \end{aligned}$$

with the last equality holding by (2.3.1). Similarly, substituting the transformation

values into (2.2.8), we see that the new coefficient b_1 is given by

$$\begin{aligned} b_1 &= -2a\left(\frac{t-bu}{2}\right)cu + b\left(1-2acu^2\right) + 2cau\left(\frac{t+bu}{2}\right) \\ &= -actu + abcu^2 + b - 2abcu^2 + actu + abcu^2 = b. \end{aligned}$$

Note that in the first equation above, we have made use of the equality

$$\frac{t^2 - b^2u^2}{4} - acu^2 = 1 - 2acu^2,$$

which follows from the equation in Remark 2) above. This confirms that the coefficients a and b are left unchanged by the action of $\mathbf{A}$. Since $D \in \mathbb{Z}^+$ is fixed, we conclude that the coefficient c remains unchanged as well. Therefore, the matrix $\mathbf{A}$ in (2.3.3) arising from a given solution (t, u) of (2.3.1) lies in $\mathrm{Aut}\,(f)$.

We have seen in the first part of the proof that any given integer pair solution (t, u) of the Fermat-Pell 4-Equation leads to a uniquely defined element of $\mathrm{Aut}\,(f)$. This second part of the proof will show that any given automorph $\mathbf{A} \in \mathrm{Aut}\,(f)$ leads to a uniquely defined integer pair solution (t, u) of the Fermat-Pell 4-Equation. For this part of the proof, we assume that

$$\mathbf{A} = \begin{pmatrix} r & s \\ m & n \end{pmatrix} \tag{2.3.4}$$

is a given element in $\mathrm{Aut}\,(f)$, and we need to prove that r, s, m, and n satisfy the formula given in (2.3.3), associated to some integer pair solution (t, u) of (2.3.1). By assumption, we have $f \cdot \mathbf{A} = f$, and so by (2.2.7), we have

$$a = ar^2 + brm + cm^2. \tag{2.3.5}$$

From (2.2.8), and a few straightforward algebraic steps, we find that

$$\begin{aligned} b &= 2ars + b + 2bsm + 2cmn, \\ 0 &= 2ars + 2bsm + 2cmn, \\ 0 &= ars + bsm + cmn. \end{aligned} \tag{2.3.6}$$

We may now use equations (2.3.5) and (2.3.6) to eliminate b. Multiplying both sides of (2.3.5) by s yields

$$as = asr^2 + bsrm + cm^2s, \tag{2.3.7}$$

and multiplying both sides of (2.3.6) by r yields

$$0 = asr^2 + bsrm + cmnr. \tag{2.3.8}$$

Subtracting (2.3.8) from (2.3.7), we have

$$as = cm^2s - cmnr = cm(ms - rn) = -cm, \tag{2.3.9}$$

with the last equality holding by virtue of the fact that $ms - rn = -1$, since the 2×2 matrix in (2.3.4) is an element of $\mathrm{SL}_2(\mathbb{Z})$. A similar argument may be used to eliminate c. Multiplying both sides of (2.3.5) by n yields

$$an = ar^2n + brmn + cm^2n, \tag{2.3.10}$$

and multiplying both sides of (2.3.6) by m yields

$$0 = arsm + bsm^2 + cm^2n. \tag{2.3.11}$$

Subtracting (2.3.11) from (2.3.10), we have $an = ar^2n + brmn - arsm - bsm^2$, which

we may rearrange to obtain

$$an = ar^2n - arsm + brmn - bsm^2. \tag{2.3.12}$$

Since the transformation matrix in (2.3.4) is an element of $\mathrm{SL}_2(\mathbb{Z})$, we have $rn - sm = 1$, and so we may factor (2.3.12) to obtain

$$\begin{aligned} an &= ar^2n - arsm + brmn - bsm^2 \\ &= ar(rn - sm) + bm(rn - sm) \\ &= ar + bm, \end{aligned}$$

which we may rewrite as

$$a(n - r) = bm. \tag{2.3.13}$$

From (2.3.9) and (2.3.13), we see that $a \mid cm$ and $a \mid bm$ (recall that $a \neq 0$); thus, there exist integers $d, e \in \mathbb{Z}$ such that $ad = cm$ and $ae = bm$.

Everything done in the proof of Theorem 2.3.3 to this point works equally well whether f is primitive or imprimitive. Finally, at this juncture, in order to complete the second part of this proof, we bring into play the assumption that $f = [a, b, c]$ is a primitive form. This just means that $\gcd(a, b, c) = 1$, which implies by Corollary 1.1.7 that there exist integers $j, k, \ell \in \mathbb{Z}$ such that $aj + bk + c\ell = 1$. If we multiply both sides of this last equation through by m, and replace bm by ae and cm by ad, we obtain

$$\begin{aligned} ajm + bkm + c\ell m &= m, \\ ajm + aek + ad\ell &= m, \\ a(jm + ek + d\ell) &= m. \end{aligned}$$

Since all of the variables here are integers by construction, we conclude that $a \mid m$. Thus, there exists an integer $u \in \mathbb{Z}$ such that

$$m = au. \tag{2.3.14}$$

Substituting this expression for m into (2.3.9), we obtain $as = -cm = -cau$, or

$$s = -cu, \tag{2.3.15}$$

where division by a is allowed since $a \neq 0$. Using (2.3.14) again, in conjunction this time with (2.3.13), yields $a(n-r) = bm = bau$, or

$$n - r = bu. \tag{2.3.16}$$

Recalling that $nr - sm = 1$, and replacing $n - r$ by bu, s by $-cu$, and m by au, we find that

$$\begin{aligned}(n+r)^2 &= n^2 + 2nr + r^2 = (n-r)^2 + 4nr \\ &= (bu)^2 + 4(1+sm) = b^2u^2 + 4\left(1 - acu^2\right) \\ &= b^2u^2 + 4 - 4acu^2 = u^2\left(b^2 - 4ac\right) + 4 \\ &= Du^2 + 4.\end{aligned}$$

If we set

$$t = n + r \in \mathbb{Z}, \tag{2.3.17}$$

then we may simplify the above to read: $t^2 - Du^2 = 4$, which shows that (t, u) is an integer pair solution of the Fermat-Pell 4-Equation. If we solve the system of two

equations given by (2.3.16) and (2.3.17) for r and n, we find that

$$r = \frac{t - bu}{2}, \tag{2.3.18}$$

and

$$n = \frac{t + bu}{2}. \tag{2.3.19}$$

By (2.3.14), (2.3.15), (2.3.18), and (2.3.19), we see that the arbitrarily given automorph in (2.3.4) taking the primitive form $f = [a, b, c]$ into itself may be expressed as follows:

$$\begin{pmatrix} r & s \\ m & n \end{pmatrix} = \begin{pmatrix} \frac{t-bu}{2} & -cu \\ au & \frac{t+bu}{2} \end{pmatrix} \in \mathrm{Aut}\,(f) \subset \mathrm{SL}_2(\mathbb{Z})\,,$$

where (t, u) is, as we saw above, an integer pair solution of the Fermat-Pell 4-Equation (2.3.1). This completes the proof of Theorem 2.3.3. □

In Example 2.3.4 below, we give two simple examples that illustrate what can go wrong with regard to Theorem 2.3.3 if the form f is not primitive. We noted in the second half of the proof of Theorem 2.3.3 that if $f = [a, b, c]$ is primitive, then $a \mid m$ [see (2.3.14)]. A review of this part of the proof highlights clearly that (2.3.14) is the linchpin result that allows us to properly carry out the second half of the proof. If f is imprimitive, then it can happen that $a \nmid m$ and that is where the trouble lies, as we see below. We will also show how easy it is to obtain nontrivial solutions to (2.3.1), if we have in hand an element in $\mathrm{Aut}\,(f)$.

Example 2.3.4. Consider the imprimitive form $f = [2, 6, 2]$ of discriminant $D = 20$. It is easy to verify that

$$\begin{pmatrix} r & s \\ m & n \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ -1 & 0 \end{pmatrix} \in \mathrm{Aut}\,(f),$$

and so $m = -1$. However, we have $a = 2 \nmid m$ in this example. From (2.3.14) and (2.3.17), we can still solve for u and t in this example to obtain $u = \frac{m}{a} = \frac{-1}{2}$ and $t = r + n = 3 + 0 = 3$. Note that $(t, u) = (3, -1/2)$ is a solution to (2.3.1), but it is not one of the *integer pair* solutions of interest to us. Similarly,

$$\begin{pmatrix} r & s \\ m & n \end{pmatrix} = \begin{pmatrix} 5 & 3 \\ -2 & -1 \end{pmatrix}$$

is an automorph of the imprimitive form $f = [4, 12, 6]$ of discriminant $D = 48$. However, $a = 4 \nmid -2 = m$. Again, $(t, u) = (4, -1/2)$ is a solution of (2.3.1), but not of the type we are seeking. Consider now the primitive form $f = [1, 6, 4]$ of discriminant $D = 20$. An easy check confirms that

$$\begin{pmatrix} r & s \\ m & n \end{pmatrix} = \begin{pmatrix} 21 & 16 \\ -4 & -3 \end{pmatrix} \in \mathrm{Aut}\,(f),$$

so that $m = -4$. Clearly, $a = 1 \mid m$ and $u = \frac{m}{a} = -4$. Also, $t = r + n = 18$, and it is easy to verify that $(t, u) = (18, -4)$ is an integer pair solution to (2.3.1) with $D = 20$.

The reader may well ask how we got our hands on the different automorphs in Example 2.3.4. We present an algorithm in Section 3.2 that allows us to produce automorphs of special "reduced" forms (all three forms in Example 2.3.4 are of this special type; see Definition 3.1.1). Theorem 2.3.3 works in both directions. If we have a systematic procedure to generate all automorphs of a given primitive form f of discriminant $D \in \mathbb{Z}^+$, then we are able to obtain all possible integer pair solutions to (2.3.1). Conversely, integer pair solutions of (2.3.1) associated to a given discriminant D translate into automorphs of forms in $Q(D)$. We consider a "brute force" method to obtain nontrivial solutions to (2.3.1) for a fixed discriminant $D \in \mathbb{Z}^+$ later in this

section. It is worth mentioning that the two trivial solutions $(t, u) = (2, 0)$ and $(-2, 0)$ to (2.3.1) lead via (2.3.3) to the two "obvious" automorphs that always lie in $\mathrm{Aut}\,(f)$, displayed at the very beginning of this section.

The following famous theorem was known to Fermat, and Weil [7] has constructed a proof of this theorem which he conjectures to be similar to what Fermat might have had in mind, even if no record exists of the proof that Fermat claims to have possessed.

Theorem 2.3.5. *For any given fixed discriminant $D \in \mathbb{Z}^+$, the Fermat-Pell 4-Equation* (2.3.1) *possesses an integer pair solution $(t, u) \in \mathbb{Z}^2$ with $u \in \mathbb{Z}^+$.*

The modern proofs of Theorem 2.3.5 fall into two distinct categories: constructive and non-constructive. Landau ([5], starting on page 76) offers a beautiful, but non constructive, proof of Theorem 2.3.5 which relies upon the pigeon-hole principle. Assuming the truth of Theorem 2.3.5, let $F(D)$ denote the *nonempty* set of all integer pair solutions $(t, u) \in \mathbb{Z}^2$ to (2.3.1) with $u \in \mathbb{Z}^+$. By the Well-Ordering Principle, there is a uniquely defined integer $u_1 \in \mathbb{Z}^+$ and an integer $t\,'$ with $(t\,', u_1) \in F(D)$ such that $u_1 \leq u$ for any given pair $(t, u) \in F(D)$. Given the existence of the integer $u_1 \in \mathbb{Z}^+$ associated to a fixed discriminant $D \in \mathbb{Z}^+$, there is a straightforward brute-force algorithm to compute it. If we set in succession $u = 1, 2, 3, \ldots$ and test each time if the positive integer $4 + Du^2$ is a perfect square, then $u = u_1$ is the first positive integer we encounter in this process where $4 + Du^2$ is a perfect square. A corresponding positive integer $t_1 \in \mathbb{Z}^+$ is then uniquely determined by the equality $t_1^2 = 4 + Du_1^2$. Note that $t_1 > 2$ since $Du_1^2 > 0$.

Definition 2.3.6. Given a fixed discriminant $D \in \mathbb{Z}^+$, the uniquely defined integer pair $(t_1, u_1) \in F(D)$ obtainable in principle from the algorithm just described is called

the *minimal solution of* (2.3.1).

The phrase "obtainable in principle" used in Definition 2.3.6 is meant to highlight the fact that the brute-force algorithm described above makes perfect sense in principle but might be hopeless in practice! In Table 2.3.1 below, we present the minimal solution to (2.3.1) for all discriminants D with $5 \leq D \leq 65$.

Table 2.3.1. Minimal solution to (2.3.1) for all discriminants D with $5 \leq D \leq 65$

D	Minimal Solution Pair (t_1, u_1)	D	Minimal Solution Pair (t_1, u_1)
5	$(3, 1)$	40	$(38, 6)$
8	$(6, 2)$	41	$(4098, 640)$
12	$(4, 1)$	44	$(20, 3)$
13	$(11, 3)$	45	$(7, 1)$
17	$(66, 16)$	48	$(14, 2)$
20	$(18, 4)$	52	$(1298, 180)$
21	$(5, 1)$	53	$(51, 7)$
24	$(10, 2)$	56	$(30, 4)$
28	$(16, 3)$	57	$(302, 40)$
29	$(27, 5)$	60	$(8, 1)$
32	$(6, 1)$	61	$(1523, 195)$
33	$(46, 8)$	65	$(258, 32)$
37	$(146, 24)$		

A cursory look at this table shows the wild variation in the size of u_1 as D is varied.

With a fast computer, the algorithm described above works fine for relatively small discriminants $D \in \mathbb{Z}^+$ in the hundreds, but there are well-documented instances where u_1 is positively enormous for values of D in the thousands and millions.

We offer a constructive proof of Theorem 2.3.5 in Section 3.2. This proof is accompanied with an effective algorithm (see Algorithm 3.2.7) that allows one to compute the minimal solution to (2.3.1) even when the integers t_1 and u_1 are gigantic.

Associated to the minimal solution (t_1, u_1) of (2.3.1) introduced in Definition 2.3.6, we define a corresponding real quadratic irrational number

$$\varepsilon_1(D) = \frac{t_1 + u_1\sqrt{D}}{2} \tag{2.3.20}$$

(such numbers are discussed in greater detail in Section 2.4). The uniquely defined real number $\varepsilon_1(D)$ associated to the discriminant $D \in \mathbb{Z}^+$ is known as the "fundamental unit of discriminant D".

2.4. Real Quadratic Irrationals

We now consider a special set of real numbers that are intimately connected with indefinite binary quadratic forms.

Definition 2.4.1. Let $d \in \mathbb{Z}^+$ be a fixed positive integer that is not a perfect square. A *real quadratic irrational* is any number β of the form

$$\beta = \frac{\ell + m\sqrt{d}}{n}, \tag{2.4.1}$$

where $\ell \in \mathbb{Z}$, and $m, n \in \mathbb{Z} \setminus \{0\}$.

It is easy to see that such a number is real, and β is irrational by the choice of d and since $m \neq 0$. The word "quadratic" appears in the definition since such a number always satisfies a quadratic equation; the number appearing in (2.4.1) is a root of the

quadratic polynomial

$$n^2x^2 - 2\ell nx + (\ell^2 - dm^2) \in \mathbb{Z}[x]. \tag{2.4.2}$$

The connection with indefinite binary quadratic forms, alluded to above, is forged in the following definition.

Definition 2.4.2. Let $D \in \mathbb{Z}^+$ be a fixed discriminant and let $f = [a, b, c]$ be a form in $Q(D)$. The real quadratic irrational number β *corresponding to* f is defined by

$$\beta = \frac{b + \sqrt{D}}{2a} \tag{2.4.3}$$

(note that by our assumptions, $a \neq 0$). This correspondence sets up a map Z having domain $Q(D)$ and codomain equal to the set of all real quadratic irrational numbers. We let $QI(D)$ denote the range of Z, and thus the function $Z : Q(D) \to QI(D)$ is surjective by construction.

This correspondence is exploited on several occasions in this thesis; the crucial choice made in (3.1.8) is but one example.

Theorem 2.4.3. *Let $D \in \mathbb{Z}^+$ be a fixed discriminant. The map*

$$Z : Q(D) \to QI(D) \tag{2.4.4}$$

described in Definition 2.4.2 is injective and therefore sets up a one-to-one correspondence between the two sets $Q(D)$ and $QI(D)$.

Proof. Let $f_1 = [a_1, b_1, c_1]$ and $f_2 = [a_2, b_2, c_2]$ be two arbitrarily given forms in $Q(D)$, and assume that $Z(f_1) = Z(f_2)$, which just says that

$$\frac{b_1 + \sqrt{D}}{2a_1} = \frac{b_2 + \sqrt{D}}{2a_2},$$

or $2a_2b_1 + 2a_2\sqrt{D} = 2a_1b_2 + 2a_1\sqrt{D}$. Rearrangement gives

$$(2a_2 - 2a_1)\sqrt{D} = 2a_1b_2 - 2a_2b_1, \tag{2.4.5}$$

and if $2a_2 - 2a_1 \neq 0$, then $\sqrt{D} \in \mathbb{Q}$, contradicting the fact that $\sqrt{D}$ is an irrational number. We conclude that $2a_2 = 2a_1$, or $a_1 = a_2$, so that from (2.4.5) we see that $2a_1b_1 = 2a_1b_2$, or $b_1 = b_2$. Since f_1 and f_2 are both in $Q(D)$, we also have $c_1 = c_2$, which shows that the map in (2.4.4) is injective. □

The most important result in this section is the following theorem.

Theorem 2.4.4. *Let $D \in \mathbb{Z}^+$ be a fixed discriminant and let $f = [a, b, c]$ be a form of discriminant D whose corresponding real quadratic irrational is $\beta = (b + \sqrt{D})/2a$. Assume furthermore that*

$$\mathbf{A} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

and that $f \cdot \mathbf{A} = f_1 = [a_1, b_1, c_1]$, where $\beta_1 = (b_1 + \sqrt{D})/2a_1$ is the real quadratic irrational corresponding to f_1. Then,

$$\beta_1 = \frac{u\beta + s}{t\beta + r}. \tag{2.4.6}$$

Remark. The quantity $t\beta + r$ appearing in the denominator of (2.4.6) is never equal to zero. To see this, assume first that $t = 0$. In this case, we must have $r \neq 0$ since $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$. Now assume that $t \neq 0$. If we have $t\beta + r = 0$, then

$$\beta = -\frac{r}{t} \in \mathbb{Q},$$

contradicting the fact that β is an irrational number.

Proof. We simply start with the expression on the right side of (2.4.6) and show after

some algebraic manipulation that this expression is equal to β_1. We have

$$\begin{aligned}\frac{s+u\left(\frac{b+\sqrt{D}}{2a}\right)}{r+t\left(\frac{b+\sqrt{D}}{2a}\right)} &= \frac{2as+u(b+\sqrt{D})}{2ar+t(b+\sqrt{D})} \\ &= \frac{\left(2as+bu+u\sqrt{D}\right)}{\left(2ar+bt+t\sqrt{D}\right)}\cdot\frac{\left(2ar+bt-t\sqrt{D}\right)}{\left(2ar+bt-t\sqrt{D}\right)}. \end{aligned} \tag{2.4.7}$$

The numerator on the right hand side of (2.4.7) multiplies out to the quantity

$$\begin{aligned} &4a^2rs+2abst-2ast\sqrt{D}+2abru+b^2tu-btu\sqrt{D}+2aru\sqrt{D}+btu\sqrt{D}-tuD \\ &\quad = 4a^2rs+2abru+2abst+b^2tu-tu\left(b^2-4ac\right)+2a\left(ru-st\right)\sqrt{D} \\ &\quad = 2a\left(2ars+b\left(ru+st\right)+2ctu+\sqrt{D}\right) \\ &\quad = 2a\left(b_1+\sqrt{D}\right), \end{aligned} \tag{2.4.8}$$

with the last equality holding by (2.2.8). The denominator on the right hand side of (2.4.7) multiplies out to the quantity

$$\begin{aligned}\left(2ar+bt\right)^2-t^2D &= 4a^2r^2+4abrt+b^2t^2-t^2\left(b^2-4ac\right) \\ &= 4a\left(ar^2+brt+ct^2\right) \\ &= 2a\cdot 2a_1, \end{aligned} \tag{2.4.9}$$

with the last equality holding by (2.2.7). Combining (2.4.7), (2.4.8), and (2.4.9), we find that

$$\frac{u\beta+s}{t\beta+r}=\frac{2a\left(b_1+\sqrt{D}\right)}{2a\cdot 2a_1}=\beta_1,$$

which confirms (2.4.6). □

The expression on the right hand side of (2.4.6) is connected to an important type of mapping.

Definition 2.4.5. Given a matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})\,,$$

the *linear fractional transformation* $L_{\mathbf{A}}$ *associated to* $\mathbf{A}$ is the mapping sending

$$x \mapsto \frac{ax+b}{cx+d}\,, \tag{2.4.10}$$

where $x \in \mathbb{R}$. If $c = 0$, then we must have $d \neq 0$ since $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$, and the domain of $L_{\mathbf{A}}$ is all real numbers. If $c \neq 0$, then the domain of $L_{\mathbf{A}}$ is $\mathbb{R} \setminus \{-\frac{d}{c}\}$. Either way, the domain of $L_{\mathbf{A}}$ always includes all irrational numbers.

Notation. Given a matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})\,,$$

and a real number γ in the domain of $L_{\mathbf{A}}$, we employ the following shorthand notation:

$$\mathbf{A} \cdot \gamma := L_{\mathbf{A}}(\gamma) = \frac{a\gamma+b}{c\gamma+d}\,. \tag{2.4.11}$$

Using this notation, we may conveniently rephrase the statement of Theorem 2.4.4 to read as follows: If

$$\mathbf{A} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})\,,$$

and $f \cdot \mathbf{A} = f_1 = [a_1, b_1, c_1]$, then (2.4.6) may be rewritten as

$$\beta_1 = \mathbf{A}^w \cdot \beta, \tag{2.4.12}$$

where use is made of Definition 1.1.22. Using the notation of Definition 2.4.2, we may

also write this last equality as

$$Z(f_1) = \mathbf{A}^w \cdot Z(f). \tag{2.4.13}$$

We are accustomed to write $f \cdot \mathbf{A}$ in this order since we have a right group action of $\mathrm{SL}_2(\mathbb{Z})$ on the set $Q(D)$. It is no accident that $\mathbf{A}^w$ is placed to the left of $\beta = Z(f)$ in (2.4.12) and (2.4.13), as it is shown below that $\mathrm{SL}_2(\mathbb{Z})$ has a natural *left* group action on the set $QI(D)$.

Theorem 2.4.6. *Let $D \in \mathbb{Z}^+$ be a fixed discriminant and assume we have an arbitrarily given $\beta \in QI(D)$. If $\mathbf{A}, \mathbf{B} \in \mathrm{SL}_2(\mathbb{Z})$, then*

$$\mathbf{A} \cdot \beta \in QI(D)\,; \tag{2.4.14}$$

we have

$$\mathbf{I} \cdot \beta = \beta, \tag{2.4.15}$$

where $\mathbf{I} \in \mathrm{SL}_2(\mathbb{Z})$ is the 2×2 identity matrix, and

$$(\mathbf{B} \cdot \mathbf{A}) \cdot \beta = \mathbf{B} \cdot (\mathbf{A} \cdot \beta). \tag{2.4.16}$$

Comparison to §1.7 *in* [1] *confirms that the group $\mathrm{SL}_2(\mathbb{Z})$ has a left group action on the set $QI(D)$.*

Proof. By Theorem 2.4.3, there exists a unique form $f \in Q(D)$ such that $Z(f) = \beta$. If we set $f_1 = f \cdot \mathbf{A}^w$, then $Z(f_1) = \mathbf{A} \cdot \beta$ by (2.4.12) and (1.1.2), and $Z(f_1) \in QI(D)$, which confirms (2.4.14). Note that (2.4.15) follows immediately from (2.4.11). To

prove (2.4.16), we set

$$\mathbf{B} = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

The right hand side of (2.4.16) may be written in the form

$$\begin{aligned} \frac{a_1\left(\frac{a\beta+b}{c\beta+d}\right)+b_1}{c_1\left(\frac{a\beta+b}{c\beta+d}\right)+d_1} &= \frac{a_1(a\beta+b)+b_1(c\beta+d)}{c_1(a\beta+b)+d_1(c\beta+d)} \\ &= \frac{(a_1a+b_1c)\beta+(a_1b+b_1d)}{(c_1a+d_1c)\beta+(c_1b+d_1d)}. \end{aligned} \tag{2.4.17}$$

Since

$$\mathbf{B}\cdot\mathbf{A} = \begin{pmatrix} a_1a+b_1c & a_1b+b_1d \\ c_1a+d_1c & c_1b+d_1d \end{pmatrix},$$

we have established (2.4.16) as well. □

Using the notation introduced in (2.4.11), we may rewrite the expression appearing in (2.4.17) as $L_{\mathbf{B}}\left[L_{\mathbf{A}}(\beta)\right]$, which is the composition of two linear fractional transformations. From this perspective, the fact that we have a *left* action here makes perfect sense.

Theorem 2.4.7. *Let $D \in \mathbb{Z}^+$ be a fixed discriminant and let $f \in Q(D)$. If we have $\mathbf{A}, \mathbf{B} \in \mathrm{SL}_2(\mathbb{Z})$, and $f_2 = f\cdot(\mathbf{A}\cdot\mathbf{B})$, then $Z(f_2) = ((\mathbf{A}\cdot\mathbf{B})^w)\cdot Z(f)$.*

Proof. Assuming that $f\cdot\mathbf{A} = f_1$, we have $Z(f_1) = \mathbf{A}^w\cdot Z(f)$ by (2.4.13). Similarly, if $f_1\cdot\mathbf{B} = f_2$, then $Z(f_2) = \mathbf{B}^w\cdot Z(f_1)$. Note that

$$f\cdot(\mathbf{A}\cdot\mathbf{B}) = (f\cdot\mathbf{A})\cdot\mathbf{B} = f_1\cdot\mathbf{B} = f_2,$$

where (2.2.18) is invoked in the first equality. We also have

$$Z(f_2) = \mathbf{B}^w\cdot Z(f_1) = \mathbf{B}^w\cdot(\mathbf{A}^w\cdot Z(f)) = \left(\mathbf{B}^w\cdot\mathbf{A}^w\right)\cdot Z(f), \tag{2.4.18}$$

with the last equality holding by (2.4.16). Combining (2.4.18) with (1.1.1), we conclude that

$$Z(f_2) = \left((\mathbf{A} \cdot \mathbf{B})^w\right) \cdot Z(f).$$

□

In Section 4.1, we make use of the following corollary which follows easily by induction from Theorem 2.4.7.

Corollary 2.4.8. *Let $D \in \mathbb{Z}^+$ be a fixed discriminant and let $f \in Q(D)$. If we have $\mathbf{A}_1, \mathbf{A}_2, \ldots \mathbf{A}_n \in \mathrm{SL}_2(\mathbb{Z})$, and $f_n = f \cdot (\mathbf{A}_1 \cdot \mathbf{A}_2 \cdots \mathbf{A}_n)$, then*

$$Z(f_n) = \left((\mathbf{A}_1 \cdot \mathbf{A}_2 \cdots \mathbf{A}_n)^w\right) \cdot Z(f).$$

2.5. Dirichlet's Class Number Formula

In order to state Dirichlet's famous "class number formula" for $h(D)$ (see the comment at the very end of Section 2.2), we require one further ingredient, known as the "Kronecker symbol", named in honor of Leopold Kronecker. We assume that $D \in \mathbb{Z}^+$ is a fixed discriminant throughout the following discussion. The Kronecker symbol is an arithmetic function, denoted by $\chi_D(n)$, and defined for all positive integers $n \in \mathbb{Z}^+$, taking on only the three values $-1, 0$, and 1. The strategy for defining $\chi_D(n)$ is straightforward. We set $\chi_D(1) = 1$ and then uniquely specify χ_D at every prime number p. Once that is done, we extend the function χ_D to all positive integers multiplicatively using the Fundamental Theorem of Arithmetic. For example, if $\chi_D(3) = 1$ and $\chi_D(7) = -1$, then we define $\chi_D(21) = \chi_D(3 \cdot 7) := \chi_D(3) \cdot \chi_D(7) = (1)(-1) = -1$. Defining χ_D at the prime number $p = 2$ is simple, so we handle it first. If p is an odd prime number, we need to employ the "Legendre symbol", named in honor of Adrien-Marie Legendre, to be defined below.

Definition 2.5.1. $p = 2 : \chi_D(2) = 0$ if $2 \mid D$, or equivalently, if $D \equiv 0 \pmod 4$.

$$\chi_D(2) = \begin{cases} 1 & \text{if } D \equiv 1 \pmod 8 \\ -1 & \text{if } D \equiv 5 \pmod 8. \end{cases}$$

Since $D \equiv 0$ or $1 \pmod 4$, this covers all possibilities.

Definition 2.5.2. If p is an odd prime number, and a is an integer not divisible by p, the *Legendre symbol*, denoted by $\left(\frac{a}{p}\right)$, is defined to have the value $+1$ if there is an integer x such that $x^2 \equiv a \pmod p$. If there is no such integer x, the symbol is set equal to -1.

Definition 2.5.3. $p =$ odd prime: $\chi_D(p) = 0$ if $p \mid D$.

$$\chi_D(p) = \left(\frac{D}{p}\right) (\leftarrow \textit{Legendre symbol}) \text{ if } p \nmid D.$$

For example, if $D = 5$ we have $\chi_5(1) = 1$, $\chi_5(2) = -1$, $\chi_5(3) = -1$, $\chi_5(4) = 1$, $\chi_5(5) = 0$, and then this pattern is repeated modulo 5 forever after so that $\chi_5(6) = 1$, $\chi_5(7) = -1$, $\chi_5(8) = -1$, $\chi_5(9) = 1$, $\chi_5(10) = 0$, et cetera. This provides a prototypical example of the behavior of χ_D for any given fixed discriminant $D \in \mathbb{Z}^+$.

It can be shown (see [5]) that the infinite series

$$\sum_{n=1}^{\infty} \frac{\chi_D(n)}{n} \tag{2.5.1}$$

is conditionally convergent. The limit value of the series in (2.5.1) is usually denoted by $L(1, \chi_D)$, since this value is equal to the L-function associated to the Dirichlet character χ_D evaluated at 1. Dirichlet's formula (certainly one of the most amazing

formulas in all of mathematics!) reads as follows:

$$h(D) = \frac{\sqrt{D}}{\log(\varepsilon_1(D))} \cdot L(1, \chi_D), \tag{2.5.2}$$

where the definition of the fundamental unit $\varepsilon_1(D)$ is given in (2.3.20). Part of the fascination of this formula is that the positive *integer* $h(D)$ is obtained through transcendental means. Using (2.5.2) to compute the integer $h(D)$ requires that enough terms are added together in (2.5.1) to ensure that the overall error on the right hand side of (2.5.2) is less than one half. For all of the examples computed in Example 2.5.4 below, the sum in (2.5.1) is taken from $n = 1$ to $n = 100,000$, which gives more than enough accuracy for the value of $L(1, \chi_D)$ in order to nail down the integer $h(D)$ precisely. The values used for t_1 and u_1 in (2.3.20) are taken from Table 2.3.1.

Example 2.5.4. If $D = 5$, then we compute the approximate value $L(1, \chi_5) = 0.43040894$, and the approximate value on the right hand side of (2.5.2) comes out to 0.99999999, so that $h(5) = 1$. If $D = 8$, then $L(1, \chi_8) = 0.62322524$, and we find that $h(8) = 1$. If $D = 12$, we obtain an approximate value $L(1, \chi_{12}) = 0.76035599$, which leads to the approximate value of 2.00002630 on the right hand side of (2.5.2), and so $h(12) = 2$. Continuing in this way gives us all of the values displayed in Table 2.5.1 below. All of the values listed for $L(1, \chi_D)$ are good to at least 3 decimal places.

Table 2.5.1. Values of $L(1, \chi_D)$ and $h(D)$ for all discriminants D with $5 \leq D \leq 65$

D	$L(1, \chi_D)$	$h(D)$	D	$L(1, \chi_D)$	$h(D)$
5	0.43040894	1	40	1.15008652	2
8	0.62322524	1	41	1.29910306	1
12	0.76035599	2	44	0.90246065	2
13	0.66275539	1	45	0.57386859	2
17	1.01608483	1	48	0.76035600	2
20	0.64561341	1	52	0.99412309	1
21	0.68379725	2	53	0.54000494	1
24	0.93587131	2	56	0.90868078	2
28	1.04646489	2	57	1.51272617	2
29	0.61178629	1	60	1.06553432	4
32	0.62322524	2	61	0.93834020	1
33	1.33280719	2	65	1.37751601	2
37	0.81927217	1			

CHAPTER III

REDUCTION THEORY

AND THE FERMAT-PELL 4-EQUATION

3.1. Zagier's Reduction Theory

Let $D \in \mathbb{Z}^+$ be a fixed discriminant satisfying the conditions in Assumption 2.2.1. From Section 2.2, we know that there are infinitely many distinct indefinite binary quadratic forms of discriminant D. Recall from Definition 2.2.2 that

$$\text{if } \; D \equiv 0 \,(\text{mod } 4), \quad \text{then } \; x^2 - \frac{D}{4}y^2 \tag{3.1.1}$$

is the principal form of discriminant D, and

$$\text{if } \; D \equiv 1 \,(\text{mod } 4), \quad \text{then } \; x^2 + xy + \frac{1-D}{4}y^2 \tag{3.1.2}$$

is the principal form of discriminant D.

We showed in Section 2.2 how to partition all forms of discriminant D into equivalence classes under the action of $\mathrm{SL}_2(\mathbb{Z})$. The equivalence class that the principal form of discriminant D falls into is called the "principal class of discriminant D". Our first goal in the present section is to show that within each equivalence class of forms of discriminant D, there exists at least one "reduced form." Some classes will actually contain several distinct reduced forms, but any given class will contain at most *finitely* many distinct reduced forms, even if such a class always contains infinitely many distinct forms in total by Theorem 2.2.8.

We also develop an algorithm which takes as input an arbitrary form of discriminant D, and after a *finite* number of unimodular transformations, produces a

reduced form lying in the same equivalence class as the form with which we started. As we saw in Section 2.2, there are only finitely many equivalence classes of forms of fixed discriminant D, and the reduced forms of discriminant D give us a means of identifying and labeling these various classes. We show at the end of this section that reduced forms allow us to decide in a finite number of steps if two arbitrarily given forms of discriminant D lie in the same equivalence class as each other or not.

The reduction algorithm mentioned above only employs unimodular transformations of a special type, which we now consider in detail. Our transformations are carried out using matrices of the form

$$\mathbf{S}(n) = \begin{pmatrix} n & 1 \\ -1 & 0 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})\,, \tag{3.1.3}$$

with $n \in \mathbb{Z}$ chosen specifically as described below. By (2.2.7), (2.2.8), and (2.2.9), such a transformation takes a form $f = [a, b, c] := ax^2 + bxy + cy^2$, with $a, b, c \in \mathbb{Z}$ and discriminant $D = b^2 - 4ac$, to a form $f\,' = [a\,', b\,', c\,']$ whose coefficients are given by

$$a\,' = an^2 - bn + c \tag{3.1.4}$$

$$b\,' = 2an - b \tag{3.1.5}$$

$$c\,' = a \tag{3.1.6}$$

with discriminant $D = (b\,')^2 - 4a\,'c\,'$.

Given a form $f = [a, b, c]$ whose discriminant $D \in \mathbb{Z}^+$ satisfies Assumption 2.2.1, we recall that we must have $a \neq 0$ and $c \neq 0$, since otherwise we would have $D = b^2$, violating the condition that D is *not* a perfect square. The significance of this restriction is immediately seen, since given such a form $f = [a, b, c]$, the transformation $\mathbf{S}(n)$ that we apply to it is uniquely determined by choosing $n \in \mathbb{Z}$ such that the

following inequalities are satisfied:

$$n > \frac{b+\sqrt{D}}{2a} > n-1. \tag{3.1.7}$$

Due to the division by $2a$, it is crucial that $2a \neq 0$, or that $a \neq 0$. The inequalities in (3.1.7) are strict, since the quantity $(b+\sqrt{D})/2a$ is irrational. This irrationality follows immediately from the fact that $\sqrt{D}$ is itself an irrational number. An equivalent way to state (3.1.7) is that $n \in \mathbb{Z}$ is chosen by setting it equal to the ceiling of the irrational number $(b+\sqrt{D})/2a$, namely,

$$n = \left\lceil \frac{b+\sqrt{D}}{2a} \right\rceil, \quad \text{or} \quad n = \lceil Z(f) \rceil, \tag{3.1.8}$$

where the description of the map Z is found in Definition 2.4.2.

After obtaining the new form $[a', b', c']$ by applying $\mathbf{S}(n)$ to $[a, b, c]$, we then apply a transformation $\mathbf{S}(n')$ to $[a', b', c']$ by choosing $n' = \left\lceil \frac{b'+\sqrt{D}}{2a'} \right\rceil$, and then iterate and continue in this way. This process constitutes the reduction algorithm mentioned above.

The following definition is central to this entire section. There are certain forms that lie in any given equivalence class that have special restrictions upon their coefficients.

Definition 3.1.1. We say that an indefinite form $[a, b, c]$ of discriminant $D > 0$ is *reduced* if

$$a > 0,\ c > 0,\ b > a + c. \tag{3.1.9}$$

It is worth noting that this is *not* the standard definition of what it means for an indefinite binary quadratic form to be reduced. Definition 3.1.1 is due to Zagier ([8], p. 122). To emphasize this point, we could say that an indefinite form satisfying

Definition 3.1.1 is "Zagier-reduced", but since this is the only definition of "reduced" that is used in this thesis, this emphasis is unnecessary. It should be noted that we follow the presentation in Zagier's book ([8], §13) rather closely throughout this whole section with the intention of giving a more leisurely and detailed exposition of his main results.

The following theorem encapsulates the claims made earlier in this section that each equivalence class contains at least one reduced form and at most finitely many such forms. It is through the use of the reduction algorithm that the proof of this theorem is obtained.

Theorem 3.1.2. *If $D \in \mathbb{Z}^+$ is a fixed discriminant, then every form of discriminant D is taken by a finite number of unimodular transformations $\mathbf{S}(n)$ to a reduced form in the same equivalence class. Furthermore, there are only finitely many reduced forms of discriminant D, and these may be explicitly described.*

Remark. Since the reduction algorithm may be applied to any given form of discriminant D, Theorem 3.1.2 shows that each equivalence class contains at least one reduced form. In Section 2.2, we set $t(D)$ equal to the number of equivalence classes of forms of discriminant D, and we now set $j(D)$ equal to the total number of reduced forms of discriminant D. Theorem 3.1.2 shows that $j(D) \in \mathbb{Z}^+$ and that $t(D) \leq j(D)$, offering an independent proof of Theorem 2.2.11, confirming again that the number of equivalence classes of forms of discriminant D is finite. It turns out that we almost always have $t(D) < j(D)$ since there is generally at least one equivalence class of forms of discriminant D containing two or more reduced forms. The reduced forms lying in a particular equivalence class are interconnected amongst themselves as part of a structure known as a "cycle of reduced forms." A detailed discussion of these

cycles appears later in this section.

Proof. We begin by proving the first statement of Theorem 3.1.2, that every form of discriminant D is taken by a finite number of unimodular transformations $\mathbf{S}(n)$ to a reduced form in the same equivalence class.

Let $[a, b, c]$ be an arbitrary form of discriminant D, and set $n = \left\lceil \frac{b+\sqrt{D}}{2a} \right\rceil$. Since n satisfies the inequalities in (3.1.7), the number θ defined by

$$\theta = n - \frac{b + \sqrt{D}}{2a} \tag{3.1.10}$$

is an irrational number that satisfies the inequalities

$$0 < \theta < 1. \tag{3.1.11}$$

We may rewrite (3.1.10) in the form $b + \sqrt{D} = 2an - 2a\theta$, or

$$-b + 2an = \sqrt{D} + 2a\theta. \tag{3.1.12}$$

The transformation $\mathbf{S}(n)$ takes the form $[a, b, c]$ to the form $[a', b', c']$ with

$$a' = an^2 - bn + c \tag{3.1.13}$$

$$b' = 2an - b \tag{3.1.14}$$

$$c' = a. \tag{3.1.15}$$

There is nothing we can do to simplify (3.1.15). By use of (3.1.12), we can rewrite (3.1.14) as

$$b' = \sqrt{D} + 2a\theta. \tag{3.1.14$'$}$$

In order to obtain a more useful version of (3.1.13), we note by (3.1.10) that

$$n^2 = \left(\frac{b+\sqrt{D}}{2a}\right)^2 + \frac{\left(b+\sqrt{D}\right)\theta}{a} + \theta^2,$$

or

$$n^2 = \frac{b^2+2b\sqrt{D}+D}{4a^2} + \frac{\left(b+\sqrt{D}\right)\theta}{a} + \theta^2. \tag{3.1.16}$$

We may now rewrite (3.1.13), using (3.1.16) and (3.1.10), in the form

$$a' = \frac{b^2+2b\sqrt{D}+D}{4a} + \left(b+\sqrt{D}\right)\theta + a\theta^2 - b\left(\frac{b+\sqrt{D}}{2a}\right) - b\theta + c, \tag{3.1.17}$$

or

$$a' = a\theta^2 + \theta\sqrt{D}, \tag{3.1.18}$$

since

$$\frac{b^2+2b\sqrt{D}+D-2b^2-2b\sqrt{D}+4ac}{4a} = 0$$

follows from $D = b^2 - 4ac$. From (3.1.18) and (3.1.11), we note that if $a > 0$, then $a' > 0$. This shows that with respect to the sequence of forms $[a,b,c]$, $[a',b',c']$, $[a'',b'',c'']\,,\ldots$ (we think of this as an infinite sequence) that are generated by the $\mathbf{S}(n)$ transformations, once (and if) the "a–coefficient" becomes positive, it stays positive in each succeeding form. Thus, it remains to show that the a–coefficient does indeed become positive.

From (3.1.11), we know that

$$0 < \theta^2 < 1, \tag{3.1.19}$$

and thus if $a < 0$, then

$$a < a\theta^2 < 0. \tag{3.1.20}$$

From (3.1.18) and (3.1.11), we have

$$a\theta^2 < a', \tag{3.1.21}$$

and combining this with (3.1.20) gives us

$$a < a' \tag{3.1.22}$$

when $a < 0$. Since $a, a', a'', \ldots$ are all integers, after a *finite* number of steps with respect to the sequence of forms, the a–coefficient will become positive (the a–coefficient will never be $= 0$ by our choice of D). Once the a–coefficient becomes positive, it will remain positive, as we saw above. Likewise, since $c' = a$ by (3.1.15), this same statement holds with respect to the c–coefficient.

Now, assume that we have reached a point within the sequence of forms where both the a–coefficient and c–coefficient are simultaneously positive. For the sake of convenience, we shall call this form $[a, b, c]$ even if in terms of our original labeling it is not the first form we started with in the sequence. Every form in the sequence of forms after and including this form $[a, b, c]$ will have a positive a–coefficient and a positive c–coefficient. By Definition 3.1.1, two of the three conditions for a reduced form are satisfied, and will remain satisfied as the sequence is carried forward ad infinitum, namely $a > 0$ and $c > 0$. Only the last condition, namely $b > a + c$, may or may not hold at this point. However, we now show that even if this condition does not currently hold, it will be satisfied after a *finite* number of further iterations of the reduction algorithm.

In the next step of reduction, we could have $0 < a' < a$, and in the step after that, we could have $0 < a'' < a'$, but only a *finite* number of such steps with a given a–coefficient strictly less than its predecessor can occur since $a, a', a'', \ldots$ are

all positive integers. This shows that after a *finite* number of steps overall in the sequence of forms, starting from the very beginning form, we will obtain a form $[a, b, c]$ (again, we are allowing for a relabeling) such that $a > 0$ *and* such that $a \leq a'$. We claim that the *next* form $[a', b', c']$ will be reduced! To see this, note that, by (3.1.18),

$$\begin{aligned} 0 \leq a' - a &= a\theta^2 + \theta\sqrt{D} - a \\ &= \theta\sqrt{D} - a\left(1 - \theta^2\right). \end{aligned} \tag{3.1.23}$$

It is clear that $0 < \theta < 1 + \theta$, and since $0 < \sqrt{D}$, we have $\theta\sqrt{D} < (1 + \theta)\sqrt{D}$, which implies that

$$\begin{aligned} \theta\sqrt{D} - a\left(1 - \theta^2\right) &< (1 + \theta)\sqrt{D} - a\left(1 - \theta^2\right) \\ &= (1 + \theta)\left[\sqrt{D} - a(1 - \theta)\right]. \end{aligned} \tag{3.1.24}$$

Combining the inequalities in (3.1.23) and (3.1.24), we have

$$0 < (1 + \theta)\left[\sqrt{D} - a(1 - \theta)\right]. \tag{3.1.25}$$

The expression $(1 + \theta)\left[\sqrt{D} - a(1 - \theta)\right]$ may be rewritten in the form

$$\frac{1 + \theta}{1 - \theta}\left[\sqrt{D}(1 - \theta) - a(1 - \theta)^2\right]. \tag{3.1.26}$$

This is allowable since by (3.1.11) we have $0 < 1 - \theta$, and so we are not dividing by 0. We also note that the quantity $\frac{1+\theta}{1-\theta}$ is a positive real number. The expression in square brackets in (3.1.26), after distribution, is equal to $\sqrt{D} - \theta\sqrt{D} - a + 2a\theta - a\theta^2$, which may be rewritten to obtain

$$\sqrt{D} + 2a\theta - \left(a\theta^2 + \theta\sqrt{D}\right) - a, \tag{3.1.27}$$

which, by (3.1.14′), (3.1.18), and (3.1.15), is equal to $b' - a' - c'$. Combining (3.1.25), (3.1.26), and (3.1.27), we finally obtain

$$0 < \left(\frac{1+\theta}{1-\theta}\right)(b' - a' - c'). \tag{3.1.28}$$

Since $0 < \frac{1+\theta}{1-\theta}$, we see from (3.1.28) that $0 < b' - a' - c'$, or equivalently, $a' + c' < b'$, which proves that $[a', b', c']$ is a reduced form as claimed, since we already know that $0 < a'$ and $0 < a = c'$. We note that this last form $[a', b', c']$, reached after finitely many steps from the starting form, lies in the same equivalence class as the starting form, since each transformation matrix $\mathbf{S}(n)$ lies in the *group* $\mathrm{SL}_2(\mathbb{Z})$. This completes the proof of the first statement in Theorem 3.1.2.

We now prove the second statement in Theorem 3.1.2, that there are only *finitely* many reduced forms of discriminant D. From the first statement proven above, it is clear that there is at least one reduced form of discriminant D, lying in the same equivalence class as the principal form of discriminant D. Let $[a, b, c]$ be a reduced form of discriminant D so that $b^2 - 4ac = D$. We set

$$k = b - 2a, \tag{3.1.29}$$

and note that

$$\begin{aligned} D - k^2 &= b^2 - 4ac - (b - 2a)^2 \\ &= -4ac + 4ab - 4a^2, \end{aligned}$$

so that

$$D - k^2 = 4a(b - a - c) > 0, \tag{3.1.30}$$

with this last inequality holding since $a > 0$ and $b > a + c$ by assumption. We conclude

that

$$k^2 < D, \tag{3.1.31}$$

or equivalently

$$|k| < \sqrt{D}. \tag{3.1.32}$$

From (3.1.30), we see that the positive integer $4a$ divides evenly into the positive integer $D - k^2$, which reads in the usual shorthand

$$4a \mid \left(D - k^2\right). \tag{3.1.33}$$

This implies that

$$k^2 \equiv D \pmod 4. \tag{3.1.34}$$

By (3.1.34), we see that the quantity $\frac{D-k^2}{4}$ is an integer, and (3.1.30) may be written as $a(b-a-c) = \frac{D-k^2}{4}$, so that

$$a \,\Big|\, \frac{D-k^2}{4}. \tag{3.1.35}$$

Now, since $4ac = 4ab - 4a^2 - (D - k^2)$, it follows that

$$c = b - a - \frac{D-k^2}{4a}, \tag{3.1.36}$$

where $\frac{D-k^2}{4a} \in \mathbb{Z}$ by (3.1.33). By (3.1.29), we note that $k + a = b - a$, so that (3.1.36) may be rewritten as

$$c = k + a - \frac{D-k^2}{4a}. \tag{3.1.37}$$

Since $a > 0$, $c > 0$, and $b > a + c$ by assumption, we have $D = b^2 - 4ac < b^2$, and so

$$\sqrt{D} < b. \tag{3.1.38}$$

From (3.1.38), we have

$$\sqrt{D} - k < b - k, \tag{3.1.39}$$

and we note from (3.1.32) that $0 < \sqrt{D} - k$. From (3.1.29), we have $b - k = 2a$, so (3.1.39) may be rewritten as

$$0 < \frac{\sqrt{D} - k}{2} < a. \tag{3.1.40}$$

If we combine (3.1.29), (3.1.32), (3.1.34), (3.1.35), (3.1.37), and (3.1.40), we may finally conclude that *any* reduced form of discriminant D according to Definition 3.1.1 may be put into the shape

$$\left[a, k + 2a, k + a - \frac{D - k^2}{4a}\right], \tag{3.1.41}$$

where a and k satisfy the following four conditions:

$$|k| < \sqrt{D}, \qquad k^2 \equiv D \pmod 4, \qquad a \,\Big|\, \frac{D - k^2}{4}, \qquad 0 < \frac{\sqrt{D} - k}{2} < a. \tag{3.1.42}$$

Only finitely many a's and k's can satisfy all of these conditions, and so there are only *finitely* many reduced forms of discriminant D. This completes the proof of Theorem 3.1.2. □

It is a simple matter to write a computer program that outputs the finite list of all forms having the shape given in (3.1.41) arising from *all* integer values of k and a that satisfy the four conditions in (3.1.42). Let $W(D)$ denote this finite list of forms, and let $R(D)$ denote the set of all reduced forms of discriminant D. We demonstrated above that $R(D) \subseteq W(D)$, confirming that $R(D)$ is a finite set, and now we wish to prove that the opposite inclusion holds: $W(D) \subseteq R(D)$. Once this claim is proven, it is guaranteed that our computer program outputs *precisely* the list of *all* reduced forms of discriminant D.

Theorem 3.1.3. *If $W(D)$ denotes the set of all forms having the shape given in* (3.1.41) *arising from all integer values of k and a that satisfy the four conditions in* (3.1.42)*, and $R(D)$ denotes the set of all reduced forms of discriminant D, then $R(D) = W(D)$.*

Proof. We already know from the proof of Theorem 3.1.2 that $R(D) \subseteq W(D)$, and thus it suffices to prove that $W(D) \subseteq R(D)$. We begin by showing that any form having coefficients as given in (3.1.41) has discriminant D:

$$\begin{aligned} &(k+2a)^2 - 4a\left[k + a - \frac{D-k^2}{4a}\right] \\ &= k^2 + 4ak + 4a^2 - 4ak - 4a^2 + \left(D - k^2\right) \\ &= k^2 + \left(D - k^2\right) = D. \end{aligned} \tag{3.1.43}$$

It follows from the condition $|k| < \sqrt{D}$ that $0 < \sqrt{D} - k$, which in turn says that $0 < \frac{\sqrt{D}-k}{2}$, and so the condition that $\frac{\sqrt{D}-k}{2} < a$ implies immediately that $a > 0$. Given $a \in \mathbb{Z}^+$, b in the form $b = k + 2a$, and some integer c such that $b^2 - 4ac = D$, we note that c is uniquely determined by the equation $c = \frac{b^2-D}{4a}$. The unique solution for c given $a \in \mathbb{Z}^+$ and $b = k + 2a$ was found in (3.1.43) as $c = k + a - \frac{D-k^2}{4a}$, where the condition $a \mid \frac{D-k^2}{4}$ guarantees that $c \in \mathbb{Z}$ (note that $D \equiv k^2 \pmod 4$ implies that $\frac{D-k^2}{4} \in \mathbb{Z}$). For the sum $a + c$, we find that

$$a + c = k + 2a - \frac{D-k^2}{4a} = b - \frac{D-k^2}{4a},$$

or

$$a + c + \frac{D-k^2}{4a} = b. \tag{3.1.44}$$

Since the second and third conditions in (3.1.42) guarantee that $\frac{D-k^2}{4a} \in \mathbb{Z}$, and the first condition (3.1.42) implies that $k^2 < D$, it follows that the integer $\frac{D-k^2}{4a}$ is positive

since $a > 0$. Using the fact that $\frac{D-k^2}{4a}$ is positive in (3.1.44) implies that $a + c < b$. All that remains to prove that $W(D) \subseteq R(D)$ is that $c > 0$.

From the fourth condition in (3.1.42) that $(\sqrt{D} - k)/2 < a$, we have $\sqrt{D} - k < 2a = b - k$, and this implies that $\sqrt{D} < b$, which shows that $b \in \mathbb{Z}^+$. Squaring both sides gives $D < b^2$, or $b^2 - 4ac < b^2$, and so $-4ac < 0$. The only way for this last inequality to hold is if $c \in \mathbb{Z}^+$, since it has already been shown that $a > 0$. We conclude that $W(D) \subseteq R(D)$, which completes the proof of Theorem 3.1.3. □

If we apply our computer program in the case where $D = 20$, we obtain exactly five reduced forms: $[4, 6, 1]$, $[5, 10, 4]$, $[2, 6, 2]$, $[4, 10, 5]$, and $[1, 6, 4]$. For four of these forms, the sum $a + c$ is exactly one less than b, and this is a very common occurrence for reduced forms of a given discriminant. The table below shows this in terms of equation (3.1.44).

Table 3.1.1. Reduced forms of discriminant $D = 20$.

Form	$k = b - 2a$	$(D - k^2)/4a$
$[4, 6, 1]$	$6 - 8 = -2$	$(20 - 4)/16 = 1$
$[5, 10, 4]$	$10 - 10 = 0$	$(20 - 0)/20 = 1$
$[2, 6, 2]$	$6 - 4 = 2$	$(20 - 4)/8 = 2$
$[4, 10, 5]$	$10 - 8 = 2$	$(20 - 4)/16 = 1$
$[1, 6, 4]$	$6 - 2 = 4$	$(20 - 16)/4 = 1$

As another follow-up to the proof of Theorem 3.1.2, we pose the following natural question: "How many steps does it take our reduction algorithm, when applied to the principal form of discriminant D as defined in (3.1.1) and (3.1.2), to obtain

a reduced form in the principal class?" Note that the principal form itself is *never* a reduced form since the c–coefficient in (3.1.1) and (3.1.2) is always negative. This implies that we always need *at least one* step in the reduction algorithm before a reduced form in the principal class is obtained. However, we show that a reduced form in the principal class is obtained in *exactly one* step if our starting point is the principal form. We first illustrate this phenomenon with several concrete examples, and then provide a proof that the reduction algorithm always outputs a reduced form in exactly one step if the starting form is the principal form.

Recall that a single step in the reduction algorithm is always effected by a simple matrix transformation $\mathbf{S}(n)$, which in turn hinges uniquely on the single integer $n \in \mathbb{Z}$. When $\mathbf{S}(n)$ takes the form $[a, b, c]$ to the form $[a', b', c']$, we illustrate this schematically as follows:

$$[a, b, c] \xrightarrow{n} [a', b', c'] .$$

The concrete examples we promised are displayed in Table 3.1.2 below.

In the table below, we see that each principal form is taken to a reduced form in exactly one step. To prove this in general, we return to the proof of Theorem 3.1.2, where we showed that if $[a, b, c]$ is a form with $a > 0$, and if $\mathbf{S}(n)$ takes this form to $[a', b', c']$ with $a \leq a'$, then the form $[a', b', c']$ will be reduced. We also showed that if $a > 0$, then $a' > 0$. Now, the principal form of discriminant D always has $a = 1$. Therefore, $a' \in \mathbb{Z}^+$, and we must have $1 = a \leq a'$. We see immediately from above that $[a', b', c']$ is a reduced form, obtained in exactly one step of the reduction algorithm starting from the principal form.

Table 3.1.2. Reduction of principal forms.

D	Principal Form to Reduced Form
5	$[1,1,-1] \xrightarrow{2} [1,3,1]$
8	$[1,0,-2] \xrightarrow{2} [2,4,1]$
12	$[1,0,-3] \xrightarrow{2} [1,4,1]$
13	$[1,1,-3] \xrightarrow{3} [3,5,1]$
17	$[1,1,-4] \xrightarrow{3} [2,5,1]$
20	$[1,0,-5] \xrightarrow{3} [4,6,1]$
21	$[1,1,-5] \xrightarrow{3} [1,5,1]$
24	$[1,0,-6] \xrightarrow{3} [3,6,1]$
28	$[1,0,-7] \xrightarrow{3} [2,6,1]$

Given a form $[a,b,c]$ of discriminant D, there is a uniquely defined corresponding transformation matrix $\mathbf{S}(n)$ that is employed in our reduction algorithm to take us to the new form $[a',b',c']$, also of discriminant D. We are naturally led to pose the following question: "If $[a,b,c]$ is a reduced form, is there anything that can be said of the form $[a',b',c']$?" The elegant answer to this question is given by the following theorem.

Theorem 3.1.4. *Let $D \in \mathbb{Z}^+$ be a fixed discriminant. If $[a,b,c]$ is a reduced form of discriminant D, then $[a',b',c']$ is also a reduced form of discriminant D.*

Remark. As we saw in the proof of Theorem 3.1.2, starting with an *arbitrary* form $[a,b,c]$ of discriminant D, not necessarily reduced, our reduction algorithm provides us with a uniquely determined infinite sequence of forms $[a,b,c]$, $[a',b',c']$, $[a'',b'',c'']$, $\ldots$, $\left[a^{(j)},b^{(j)},c^{(j)}\right]$, $\ldots$ for $j = 0,1,2,3,\ldots$. Theorem 3.1.2 says that for

some integer $m \in \mathbb{Z}^{\geq 0}$, the form $\left[a^{(m)}, b^{(m)}, c^{(m)}\right]$ will be reduced. If we specify $m \in \mathbb{Z}^{\geq 0}$ to be the first such integer for which $\left[a^{(m)}, b^{(m)}, c^{(m)}\right]$ is a reduced form, Theorem 3.1.4 states that for *every* integer $j \geq m$, each form $\left[a^{(j)}, b^{(j)}, c^{(j)}\right]$ will be reduced as well.

As a preliminary step towards proving Theorem 3.1.4, we first prove the following lemma, which is a result of interest in itself.

Lemma 3.1.5. *A quadratic form*

$$f(x, y) = ax^2 + bxy + cy^2 := [a, b, c] \tag{3.1.45}$$

of discriminant $b^2 - 4ac = D \in \mathbb{Z}^+$ (recall that $a \neq 0$ and $c \neq 0$ for any form whose discriminant is as specified in Theorem 3.1.4) is reduced if and only if the two roots of the quadratic equation $f(x, -1) = ax^2 - bx + c = 0$, namely $(b \pm \sqrt{D})/2a$, satisfy the following inequalities:

$$0 < \frac{b - \sqrt{D}}{2a} < 1 < \frac{b + \sqrt{D}}{2a}. \tag{3.1.46}$$

Proof. We first prove the forward direction. Assume that $[a, b, c]$ is a reduced form, namely $a > 0$, $c > 0$, and $b > a + c$. From equations (3.1.29) and (3.1.32), the quantity $k = b - 2a$ satisfies the inequality $|k| < \sqrt{D}$, or

$$|b - 2a| = |2a - b| < \sqrt{D}, \tag{3.1.47}$$

which may be rewritten as

$$-\sqrt{D} < 2a - b < \sqrt{D}. \tag{3.1.48}$$

Adding b to both sides gives

$$b - \sqrt{D} < 2a < b + \sqrt{D}. \tag{3.1.49}$$

From (3.1.38), we know that $\sqrt{D} < b$, and so (3.1.49) may be extended to

$$0 < b - \sqrt{D} < 2a < b + \sqrt{D}. \tag{3.1.50}$$

Since $2a > 0$, we may divide every expression in (3.1.50) by $2a$ to obtain

$$0 < \frac{b - \sqrt{D}}{2a} < 1 < \frac{b + \sqrt{D}}{2a}. \tag{3.1.51}$$

This completes the proof of the forward direction.

Next, we prove the reverse direction. Assume that the inequalities in (3.1.51) hold for the two roots of the quadratic equation $f(x, -1) = 0$. From (3.1.51), we have

$$0 < \frac{b + \sqrt{D}}{2a} - \frac{b - \sqrt{D}}{2a} = \frac{\sqrt{D}}{a}. \tag{3.1.52}$$

Since $\sqrt{D} > 0$, we see that we must also have $a > 0$ by (3.1.52). Both roots are assumed to be positive, and so their product must be positive as well:

$$0 < \left(\frac{b + \sqrt{D}}{2a}\right)\left(\frac{b - \sqrt{D}}{2a}\right) = \frac{b^2 - D}{4a^2} = \frac{4ac}{4a^2} = \frac{c}{a}. \tag{3.1.53}$$

Since $a > 0$, we see that we must also have $c > 0$ by (3.1.53). Since $2a > 0$, the first inequality in (3.1.51) implies that $0 < b - \sqrt{D}$, or that $\sqrt{D} < b$, which shows that $b > 0$. However, it remains to prove a stronger inequality, namely $b > a + c$. Multiplying the inequalities in (3.1.51) through by the positive quantity $2a$, we obtain $b - \sqrt{D} < 2a < b + \sqrt{D}$, or $-\sqrt{D} < 2a - b < \sqrt{D}$, which implies that $|2a - b| = |b - 2a| < \sqrt{D}$. Setting $k = b - 2a$ and squaring gives $|k|^2 = k^2 < D$, or

$$0 < D - k^2 = b^2 - 4ac - (b - 2a)^2 = 4a(b - a - c). \tag{3.1.54}$$

Since $4a > 0$, (3.1.54) implies that $b - a - c > 0$, or $b > a + c$. This completes the proof of Lemma 3.1.5. □

Proof of Theorem 3.1.4. Assume that $[a, b, c]$ is a reduced form of discriminant D. The value of n in the matrix $\mathbf{S}(n)$ which takes us to $[a', b', c']$ is given by $n = \left\lceil \frac{b+\sqrt{D}}{2a} \right\rceil$. From the forward direction of Lemma 3.1.5, we see that n satisfies the inequality $n \geq 2$. Since $0 < (b - \sqrt{D})/2a < 1$ by (3.1.51), the inequality

$$1 < n - \frac{(b - \sqrt{D})}{2a} \tag{3.1.55}$$

holds based upon the distance between the two numbers on the right hand side of (3.1.55) as illustrated in the picture below:

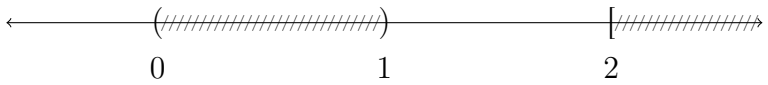

Figure 3.1. Proof by picture

From (3.1.18) and (3.1.11), we already noted that if $a > 0$, then $a' > 0$. By (3.1.6), we also have $c' = a > 0$, and so it only remains to show that $b' > a' + c'$ in order to confirm that $[a', b', c']$ is also a reduced form.

From the definition of the irrational number θ in (3.1.10), we have $n - \theta = (b + \sqrt{D})/2a$ and so

$$\frac{\sqrt{D}}{a} = \frac{b + \sqrt{D}}{2a} - \frac{b - \sqrt{D}}{2a} = n - \theta - \frac{b - \sqrt{D}}{2a}. \tag{3.1.56}$$

From (3.1.11), we have

$$0 < 1 - \theta. \tag{3.1.57}$$

If we subtract θ from both sides of (3.1.55), and make use of both (3.1.56) and (3.1.57), we obtain

$$0 < 1 - \theta < \frac{\sqrt{D}}{a}. \tag{3.1.58}$$

Since $a > 0$, this implies in turn that $a(1-\theta) < \sqrt{D}$, or

$$0 < \sqrt{D} - a(1-\theta). \tag{3.1.59}$$

By (3.1.26) and (3.1.27), we recall that

$$\left[\sqrt{D}(1-\theta) - a(1-\theta)^2\right] = b' - a' - c'. \tag{3.1.60}$$

Factoring the left hand side of (3.1.60) as $(1-\theta)[\sqrt{D} - a(1-\theta)]$, and using the inequalities in (3.1.57) and (3.1.59), we conclude from (3.1.60) that $b' - a' - c' > 0$, or $b' > a' + c'$. This completes the proof of Theorem 3.1.4. □

As an illustration of Theorem 3.1.4, we consider the set of all reduced forms of discriminant $D = 28$. We have

$$R(28) = \{[6,10,3],\ [7,14,6],\ [2,6,1],\ [3,8,3],\ [6,14,7],\ [1,6,2],\ [3,10,6]\}.$$

If we start with the reduced form $[1,6,2]$, Theorem 3.1.4 guarantees that the first step of the reduction algorithm applied to this form takes us to exactly one of the forms in the set $R(28)$ above. We find that $[1,6,2] \xrightarrow{6} [2,6,1]$. In turn, the reduced form $[2,6,1]$ must be sent to a form in $R(28)$, and we find that $[2,6,1] \xrightarrow{3} [1,6,2]$. If we consider the infinite sequence of forms generated by the reduction algorithm, starting with $[1,6,2]$, we obtain the following "purely periodic" repeating pattern:

$$[1,6,2] \xrightarrow{6} [2,6,1] \xrightarrow{3} [1,6,2] \xrightarrow{6} [2,6,1] \xrightarrow{3} [1,6,2] \xrightarrow{6} [2,6,1] \xrightarrow{3} \ldots. \tag{3.1.61}$$

Our goal is to show that such a purely periodic repeating pattern *always* occurs when the reduction algorithm is applied to a starting form f that is *reduced*. Since the repeating pattern in (3.1.61) is the norm when we apply the reduction algorithm to a reduced form, we find it useful to introduce the following diagram that summarizes

compactly the overall pattern:

$$[1,6,2] \xrightarrow{6} [2,6,\underbrace{1}_{3}]. \tag{3.1.62}$$

The curved arrow at the end of this diagram indicates that the form $[2,6,1]$ is sent back to the starting form $[1,6,2]$, and this cycle of length 2 repeats over and over as the reduction algorithm is applied indefinitely.

Since the pattern in (3.1.62) is the norm for reduced forms, we have a special name for it. We call such a pattern a "cycle of reduced forms." There are five other reduced forms of discriminant $D = 28$, and they all form a separate cycle:

$$[3,8,3] \xrightarrow{3} [6,10,3] \xrightarrow{2} [7,14,6] \xrightarrow{2} [6,14,7] \xrightarrow{2} [3,10,\underbrace{6}_{3}]. \tag{3.1.63}$$

It is clear that all reduced forms in a given cycle lie in the same equivalence class of forms under the action of $\mathrm{SL}_2(\mathbb{Z})$. For example, the two forms in (3.1.62) both lie in the principal class. However, we pose the question: Is the same true for the forms in (3.1.63)? Do they all lie in the principal class as well, or do they all lie in a separate class from the principal class? In the first case, all forms of discriminant $D = 28$ would fall into a single class, and in the second case, there would be exactly two classes of forms. The fact that the second case holds is guaranteed by Theorem 3.1.11, which is stated and proved later in this section.

In order to prove that the reduced forms of a given discriminant are naturally aligned into disjoint cycles, we find it useful to introduce some new terminology. Let $D \in \mathbb{Z}^+$ be a discriminant as previously defined, and let $f = [a,b,c]$ be an arbitrary form of discriminant D. Using our reduction algorithm, we obtain the new form $f' = [a',b',c']$ when we apply $\mathbf{S}(n)$ to $f = [a,b,c]$, where $n = \left\lceil \frac{b+\sqrt{D}}{2a} \right\rceil$. Given f,

the form f', also of discriminant D, is uniquely defined, and therefore we have a well-defined map, denoted by $I(f) = f'$. Theorem 3.1.4 guarantees that if f is a reduced form, then $I(f)$ is also a reduced form.

Definition 3.1.6. Given an arbitrary form f of discriminant D, the uniquely defined form $f' = I(f)$ is called the *right neighbor* of f. The shorthand used previously was $f \xrightarrow{n} f'$, and the "right neighbor" terminology reflects this diagrammatic picture.

It is convenient to also introduce the notion of the "left neighbor" of f, which we now define. Given an arbitrary form $f = [a, b, c]$ of discriminant $D \in \mathbb{Z}^+$, we set

$$m = \left\lceil \frac{b + \sqrt{D}}{2c} \right\rceil \tag{3.1.64}$$

(recall that $c \neq 0$ by our choice of D). The *left neighbor* of f, which we denote by $f^v := J(f)$, is obtained from f by applying the transformation matrix

$$\mathbf{T}(m) = \begin{pmatrix} 0 & -1 \\ 1 & m \end{pmatrix} \tag{3.1.65}$$

with $\det \mathbf{T}(m) = 1$ to f, where m is given as in (3.1.64). The form $f^v = [a^v, b^v, c^v]$ we obtain has coefficients given by

$$a^v = c \tag{3.1.66}$$

$$b^v = -b + 2cm \tag{3.1.67}$$

$$c^v = a - bm + cm^2, \tag{3.1.68}$$

where $(b^v)^2 - 4a^v c^v = D$. Given f, the form f^v is clearly uniquely defined. The following theorem, which mimics Theorem 3.1.4, should come as no surprise given our past experience.

Theorem 3.1.7. *Let $D \in \mathbb{Z}^+$ be a fixed discriminant. If $f = [a, b, c]$ is a reduced form of discriminant D, then $f^v = [a^v, b^v, c^v]$ is also a reduced form of discriminant D.*

Proof. Let $f = [a, b, c]$ be a reduced form of discriminant D, and set $\ell = b - 2c$. Note that

$$\begin{aligned} D - \ell^2 &= b^2 - 4ac - (b - 2c)^2 \\ &= b^2 - 4ac - \left[b^2 - 4bc + 4c^2\right] \\ &= 4c(b - a - c) > 0, \end{aligned}$$

with the last inequality holding since $c > 0$ and $b > a + c$ by assumption. We conclude that $\ell^2 < D$, or equivalently

$$|\ell| < \sqrt{D}. \tag{3.1.69}$$

Therefore, $|b - 2c| = |2c - b| < \sqrt{D}$, which may be rewritten as $-\sqrt{D} < 2c - b < \sqrt{D}$, or equivalently

$$b - \sqrt{D} < 2c < b + \sqrt{D}. \tag{3.1.70}$$

From (3.1.38), we know that $\sqrt{D} < b$, and so (3.1.70) may be extended to

$$0 < b - \sqrt{D} < 2c < b + \sqrt{D}. \tag{3.1.71}$$

Since $2c > 0$, we may divide (3.1.71) throughout by $2c$ to obtain

$$0 < \frac{b - \sqrt{D}}{2c} < 1 < \frac{b + \sqrt{D}}{2c}. \tag{3.1.72}$$

Given the inherent symmetry in Definition 3.1.1 for a reduced form, we note that if $[a, b, c]$ is a reduced form of discriminant D, so is $[c, b, a]$. This observation gives us a direct deduction of the inequalities in (3.1.72) from those in (3.1.51). Comparing

(3.1.64) and (3.1.72), we conclude that

$$m \geq 2 \tag{3.1.73}$$

in (3.1.64) when f is a reduced form. We also note that the number φ defined by

$$\varphi = m - \frac{b + \sqrt{D}}{2c} \tag{3.1.74}$$

is an irrational number that satisfies the inequalities

$$0 < \varphi < 1. \tag{3.1.75}$$

We may rewrite (3.1.74) in the form $b + \sqrt{D} = 2cm - 2c\varphi$, or

$$-b + 2cm = \sqrt{D} + 2c\varphi. \tag{3.1.76}$$

By use of (3.1.76), we can rewrite (3.1.67) as

$$b^v = \sqrt{D} + 2c\varphi. \tag{3.1.77}$$

We may also express c^v compactly in terms of φ by using (3.1.74) to obtain

$$m^2 = \left(\frac{b + \sqrt{D}}{2c}\right)^2 + \frac{(b + \sqrt{D})\varphi}{c} + \varphi^2,$$

or

$$m^2 = \frac{b^2 + 2b\sqrt{D} + D}{4c^2} + \frac{(b + \sqrt{D})\varphi}{c} + \varphi^2. \tag{3.1.78}$$

We may now rewrite (3.1.68), using (3.1.78) and (3.1.74), in the form

$$c^v = \frac{b^2 + 2b\sqrt{D} + D}{4c} + (b + \sqrt{D})\varphi + c\varphi^2 - b\left(\frac{b + \sqrt{D}}{2c}\right) - b\varphi + a,$$

or

$$c^v = c\varphi^2 + \varphi\sqrt{D}, \tag{3.1.79}$$

since

$$\frac{b^2 + 2b\sqrt{D} + D - 2b^2 - 2b\sqrt{D} + 4ac}{4c} = 0$$

follows from $D = b^2 - 4ac$. From (3.1.75) and (3.1.79), we note that if $c > 0$, then $c^v > 0$. Since $f = [a, b, c]$ is a reduced form by assumption, we have $c > 0$ and so $c^v > 0$ in this case. Since $a^v = c$ by (3.1.66), we also have $a^v > 0$ in this case. To prove that f^v is a reduced form, it remains to show that $b^v > a^v + c^v$.

Combining (3.1.72) and (3.1.73) and using the same "proof by picture" as illustrated in Figure 3.1, we find that

$$1 < m - \frac{b - \sqrt{D}}{2c}. \tag{3.1.80}$$

From (3.1.74), we have $m - \varphi = (b + \sqrt{D})/2c$, and so

$$\frac{\sqrt{D}}{c} = \frac{b + \sqrt{D}}{2c} - \frac{b - \sqrt{D}}{2c} = m - \varphi - \frac{b - \sqrt{D}}{2c}. \tag{3.1.81}$$

From (3.1.75), we have

$$0 < 1 - \varphi, \tag{3.1.82}$$

and if we subtract φ from both sides of (3.1.80) and make use of (3.1.81), we obtain

$$0 < 1 - \varphi < \frac{\sqrt{D}}{c}. \tag{3.1.83}$$

Since $c > 0$, this implies in turn that $c(1 - \varphi) < \sqrt{D}$, or

$$0 < \sqrt{D} - c(1 - \varphi). \tag{3.1.84}$$

If we combine (3.1.77), (3.1.66), and (3.1.79), we obtain

$$\begin{aligned} b^v - a^v - c^v &= \sqrt{D} + 2c\varphi - c - \left(c\varphi^2 + \varphi\sqrt{D}\right) \\ &= \sqrt{D} - \varphi\sqrt{D} - c + 2c\varphi - c\varphi^2 \\ &= \left[\sqrt{D}(1-\varphi) - c(1-\varphi)^2\right] \\ &= (1-\varphi)\left[\sqrt{D} - c(1-\varphi)\right]. \end{aligned}$$

Using the inequalities in (3.1.82) and (3.1.84), we conclude that $b^v - a^v - c^v > 0$, or $b^v > a^v + c^v$. This completes the proof of Theorem 3.1.7. □

If $D \in \mathbb{Z}^+$ is a fixed discriminant as previously defined, we let $R(D)$ denote the set of all reduced forms of discriminant D. By Theorem 3.1.2, we know that $R(D)$ is a finite nonempty set. If we restrict the map I sending a form f to its right neighbor f' to the set $R(D)$, we have

$$I : R(D) \to R(D). \tag{3.1.85}$$

This follows from Theorem 3.1.4, which states that $I(f) \in R(D)$ if $f \in R(D)$. This restricted map has several interesting properties, the first of which is given by

Proposition 3.1.8. *The restricted map $I : R(D) \to R(D)$ is a bijection.*

In order to prove Proposition 3.1.8, we make use of the left neighbor map J defined immediately after Definition 3.1.6. More specifically, we make use of the restriction of J to $R(D)$. By Theorem 3.1.7, we know that the restricted map

$$J : R(D) \to R(D) \tag{3.1.86}$$

is well-defined since $J(f) \in R(D)$, assuming that $f \in R(D)$. We also show that the restricted map defined by (3.1.86) is a bijection, and is in fact the inverse of the map

defined in (3.1.85).

The proof of Proposition 3.1.8 depends upon the lemma below, which is of interest in its own right.

Lemma 3.1.9. *Assume that* $f = [a, b, c] \in R(D)$ *and* $n = \left\lceil \frac{b+\sqrt{D}}{2a} \right\rceil$. *Also, assume that* $I(f) = f_1 = [a_1, b_1, c_1]$. *If we set* $m = \left\lceil \frac{b_1+\sqrt{D}}{2c_1} \right\rceil$, *then* $n = m$.

Remark. By (3.1.15), we know that $c_1 = a$ and so $m = \left\lceil \frac{b_1+\sqrt{D}}{2a} \right\rceil$. Even so, it is not obvious that $n = m$ since we generally have $b \neq b_1$ (recall from (3.1.14) that $b_1 = -b + 2an$). This is best illustrated by an example. Suppose that $D = 33$ and $f = [4, 9, 3]$, which is clearly reduced. We find that $(9 + \sqrt{33})/(2 \cdot 4) = 1.843\ldots$, and so $n = 2$. If we apply $\mathbf{S}(2)$ to f, we obtain $f_1 = [1, 7, 4]$. We find that $(7 + \sqrt{33})/(2 \cdot 4) = 1.593\ldots$, and so $m = 2$, in accordance with Lemma 3.1.9. But the decimal answers to which we are applying the ceiling operation are distinct from each other, and it is not immediately clear why one number could not be just below 2, say, and the other just above 2, producing a situation where $m \neq n$.

Proof. Since f is a reduced form, we know that $a > 0$, and that $2 \leq n$ by (3.1.51). From (3.1.14), we have

$$-b + 2an = b_1. \tag{3.1.87}$$

We know that f_1 is a reduced form by Theorem 3.1.4. By the definition of the ceiling of a number, the integer $m \geq 2$ (see (3.1.73)) is uniquely determined by the following inequalities (recall that $c_1 = a$):

$$m - 1 < \frac{b_1 + \sqrt{D}}{2a} < m. \tag{3.1.88}$$

If we multiply every term in (3.1.88) through by the positive integer $2a$, we obtain

$$2a(m-1) < b_1 + \sqrt{D} < 2am, \tag{3.1.89}$$

or

$$-b_1 + 2a(m-1) < \sqrt{D} < -b_1 + 2am. \tag{3.1.90}$$

The inequalities in (3.1.90) show that m is the smallest positive integer such that the expression $-b_1 + 2az$ is greater than $\sqrt{D}$ when m is substituted in for z. By (3.1.87), we have

$$-b_1 + 2an = b, \tag{3.1.91}$$

and since $\sqrt{D} < b$ by (3.1.38), we have $\sqrt{D} < -b_1 + 2an$. Since n is a positive integer such that the expression $-b_1 + 2az$ is greater than $\sqrt{D}$ when n is substituted in for z, and m is the *smallest* positive integer with this property, we conclude that

$$m \leq n. \tag{3.1.92}$$

Subtracting $2a$ from both sides of (3.1.91) gives

$$-b_1 + 2a(n-1) = b - 2a,$$

and we know that $b - 2a < \sqrt{D}$ by (3.1.47). This shows that if we set z equal to any positive integer less than n, the expression $-b_1 + 2az$ is less than $\sqrt{D}$. Since $\sqrt{D} < -b_1 + 2am$ by (3.1.90), we can not have $m < n$. Combined with (3.1.92), we conclude that $m = n$. This completes the proof of Lemma 3.1.9. □

Proof of Proposition 3.1.8. In terms of the notation used in Lemma 3.1.9, we claim that $J(f_1) = f$. By Lemma 3.1.9, we have $m = n$, and so the form $J(f_1)$ is obtained from f_1 by applying the transformation matrix $\mathbf{T}(n)$ to f_1. We have $f \cdot \mathbf{S}(n) = f_1$

and $f_1 \cdot \mathbf{T}(n) = J(f_1)$, and so $f \cdot \mathbf{S}(n) \cdot \mathbf{T}(n) = J(f_1)$. This shows that the matrix product $\mathbf{S}(n) \cdot \mathbf{T}(n)$ takes f to $J(f_1)$, and an easy calculation shows that

$$\mathbf{S}(n) \cdot \mathbf{T}(n) = \begin{pmatrix} 1 & 0 \\ \\ 0 & 1 \end{pmatrix},$$

the 2×2 identity matrix, so indeed $J(f_1) = f$. This shows that for any given $f \in R(D)$, we have $J(I(f)) = f$, which implies that the restricted map $I : R(D) \to R(D)$ is one-to-one. To see this, assume that $f, g \in R(D)$, and that $I(f) = I(g)$. We then have $J(I(f)) = J(I(g))$, or $f = g$. Recalling that a one-to-one map from a *finite* set to *itself* is always surjective, we conclude that the restricted map $I : R(D) \to R(D)$ is a bijection, which completes the proof of Proposition 3.1.8. □

Proposition 3.1.8 is crucial to showing how (and why!) the set of reduced forms $R(D)$ is partitioned into disjoint cycles under the action of the restricted map $I : R(D) \to R(D)$. Assume that we apply our reduction algorithm to a given *reduced* form $f = f_0$. If $I(f_0) = f_0$, then f_0 will be in a cycle that consists only of itself, and there is nothing more to say. Assume on the other hand that $I(f_0) = f_1 \neq f_0$, and that a maximum number of t steps in the reduction algorithm can be made in such a way that all of the forms appearing in the beginning of the sequence are mutually distinct from each other. In that case, the sequence of forms would begin as follows:

$$f_0 \to f_1 \to \cdots \to f_t, \quad t \geq 1, \tag{3.1.93}$$

where $f_0, f_1, \ldots, f_t$ are all distinct from each other, but $f_{t+1} \in \{f_0, f_1, \ldots, f_t\}$. By Theorem 3.1.4, and the fact that $R(D)$ is a finite set, we know that there exists a uniquely defined *positive* integer $t \in \mathbb{Z}^+$ with exactly this property (again, we are assuming that $I(f_0) \neq f_0$). We claim that in this case, we must have $f_{t+1} = f_0$. To

prove this claim, assume on the contrary that $I(f_t) = f_{t+1} = f_j$ for some $j \in \mathbb{Z}^+$ with $1 \leq j \leq t$. Note that we also have $I(f_{j-1}) = f_j$, and by assumption f_{j-1} and f_t are distinct from each other since $0 \leq j-1 \leq t-1$. This contradicts the fact that the restricted map $I : R(D) \to R(D)$ is one-to-one by Proposition 3.1.8, and so $f_{t+1} = f_0$, establishing our claim. Having established that $I(f_t) = f_0$ here means that we have attained the goal enunciated just below equation (3.1.61), showing that a purely periodic repeating pattern *always* occurs when the reduction algorithm is applied to a starting form $f = f_0$ that is *reduced*. Again, we use the following diagram to compactly summarize this pattern:

$$f_0 \xrightarrow{n_0} f_1 \xrightarrow{n_1} \cdots \xrightarrow{n_{t-1}} \underbrace{f_t}_{n_t}. \tag{3.1.94}$$

If $t \geq 1$, then f_{j+1} is obtained from f_j by the application of $\mathbf{S}(n_j)$ to f_j for each j with $0 \leq j \leq t-1$. As a special case of (3.1.94), it is entirely possible to have just a single reduced form f_0 in the diagram, which we denote by

$$\underbrace{f_0}_{n_0}. \tag{3.1.95}$$

Again, we call the pattern displayed in (3.1.94) a "cycle of reduced forms," where by construction all $t+1$ of the forms appearing in (3.1.94) are assumed to be *distinct* from each other.

The question still remains as to why these cycles of reduced forms create a partitioning of the set of reduced forms $R(D)$. Any mathematical partition can be established via an equivalence relation.

Definition 3.1.10. We write $f \approx g$ if and only if $f, g \in R(D)$ are such that g lies in the cycle generated from the starting form f.

In order to show that this definition gives an equivalence relation on the forms in $R(D)$, we must prove that the following three properties hold assuming that f, g, and h are arbitrarily chosen forms in $R(D)$:

i) We have $f \approx f$ (reflexive property).

ii) If $f \approx g$, then $g \approx f$ (symmetric property).

iii) If $f \approx g$ and $g \approx h$, $f \approx h$ (transitive property).

Proof. i) We just proved (and (3.1.94) gives a pictorial representation) that f lies in the cycle obtained from applying the reduction algorithm to the starting reduced form f since the cycle always *returns back* to f. ii) The assumption that $f \approx g$ is illustrated in the following diagram:

$$f \to \cdots \to g \to \underset{\hookleftarrow}{\cdots}.$$

What is crucial here is that we *always* cycle back to the starting form f and so f lies in the cycle generated from the starting form g. iii) This proof is also easy to visualize by using diagrams. Given our assumptions, we could have a diagram like this:

$$f \to \cdots \to g \to \cdots \to h \to \underset{\hookleftarrow}{\cdots},$$

or like this:

$$f \to \cdots \to h \to \cdots \to g \to \underset{\hookleftarrow}{\cdots}.$$

Either way, it is clear that $f \approx h$. This completes the proof that the relationship given in Definition 3.1.10 provides an equivalence relation on the forms in $R(D)$. □

From an algorithmic standpoint, we may carve out the partitioning of $R(D)$ into disjoint cycles as follows. Start with any given form $f \in R(D)$ and generate

the corresponding cycle of reduced forms using f as a starting form. If all forms in $R(D)$ appear in this one cycle, then we are done. Otherwise, take a reduced form g not in this cycle and generate the corresponding cycle of reduced forms using g as a starting form. Since Definition 3.1.10 gives an equivalence relation on the forms in $R(D)$, these two cycles are disjoint. If all the forms in $R(D)$ are now accounted for, we are done. Otherwise, we continue in this way until we have completely exhausted the *finite* nonempty set $R(D)$ and have obtained a representation of $R(D)$ as a union of mutually disjoint cycles of forms.

We next move on to a very important theorem, which boils down to say that the number of disjoint cycles that the forms in $R(D)$ fall into is exactly equal to $t(D)$, this being the number of equivalence classes of forms of discriminant D under the action of $\mathrm{SL}_2(\mathbb{Z})$.

Theorem 3.1.11. *Let $D \in \mathbb{Z}^+$ be a fixed discriminant. Assume that $f = [a, b, c]$ is a reduced form of discriminant D, and let*

$$f = f_0, f_1, f_2, \ldots \tag{3.1.96}$$

be the infinite sequence of reduced forms generated when the reduction algorithm is applied to the form f as the starting form. If f^* is any reduced form of discriminant D lying in the same equivalence class as f, then f^* must appear in the sequence (3.1.96).

Remark. If g is any form that appears in the sequence (3.1.96), then we know that g is a reduced form by Theorem 3.1.4, and that g lies in the same equivalence class as f under the action of $\mathrm{SL}_2(\mathbb{Z})$. Theorem 3.1.11 tells us that the converse result holds; namely, if g is a reduced form lying in the same equivalence class as f, then g is one of the forms appearing in the sequence (3.1.96). Therefore, if g is reduced and it does not

appear in the sequence (3.1.96), then g is not in the same equivalence class as f. This implies that each equivalence class of forms under the action of $\mathrm{SL}_2(\mathbb{Z})$ corresponds to exactly one cycle of forms; so if we compute how many disjoint cycles of reduced forms there are of discriminant D, then we have at the same time computed the number of equivalence classes of forms $t(D)$. In our earlier example involving reduced forms of discriminant $D = 28$, we saw that the reduced form $[3, 8, 3]$ does not appear in the cycle (3.1.62), and therefore $[3, 8, 3]$ is not in the principal class by Theorem 3.1.11. Given that all of the reduced forms of discriminant $D = 28$ are accounted for in the two cycles (3.1.62) and (3.1.63), we see that there are exactly two classes of forms when $D = 28$, namely, $t(28) = 2$.

Proof of Theorem 3.1.11. By assumption, $f = [a, b, c] = f_0$ is a given reduced form of discriminant D. We are also assuming that $f^* = [a^*, b^*, c^*]$ is a reduced form of discriminant D lying in the *same* equivalence class as f. By definition, this means that there exists a matrix

$$\mathbf{A} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \tag{3.1.97}$$

such that $f \cdot \mathbf{A} = f^*$. From page 173 of [5], we have

$$a^* = f(r, t) = ar^2 + brt + ct^2 \tag{3.1.98}$$

$$b^* = 2ars + b(ru + st) + 2ctu \tag{3.1.99}$$

$$c^* = f(s, u) = as^2 + bsu + cu^2. \tag{3.1.100}$$

We also claim that

$$a^* + c^* - b^* = f(r - s, t - u). \tag{3.1.101}$$

To see this, we note that

$$\begin{aligned} f(r-s,t-u) &= a(r-s)^2 + b(r-s)(t-u) + c(t-u)^2 \\ &= a\left[r^2 - 2rs + s^2\right] + b\left[rt - ru - st + su\right] + c\left[t^2 - 2tu + u^2\right] \\ &= f(r,t) + f(s,u) - 2ars - b(ru+st) - 2ctu \\ &= a^* + c^* - b^* \end{aligned}$$

by (3.1.98), (3.1.99), and (3.1.100). Since f^* is a reduced form by assumption, we have $a^* > 0$, $c^* > 0$, and

$$a^* + c^* - b^* < 0. \tag{3.1.102}$$

We claim that we must have

$$t \neq u. \tag{3.1.103}$$

Assume on the contrary that $t = u$. This implies that

$$\begin{aligned} f(r-s,t-u) &= f(r-s,0) \\ &= a(r-s)^2 \\ &\geq 0, \end{aligned}$$

since $a > 0$ (recall that f is reduced). On the other hand, (3.1.101) and (3.1.102) combine to say that $f(r-s,t-u) < 0$, which shows the contradiction. Thus, we must have $t \neq u$.

From (3.1.103), we see that we must have either $t > u$ or $t < u$. Without loss of generality, we now show that our situation may be arranged in such a way that

$$t < u. \tag{3.1.104}$$

If (3.1.104) already holds within the matrix $\mathbf{A}$ in (3.1.97), then we are done. If we have

$$t > u \tag{3.1.105}$$

in (3.1.97), then we replace the matrix $\mathbf{A}$ by

$$-\mathbf{A} = \begin{pmatrix} -r & -s \\ -t & -u \end{pmatrix}.$$

This causes no trouble since $-\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$ as well, and it easy to check that $-\mathbf{A}$ takes f to f^* just as the matrix $\mathbf{A}$ does. If (3.1.105) holds, then we have $-t < -u$, which shows that (3.1.104) holds with respect to the replacement matrix $-\mathbf{A}$. We therefore assume from this point forward that (3.1.104) holds with respect to the matrix $\mathbf{A}$ in (3.1.97).

From this point on, our proof of Theorem 3.1.11 breaks down into three cases, depending upon the value of t.

Case I: $t = 0$. Recall that $ru - st = 1$ since $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$. If $t = 0$, then $ru = 1$, which implies that either $r = u = 1$ or $r = u = -1$, since r and u are both integers. By (3.1.104), we have $0 < u$, and so $r = u = 1$. We have $t = 0$, and we claim that we must have $s = 0$ in this case as well. Since $c^* > 0$ and $0 > a^* + c^* - b^*$, we have

$$c^* = f(s,u) = f(s,1) > 0 > f(r-s, t-u) = f(1-s,-1) = f(s-1,1).$$

The key inequalities here may be rewritten more explicitly as

$$as^2 + bs + c > 0 > a(s-1)^2 + b(s-1) + c. \tag{3.1.106}$$

If we set $s = 0$ in (3.1.106), we obtain

$$c > 0 > a - b + c, \tag{3.1.107}$$

and these two inequalities are consistent with our assumption that f is a reduced form. If we can show that *at most* one integer value for s can be plugged in such that the inequalities in (3.1.106) are satisfied, then we may conclude that $s = 0$. The easiest way to show this is to consider the graph of the quadratic polynomial $\phi(x) = f(x, 1) = ax^2 + bx + c$. Since $b^2 - 4ac > 0$ by assumption, and $a > 0$, this graph is a parabola that opens upward and crosses the x-axis at two distinct points, say at x_1 and x_2, with $x_1 < x_2$. Note that (3.1.106) says that the integer s must satisfy

$$\phi(s - 1) < 0 < \phi(s). \tag{3.1.108}$$

Since $s - 1 < s$, this can only happen if $s - 1 < x_2 < s$, and this in turn can only happen for at most one *integer* value of s, namely $s = \lceil x_2 \rceil$, where x_2 is an irrational number since $D = b^2 - 4ac$ is not a perfect square. Since $r = u = 1$ and $s = t = 0$ in this case, we conclude that

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and so $f = f^*$ by (3.1.97), which implies that f^* lies in the sequence (3.1.96).

Case II: $t < 0$. We have $f \cdot \mathbf{A} = f^*$, with certain assumptions in place concerning the matrix $\mathbf{A}$. The first step of the reduction algorithm, applied to f, is carried out via the matrix $\mathbf{S}(n)$, where n is given by (3.1.8). We have $f \cdot \mathbf{S}(n) = f_1 = [a_1, b_1, c_1]$, where f_1 is a reduced form as well by Theorem 3.1.4. We now wish to compute the matrix transformation that takes us from f_1 to f^*. First note that $f_1 \cdot \mathbf{S}(n)^{-1} = f$, and it is

easy to check that

$$\mathbf{S}(n)^{-1} = \begin{pmatrix} 0 & -1 \\ 1 & n \end{pmatrix},$$

which we earlier denoted by $\mathbf{T}(n)$ in (3.1.65). Combining $f_1 \cdot \mathbf{S}(n)^{-1} = f$ with $f \cdot \mathbf{A} = f^*$ gives

$$f_1 \cdot \mathbf{S}(n)^{-1} \cdot \mathbf{A} = f^*.$$

We have

$$\mathbf{S}(n)^{-1} \cdot \mathbf{A} = \begin{pmatrix} -t & -u \\ r + nt & s + nu \end{pmatrix}, \tag{3.1.109}$$

and if this last matrix is set equal to

$$\mathbf{A}_1 = \begin{pmatrix} r_1 & s_1 \\ t_1 & u_1 \end{pmatrix},$$

we have

$$f_1 \cdot \mathbf{A}_1 = f^*, \tag{3.1.110}$$

which gives us the transformation matrix $\mathbf{A}_1$ taking f_1 to f^*. Note that

$$r_1 - s_1 = -t + u > 0, \tag{3.1.111}$$

with this last inequality holding by (3.1.104). Since $\mathbf{A}_1 \in \mathrm{SL}_2(\mathbb{Z})$ is such that $f_1 \cdot \mathbf{A}_1 = f^*$, we see by (3.1.101) that

$$a^* + c^* - b^* = f_1 (r_1 - s_1, t_1 - u_1). \tag{3.1.112}$$

By (3.1.102), we have

$$f_1 (r_1 - s_1, t_1 - u_1) < 0. \tag{3.1.113}$$

The left side of (3.1.113) is equal to

$$a_1 (r_1 - s_1)^2 + b_1 (r_1 - s_1)(t_1 - u_1) + c_1 (t_1 - u_1)^2 . \tag{3.1.114}$$

Since f_1 is a reduced form, we have $a_1 > 0$ and $c_1 > 0$, and thus the only way that the expression in (3.1.114) can be negative is if $t_1 - u_1 < 0$ (recall from (3.1.111) that $r_1 - s_1 > 0$, and we also have $b_1 > 0$). Thus, we have

$$t_1 < u_1, \tag{3.1.115}$$

which shows that (3.1.104) holds with respect to the matrix $\mathbf{A}_1$ that is employed in (3.1.97) taking us from f_1 to f^*. This is important because if it so happened that $t_1 = 0$, we could conclude by Case I above that

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and so $f_1 = f^*$ (recall that (3.1.104) was a crucial ingredient used in the argument to verify Case I). From (3.1.109), we have $t_1 = r + nt$, and it is critical at this stage in the proof to establish that

$$t < t_1 \leq 0,$$

or equivalently

$$t < r + nt \leq 0. \tag{3.1.116}$$

The two inequalities in (3.1.116) may be replaced by two other inequalities that are easier to establish. We claim that the inequality

$$n - 1 < \frac{r}{(-t)} \tag{3.1.117}$$

implies the left inequality in (3.1.116). Assuming that (3.1.117) holds, and recalling that in the present Case II that $t < 0$, we have $0 < -t$, and so $(-t)(n-1) < r$, or $t - nt < r$, or finally $t < r + nt$, which proves our claim. We next claim that the inequality

$$\frac{r}{(-t)} \leq n \tag{3.1.118}$$

implies the right inequality in (3.1.116). Assuming that (3.1.118) holds, and with $0 < -t$, we have $r \leq (-t)n$, or $r + nt \leq 0$, which proves the claim. Therefore, if we can prove the inequalities in (3.1.117) and (3.1.118), then we will have established both inequalities in (3.1.116).

We first prove (3.1.117). This involves the clever use of the quadratic polynomial

$$\psi(x) = f(x, -1) = ax^2 - bx + c, \tag{3.1.119}$$

which was first encountered in Lemma 3.1.5. Since $a > 0$ and $(-b)^2 - 4ac > 0$, the graph of this polynomial is a parabola that opens upward and crosses the x-axis at two distinct points, namely $(b \pm \sqrt{D})/2a$, with $(b + \sqrt{D})/2a$ being the rightmost root by the forward direction of Lemma 3.1.5, in particular (3.1.46). We now claim that if we evaluate $\psi(x)$ at the rational number $r/(-t)$, we obtain a positive answer. To see this, recall from (3.1.98) that $f(r, t) > 0$ since f^* is a reduced form. Thus, we have $0 < ar^2 + brt + ct^2$. If we divide both sides of this last inequality by the positive integer t^2, we obtain

$$0 < a\left(\frac{r}{t}\right)^2 + b\left(\frac{r}{t}\right) + c = a\left[\frac{r}{(-t)}\right]^2 - b\left[\frac{r}{(-t)}\right] + c = \psi\left(\frac{r}{(-t)}\right), \tag{3.1.120}$$

which proves our claim. We also claim that if we evaluate $\psi(x)$ at the rational number $(r - s)/(-t + u)$, noting that $-t + u \neq 0$ by (3.1.104), we obtain a negative answer.

To see this, recall from (3.1.101) and (3.1.102) that $f(r-s, t-u) < 0$ since f^* is a reduced form. Thus, we have $0 > a(r-s)^2 + b(r-s)(t-u) + c(t-u)^2$. If we divide both sides of this last inequality by the positive integer $(t-u)^2$, we obtain

$$0 > a\left[\frac{(r-s)}{(t-u)}\right]^2 + b\left[\frac{(r-s)}{(t-u)}\right] + c = a\left[\frac{(r-s)}{(-t+u)}\right]^2 - b\left[\frac{(r-s)}{(-t+u)}\right] + c, \quad (3.1.121)$$

and thus we have

$$\psi\left(\frac{(r-s)}{(-t+u)}\right) < 0,$$

which verifies our claim. Our next claim is that

$$\frac{(r-s)}{(-t+u)} < \frac{r}{(-t)}. \quad (3.1.122)$$

To see this, note that $0 < ru - st = 1$, which implies $-rt + st < -rt + ru$, or $(-t)(r-s) < r(-t+u)$. Recalling that $0 < -t$ in the present Case II, and that $0 < -t + u$ by (3.1.104), we obtain a verification of (3.1.122) by cross division. In terms of the parabolic graph of $\psi(x)$, since $\psi(x)$ takes on a negative value when $x = (r-s)/(-t+u)$ and $(b+\sqrt{D})/2a$ is the rightmost root of $\psi(x)$, we must have $(r-s)/(-t+u) < (b+\sqrt{D})/2a$. Combining (3.1.120) and (3.1.122) shows that $r/(-t)$ must lie to the right of this rightmost root of $\psi(x)$. We conclude that

$$\frac{(r-s)}{(-t+u)} < \frac{(b+\sqrt{D})}{2a} < \frac{r}{(-t)}. \quad (3.1.123)$$

Combining (3.1.123) with (3.1.7) confirms that $n - 1 < r/(-t)$, which proves the inequality in (3.1.117).

To prove the inequality in (3.1.118), assume on the contrary that we have $n < r/(-t)$. By (3.1.7), we have $(b+\sqrt{D})/2a < n$, and putting these last two

inequalities together with (3.1.123) gives

$$\frac{(r-s)}{(-t+u)} < n < \frac{r}{(-t)}. \tag{3.1.124}$$

Recalling that $0 < -t$ in the present Case II, and that $0 < -t+u$ by (3.1.104), if we multiply the inequalities in (3.1.124) through by the product of positive integers $(-t)\cdot(-t+u)$, we obtain

$$(-t)(r-s) < n(-t)(-t+u) < r(-t+u),$$

or

$$-rt+st < nt(t-u) < -rt+ru. \tag{3.1.125}$$

Each of the three expressions appearing in this string of inequalities is an *integer*, and since each of the inequalities in (3.1.125) is *strict*, we may conclude that $-rt+st+2 \leq -rt+ru$, which in turn implies that $2 \leq ru-st$, which contradicts the known value $ru-st=1$. Since we arrived at this contradiction under the assumption that $n < r/(-t)$, the inequality in (3.1.118) must hold. Thus, we have established both inequalities in (3.1.116).

With the inequalities in (3.1.116) now established, we are ready to conclude our analysis of Case II, with the help of Case I. We carry out this final part of the argument by a process of induction. Before we begin, let us review our progress to this point. We started with a reduced form $f=[a,b,c]$ and a transformation matrix

$$\mathbf{A} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

with $t<u$ taking f to the reduced form $f^*=[a^*,b^*,c^*]$. Case I shows that under

these circumstances, if $t = 0$, then

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$f = f^*$, and f^* appears in the sequence (3.1.96). If $t < 0$, in Case II, we consider the right neighbor f_1 of f, which is also reduced. We then have a well-defined transformation matrix

$$\mathbf{A}_1 = \begin{pmatrix} r_1 & s_1 \\ t_1 & u_1 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

with $t_1 < u_1$ taking f_1 to f^* such that, by (3.1.116), we have $t < t_1 \leq 0$. If $t_1 = 0$, then by Case I we have

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$f_1 = f^*$, and again f^* appears in the sequence (3.1.96). If $t_1 < 0$, we arrive back in Case II, and we then consider the right neighbor f_2 of f_1, which itself is a reduced form. Again, we have a well-defined transformation matrix

$$\mathbf{A}_2 = \begin{pmatrix} r_2 & s_2 \\ t_2 & u_2 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

with $t_2 < u_2$ taking f_2 to f^* such that $t_1 < t_2 \leq 0$. If $t_2 = 0$, then we have $f_2 = f^*$, and so f^* again appears in the sequence (3.1.96). If $t_2 < 0$, we consider the right neighbor f_3 of f_2, and proceed in exactly the same manner. Since $t, t_1, t_2, \ldots$ are all negative integers, and the inequalities $t < t_1 < t_2 < \cdots$ are all strict, this inductive procedure *must* terminate after a *finite* number of steps with $t_j = 0$ and $f_j = f^*$ for some form f_j in the sequence (3.1.96). In this way, the proof of Theorem 3.1.11 is

complete if we land in either Case I or Case II.

If we do not land in either Case I or Case II, we finally consider the last possible case.

Case III: $t > 0$. If $t > 0$, note that we still have (see (3.1.120))

$$0 < \psi\left(\frac{r}{(-t)}\right), \tag{3.1.126}$$

since this inequality holds regardless of the sign of t, provided of course that $t \neq 0$. We also have

$$\psi\left(\frac{(r-s)}{(-t+u)}\right) < 0, \tag{3.1.127}$$

by use of the inequality in (3.1.121). On the other hand, the inequality in (3.1.122) no longer holds, so we must start from scratch in this case. We have $0 < ru - st = 1$, which leads to $-rt + st < -rt + ru$, or $(-t)(r-s) < r(-t+u)$. Since $0 < -t + u$ by (3.1.104), we have $[(-t)(r-s)]/(-t+u) < r$, and since $-t < 0$ in Case III, we have

$$\frac{r}{(-t)} < \frac{(r-s)}{(-t+u)}. \tag{3.1.128}$$

In terms of the parabolic graph of $\psi(x)$, we see from (3.1.127) that the rational number $(r-s)/(-t+u)$ must fall between the two distinct roots of $\psi(x)$, and so

$$\frac{b-\sqrt{D}}{2a} < \frac{(r-s)}{(-t+u)}.$$

Combining (3.1.126) and (3.1.128) shows that $r/(-t)$ must lie to the left of the leftmost root of $\psi(x)$. We conclude that

$$\frac{r}{(-t)} < \frac{b-\sqrt{D}}{2a} < \frac{(r-s)}{(-t+u)}. \tag{3.1.129}$$

From (3.1.46), we have $(b-\sqrt{D})/2a < 1$, which depends directly upon $f = [a, b, c]$

being a reduced form. Combining this last inequality with (3.1.129) gives $r/(-t) < 1$, and since $-t < 0$, we conclude that

$$-t < r, \tag{3.1.130}$$

which is the crucial inequality that we need in Case III. From the beginning, we started with a transformation matrix

$$\mathbf{A} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

with $t < u$ which takes the reduced form $f = [a, b, c]$ to the reduced form $f^* = [a^*, b^*, c^*]$, namely

$$f \cdot \mathbf{A} = f^*. \tag{3.1.131}$$

From (3.1.131), we see that

$$f^* \cdot \mathbf{A}^{-1} = f, \tag{3.1.132}$$

and an easy calculation shows that

$$\mathbf{A}^{-1} = \begin{pmatrix} u & -s \\ -t & r \end{pmatrix}. \tag{3.1.133}$$

From this new angle, we imagine starting with the reduced form f^*, and using the transformation matrix $\mathbf{A}^{-1}$ to take us to the reduced form $f = [a, b, c]$. The inequality in (3.1.130) is the exact analogue to (3.1.104) when we earlier went from f to f^*. Since the lower left entry in $\mathbf{A}^{-1}$ is negative, we are in the exact setting we were in earlier going from f to f^* in Case II. By our earlier work in Cases I and II, we may conclude that f lies in the cycle generated from the starting form f^*. Appealing to Definition 3.1.10, in particular to the symmetric property of that definition, we

conclude that f^* appears in the sequence (3.1.96) in Case III as well, which completes the proof of Theorem 3.1.11. □

The theory presented above allows us to answer a question posed at the beginning of this section: "How does one determine when two *arbitrary* forms of discriminant D lie in the same equivalence class?" Theorem 3.1.11, in particular, tells us that two arbitrarily given forms of discriminant D lie in the same equivalence class if and only if the reduction algorithm takes them to reduced forms which are members of the same cycle.

As an illustration of the theory developed in this section, we present a list of all cycles of reduced forms in Table 3.1.3 for those discriminants explicitly described at the beginning of Section 2.2, beginning with $D = 5$ and ending with $D = 65$. The "principal cycle of discriminant D", by definition, is the cycle of reduced forms obtained when the reduction algorithm is applied to the principal form of discriminant D (recall that the principal form itself is never a reduced form). In Table 3.1.3, we number each cycle separately, with the first cycle displayed for a given discriminant being always the principal cycle of that discriminant. Imprimitive cycles are represented with an asterisk $*$ before their associated number, while the primitive cycles are always listed first and are unmarked. Being aware of these conventions allows one to quickly scan through Table 3.1.3 and work out the value of $h(D)$ and $t(D)$ for each value of D included. A leisurely comparison between Table 3.1.3 and Table 2.5.1 shows perfect agreement in the value of $h(D)$ for each D listed. The numbers above the arrows in Table 3.1.3 were already defined earlier in this section. An arrow going straight out at the end of a line with a number above it takes you to the first form on the next line. Using the convention established earlier in this section, the curved arrows take you all the way back to the beginning of the cycle.

Table 3.1.3. All cycles of reduced forms of discriminant D with $5 \leq D \leq 65$.

D	Cycles of All Reduced Forms
5	1) $[1,3,\underset{\underset{3}{\longleftarrow}}{1}]$
8	1) $[1,4,2] \overset{4}{\to} [2,4,\underset{\underset{2}{\longleftarrow}}{1}]$
12	1) $[1,4,\underset{\underset{4}{\longleftarrow}}{1}]$
	2) $[3,6,2] \overset{2}{\to} [2,6,\underset{\underset{3}{\longleftarrow}}{3}]$
13	1) $[1,5,3] \overset{5}{\to} [3,5,1] \overset{2}{\to} [3,7,\underset{\underset{2}{\longleftarrow}}{3}]$
17	1) $[1,5,2] \overset{5}{\to} [2,5,1] \overset{3}{\to} [4,7,2] \overset{2}{\to} [4,9,4] \overset{2}{\to} [2,7,\underset{\underset{3}{\longleftarrow}}{4}]$
20	1) $[1,6,4] \overset{6}{\to} [4,6,1] \overset{2}{\to} [5,10,4] \overset{2}{\to} [4,10,\underset{\underset{2}{\longleftarrow}}{5}]$
	$*$2) $[2,6,\underset{\underset{3}{\longleftarrow}}{2}]$
21	1) $[1,5,\underset{\underset{5}{\longleftarrow}}{1}]$
	2) $[5,9,3] \overset{2}{\to} [5,11,5] \overset{2}{\to} [3,9,\underset{\underset{3}{\longleftarrow}}{5}]$
24	1) $[1,6,3] \overset{6}{\to} [3,6,\underset{\underset{2}{\longleftarrow}}{1}]$
	2) $[5,8,2] \overset{2}{\to} [6,12,5] \overset{2}{\to} [5,12,6] \overset{2}{\to} [2,8,\underset{\underset{4}{\longleftarrow}}{5}]$
28	1) $[1,6,2] \overset{6}{\to} [2,6,\underset{\underset{3}{\longleftarrow}}{1}]$
	2) $[6,10,3] \overset{2}{\to} [7,14,6] \overset{2}{\to} [6,14,7] \overset{2}{\to} [3,10,6] \overset{3}{\to} [3,8,\underset{\underset{3}{\longleftarrow}}{3}]$
29	1) $[1,7,5] \overset{7}{\to} [5,7,1] \overset{2}{\to} [7,13,5] \overset{2}{\to} [7,15,7] \overset{2}{\to} [5,13,\underset{\underset{2}{\longleftarrow}}{7}]$

D	Cycles of All Reduced Forms
32	1) $[1,6,\underset{\overset{\longleftarrow}{6}}{1}]$
	2) $[7,12,4] \xrightarrow{2} [8,16,7] \xrightarrow{2} [7,16,8] \xrightarrow{2} [4,12,\underset{\overset{\longleftarrow}{3}}{7}]$
	$*3)$ $[4,8,2] \xrightarrow{2} [2,8,\underset{\overset{\longleftarrow}{4}}{4}]$
33	1) $[1,7,4] \xrightarrow{7} [4,7,1] \xrightarrow{2} [3,9,4] \xrightarrow{3} [4,9,\underset{\overset{\longleftarrow}{2}}{3}]$
	2) $[6,9,2] \xrightarrow{2} [8,15,6] \xrightarrow{2} [8,17,8] \xrightarrow{2} [6,15,8] \xrightarrow{2} [2,9,6] \xrightarrow{4} [2,7,\underset{\overset{\longleftarrow}{4}}{2}]$
37	1) $[1,7,3] \xrightarrow{7} [3,7,1] \xrightarrow{3} [7,11,3] \xrightarrow{2} [9,17,7] \xrightarrow{2}$
	$[9,19,9] \xrightarrow{2} [7,17,9] \xrightarrow{2} [3,11,\underset{\overset{\longleftarrow}{3}}{7}]$
40	1) $[1,8,6] \xrightarrow{8} [6,8,1] \xrightarrow{2} [9,16,6] \xrightarrow{2} [10,20,9] \xrightarrow{2} [9,20,10] \xrightarrow{2} [6,16,\underset{\overset{\longleftarrow}{2}}{9}]$
	2) $[5,10,3] \xrightarrow{2} [3,10,5] \xrightarrow{3} [2,8,3] \xrightarrow{4} [3,8,\underset{\overset{\longleftarrow}{3}}{2}]$
41	1) $[1,7,2] \xrightarrow{7} [2,7,1] \xrightarrow{4} [5,9,2] \xrightarrow{2} [4,11,5] \xrightarrow{3} [8,13,4] \xrightarrow{2} [10,19,8] \xrightarrow{2}$
	$[10,21,10] \xrightarrow{2} [8,19,10] \xrightarrow{2} [4,13,8] \xrightarrow{3} [5,11,4] \xrightarrow{2} [2,9,\underset{\overset{\longleftarrow}{4}}{5}]$
44	1) $[1,8,5] \xrightarrow{8} [5,8,1] \xrightarrow{2} [5,12,\underset{\overset{\longleftarrow}{2}}{5}]$
	2) $[7,10,2] \xrightarrow{2} [10,18,7] \xrightarrow{2} [11,22,10] \xrightarrow{2} [10,22,11] \xrightarrow{2} [7,18,10] \xrightarrow{2} [2,10,\underset{\overset{\longleftarrow}{5}}{7}]$
45	1) $[1,7,\underset{\overset{\longleftarrow}{7}}{1}]$
	2) $[9,15,5] \xrightarrow{2} [11,21,9] \xrightarrow{2} [11,23,11] \xrightarrow{2} [9,21,11] \xrightarrow{2} [5,15,\underset{\overset{\longleftarrow}{3}}{9}]$
	$*3)$ $[3,9,\underset{\overset{\longleftarrow}{3}}{3}]$

D	Cycles of All Reduced Forms
48	1) $[1,8,4] \overset{8}{\to} [4,8,\underset{\underset{2}{\hookleftarrow}}{1}]$
	2) $[8,12,3] \overset{2}{\to} [11,20,8] \overset{2}{\to} [12,24,11] \overset{2}{\to} [11,24,12] \overset{2}{\to} [8,20,11] \overset{2}{\to} [3,12,\underset{\underset{4}{\hookleftarrow}}{8}]$
	$*3)$ $[2,8,\underset{\underset{4}{\hookleftarrow}}{2}]$
	$*4)$ $[6,12,4] \overset{2}{\to} [4,12,\underset{\underset{3}{\hookleftarrow}}{6}]$
52	1) $[1,8,3] \overset{8}{\to} [3,8,1] \overset{3}{\to} [4,10,3] \overset{3}{\to} [9,14,4] \overset{2}{\to} [12,22,9] \overset{2}{\to}$
	$[13,26,12] \overset{2}{\to} [12,26,13] \overset{2}{\to} [9,22,12] \overset{2}{\to} [4,14,9] \overset{3}{\to} [3,10,\underset{\underset{3}{\hookleftarrow}}{4}]$
	$*2)$ $[6,10,2] \overset{2}{\to} [6,14,6] \overset{2}{\to} [2,10,\underset{\underset{5}{\hookleftarrow}}{6}]$
53	1) $[1,9,7] \overset{9}{\to} [7,9,1] \overset{2}{\to} [11,19,7] \overset{2}{\to} [13,25,11] \overset{2}{\to}$
	$[13,27,13] \overset{2}{\to} [11,25,13] \overset{2}{\to} [7,19,\underset{\underset{2}{\hookleftarrow}}{11}]$
56	1) $[1,8,2] \overset{8}{\to} [2,8,\underset{\underset{4}{\hookleftarrow}}{1}]$
	2) $[10,16,5] \overset{2}{\to} [13,24,10] \overset{2}{\to} [14,28,13] \overset{2}{\to} [13,28,14] \overset{2}{\to}$
	$[10,24,13] \overset{2}{\to} [5,16,10] \overset{3}{\to} [7,14,5] \overset{2}{\to} [5,14,\underset{\underset{3}{\hookleftarrow}}{7}]$
57	1) $[1,9,6] \overset{9}{\to} [6,9,1] \overset{2}{\to} [7,15,6] \overset{2}{\to} [4,13,7] \overset{3}{\to}$
	$[4,11,4] \overset{3}{\to} [7,13,4] \overset{2}{\to} [6,15,\underset{\underset{2}{\hookleftarrow}}{7}]$
	2) $[8,11,2] \overset{2}{\to} [12,21,8] \overset{2}{\to} [14,27,12] \overset{2}{\to} [14,29,14] \overset{2}{\to}$
	$[12,27,14] \overset{2}{\to} [8,21,12] \overset{2}{\to} [2,11,8] \overset{5}{\to} [3,9,2] \overset{3}{\to} [2,9,\underset{\underset{5}{\hookleftarrow}}{3}]$

D	Cycles of All Reduced Forms
60	1) $[1, 8, \underset{\hookleftarrow 8}{1}]$
	2) $[11, 18, 6] \xrightarrow{2} [14, 26, 11] \xrightarrow{2} [15, 30, 14] \xrightarrow{2} [14, 30, 15] \xrightarrow{2} [11, 26, 14] \xrightarrow{2} [6, 18, \underset{\hookleftarrow 3}{11}]$
	3) $[7, 12, 3] \xrightarrow{2} [7, 16, 7] \xrightarrow{2} [3, 12, \underset{\hookleftarrow 4}{7}]$
	4) $[5, 10, 2] \xrightarrow{2} [2, 10, \underset{\hookleftarrow 5}{5}]$
61	1) $[1, 9, 5] \xrightarrow{9} [5, 9, 1] \xrightarrow{2} [3, 11, 5] \xrightarrow{4} [9, 13, 3] \xrightarrow{2} [13, 23, 9] \xrightarrow{2} [15, 29, 13] \xrightarrow{2}$ $[15, 31, 15] \xrightarrow{2} [13, 29, 15] \xrightarrow{2} [9, 23, 13] \xrightarrow{2} [3, 13, 9] \xrightarrow{4} [5, 11, \underset{\hookleftarrow 2}{3}]$
65	1) $[1, 9, 4] \xrightarrow{9} [4, 9, 1] \xrightarrow{3} [10, 15, 4] \xrightarrow{2} [14, 25, 10] \xrightarrow{2} [16, 31, 14] \xrightarrow{2}$ $[16, 33, 16] \xrightarrow{2} [14, 31, 16] \xrightarrow{2} [10, 25, 14] \xrightarrow{2} [4, 15, \underset{\hookleftarrow 3}{10}]$
	2) $[7, 11, 2] \xrightarrow{2} [8, 17, 7] \xrightarrow{2} [5, 15, 8] \xrightarrow{3} [8, 15, 5] \xrightarrow{2}$ $[7, 17, 8] \xrightarrow{2} [2, 11, 7] \xrightarrow{5} [2, 9, \underset{\hookleftarrow 5}{2}]$

3.2. Solving the Fermat-Pell 4-Equation using Reduced Forms

In this section, we use the theory presented in Section 3.1, and specifically Theorem 3.1.11, to find all solutions to the Fermat-Pell 4-Equation (2.3.1). Theorem 2.3.3 plays a crucial role in this discussion since it allows us to obtain all solutions to (2.3.1) from a knowledge of all automorphs of a given primitive form f of discriminant $D \in \mathbb{Z}^+$. Theorem 3.1.11 is particularly helpful in terms of describing $\mathrm{Aut}\,(f)$ when the form f is reduced. With these comments in mind, we often assume throughout this section that $f \in Q(D)$ is a form that is both primitive and reduced. We have already verified that such a form always exists for any given discriminant $D \in \mathbb{Z}^+$.

The basic idea is fairly simple. Given a reduced form $f \in Q(D)$, we move through the cycle of reduced forms to which f belongs, multiplying the $\mathbf{S}(n)$ matrices together as we go, and every time we return to f in the cycle we obtain an automorph of f. It is easy to see that infinitely many distinct automorphs of f are obtained as we run through the cycle over and over again.

We first require a lemma, and for this initial result we can momentarily relax all restrictions on the form f.

Lemma 3.2.1. *Let $f = f_0$ be an arbitrary form in $Q(D)$, and assume that when the Zagier reduction algorithm is applied to f_0 that we obtain the following uniquely defined infinite sequence of forms:*

$$f_0 \xrightarrow{n_0} f_1 \xrightarrow{n_1} f_2 \xrightarrow{n_2} \cdots . \tag{3.2.1}$$

By the right group action of $\mathrm{SL}_2(\mathbb{Z})$ on $Q(D)$, we know for each $\ell \in \mathbb{Z}^{\geq 0}$ that $f_0 \cdot \mathbf{U}_\ell = f_{\ell+1}$, where

$$\mathbf{U}_\ell = \prod_{j=0}^{\ell} \mathbf{S}(n_j). \tag{3.2.2}$$

If we set $p_{-2} = 0,\ p_{-1} = 1,\ q_{-2} = -1,\ q_{-1} = 0$, and define recursively

$$p_j = n_j p_{j-1} - p_{j-2} \quad \text{for } j = 0, 1, 2, \ldots \tag{3.2.3}$$

$$q_j = n_j q_{j-1} - q_{j-2} \quad \text{for } j = 0, 1, 2, \ldots, \tag{3.2.4}$$

then for each $\ell \in \mathbb{Z}^{\geq 0}$ we have

$$\mathbf{U}_\ell = \begin{pmatrix} p_\ell & p_{\ell-1} \\ -q_\ell & -q_{\ell-1} \end{pmatrix}. \tag{3.2.5}$$

Remark. The values $p_{-2} = 0,\ p_{-1} = 1,\ q_{-2} = -1,\ q_{-1} = 0$, as well as the recursive

formulas (3.2.3) and (3.2.4), play a crucial role in the theory of minus continued fractions as presented in Section 4.2. Indeed, Lemma 3.2.1 is our first link tying together Zagier's reduction theory for indefinite binary quadratic forms and the theory of minus continued fractions. For future reference, we observe that

$$p_0 = n_0 p_{-1} - p_{-2} = n_0 \tag{3.2.6}$$

and

$$q_0 = n_0 q_{-1} - q_{-2} = 1. \tag{3.2.7}$$

We also note that if we look at the matrix on the right hand side of (3.2.5) when $\ell = -1$ that we obtain the identity matrix in $\mathrm{SL}_2(\mathbb{Z})$:

$$\begin{pmatrix} p_{-1} & p_{-2} \\ -q_{-1} & -q_{-2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I}.$$

Proof of Lemma 3.2.1. We proceed by induction. For the base case, when $\ell = 0$, we may use (3.2.6) and (3.2.7) to obtain

$$\begin{pmatrix} p_0 & p_{-1} \\ -q_0 & -q_{-1} \end{pmatrix} = \begin{pmatrix} n_0 & 1 \\ -1 & 0 \end{pmatrix} = \mathbf{S}(n_0),$$

which equals $\mathbf{U}_0$ by (3.2.2). Now assume that (3.2.5) holds for some integer $\ell \in \mathbb{Z}^{\geq 0}$. Making use of the induction hypothesis, we obtain

$$\mathbf{U}_\ell \cdot \mathbf{S}(n_{\ell+1}) = \begin{pmatrix} p_\ell & p_{\ell-1} \\ -q_\ell & -q_{\ell-1} \end{pmatrix} \cdot \begin{pmatrix} n_{\ell+1} & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} n_{\ell+1} p_\ell - p_{\ell-1} & p_\ell \\ -(n_{\ell+1} q_\ell - q_{\ell-1}) & -q_\ell \end{pmatrix},$$

and we note that this last matrix is equal to

$$\begin{pmatrix} p_{\ell+1} & p_\ell \\ -q_{\ell+1} & -q_\ell \end{pmatrix}$$

by (3.2.3) and (3.2.4). Since the first matrix product in the string of equalities above is equal to $\mathbf{U}_{\ell+1}$ by (3.2.2), the induction step is complete. □

By use of this lemma, the following easy corollary may be derived.

Corollary 3.2.2. *If $f = f_0 \in Q(D)$ is a reduced form, then in terms of the notation instituted in Lemma 3.2.1, we have*

$$1 = q_0 < q_1 < q_2 < \cdots . \tag{3.2.8}$$

Proof. Since f_0 is reduced by assumption, every form in the diagram (3.2.1) is reduced as well by Theorem 3.1.4. By the forward direction of Lemma 3.1.5, we see that $n_j \in \mathbb{Z}^{\geq 2}$ for $j = 0, 1, 2, \ldots$. Looking ahead to Section 4.2, we prove (3.2.8) by induction under the weaker assumption that $n_j \in \mathbb{Z}^{\geq 2}$ for $j = 1, 2, 3, \ldots$ [see (4.2.3)]. The referred to induction proof is laid out just under (4.2.7a). □

We now have all of the results needed to give a constructive proof of Theorem 2.3.5, which states that the Fermat-Pell 4-Equation $t^2 - Du^2 = 4$ has an integer pair solution $(t_0, u_0) \in \mathbb{Z}^2$ with $u_0 \in \mathbb{Z}^+$.

Proof of Theorem 2.3.5. Let $f = [a, b, c] = f_0 \in Q(D)$ be a form that is both primitive and reduced. Assume that there are exactly $m \in \mathbb{Z}^+$ forms lying in the cycle of reduced forms to which f belongs, and therefore that we have the following diagram:

$$f_0 \xrightarrow{n_0} f_1 \xrightarrow{n_1} \cdots \to f_{m-1} \underset{n_{m-1}}{\longrightarrow} . \tag{3.2.9}$$

From this diagram, we see that $\mathbf{U}_{m-1} = \prod_{j=0}^{m-1} \mathbf{S}(n_j)$ is an automorph of the primitive form f. Combining (3.2.5) and Theorem 2.3.3, we see that

$$au = -q_{m-1}, \tag{3.2.10}$$

where $a \in \mathbb{Z}^+$ and u is an integer that along with another integer t_0 satisfies the Diophantine equation $t_0^2 - Du^2 = 4$. By Corollary 3.2.2, we have $0 < q_{m-1}$. By (3.2.10), we see that $u = -q_{m-1}/a$ is a negative integer. If we set $u_0 = -u \in \mathbb{Z}^+$, then $t_0^2 - Du_0^2 = 4$, completing the proof. This proof is constructive since the theory presented in Section 3.1 allows us to assemble the diagram in (3.2.9), which in turn leads us directly to being able to compute the integers u_0 and t_0 as shown above. □

Proceeding along the same lines laid out in the proof of Theorem 2.3.5, we may prove the following stronger result.

Corollary 3.2.3. *Given a fixed discriminant $D \in \mathbb{Z}^+$, the Fermat-Pell 4-Equation* (2.3.1) *possesses an infinite number of distinct integer pair solutions. Also, if $g \in Q(D)$ is a primitive form, then* $\mathrm{Aut}\,(g)$ *is a set with infinitely many distinct elements.*

Proof. Just as in the proof of Theorem 2.3.5, let $f = [a, b, c] = f_0 \in Q(D)$ be a form that is both primitive and reduced, and assume again that we have diagram (3.2.9) in effect. Not only is $\mathbf{U}_{m-1}$ an automorph of the primitive form f, but we also see from the diagram that $\mathbf{U}_{2m-1}, \mathbf{U}_{3m-1}, \ldots$ are all in $\mathrm{Aut}\,(f)$ as well. By (3.2.8), we see that $q_{m-1}, q_{2m-1}, q_{3m-1}, \ldots$ constitutes an infinite sequence of *mutually distinct* positive integers. Combining (3.2.5) and Theorem 2.3.3, we see that each positive integer

$$\frac{q_{km-1}}{a} \quad \text{for } k = 1, 2, 3, \ldots, \tag{3.2.11}$$

is the u–value in an integer pair solution (t, u) of (2.3.1). Since all of the integers

in (3.2.11) are mutually distinct, we see that (2.3.1) possesses an infinite number of distinct integer pair solutions.

Now assume that $g = [a^*, b^*, c^*] \in Q(D)$ is a primitive form, but not necessarily reduced. We saw above that there exists an infinite sequence of integer pair solutions to (2.3.1) which may be put in to the form $(t_1, u_1), (t_2, u_2), (t_3, u_3), \ldots$, where we have $0 < u_1 < u_2 < u_3 < \cdots$. Looking at the 2×2 matrices in (2.3.3) that make up the set $\mathrm{Aut}\,(g)$, we have an infinite number of distinct integers $a^*u_1,\ a^*u_2,\ a^*u_3, \ldots$ appearing in the lower left entry and so $\mathrm{Aut}\,(g)$ is a set with infinitely many distinct elements. □

Looking back at Example 2.3.4, the automorph given for the form $f = [2, 6, 2]$ of discriminant $D = 20$ is simply $\mathbf{S}(3)$, which is equal to $\mathbf{U}_0$ (in Table 3.1.3, we see that $[2, 6, 2]$ is the only reduced form in its cycle). The automorph given for the form $f = [4, 12, 6]$ of discriminant $D = 48$ is equal to $\mathbf{S}(3) \cdot \mathbf{S}(2)$, which is $\mathbf{U}_1$ in this case (in Table 3.1.3, we see that $[4, 12, 6]$ is one of two reduced forms in a cycle). Finally, the automorph given for the form $f = [1, 6, 4]$ of discriminant $D = 20$ is equal to $\mathbf{S}(6) \cdot \mathbf{S}(2) \cdot \mathbf{S}(2) \cdot \mathbf{S}(2)$, which is $\mathbf{U}_3$ in this case (in Table 3.1.3, we see that $[1, 6, 4]$ is one of four reduced forms in a cycle).

For the remainder of this section, our goal is to start with a reduced (not necessarily primitive) form $f = f_0$ of discriminant $D \in \mathbb{Z}^+$ and to describe the countably large infinite set of all matrices in $\mathrm{Aut}\,(f)$. The basic structure of $\mathrm{Aut}\,(f)$ is easy to describe, as seen in the following theorem. In the statement below, to obtain a matrix in the form $-\mathbf{U}^{-3}$, for example, we raise the multiplicative inverse $\mathbf{U}^{-1}$ to the 3rd power and then multiply every entry of the resulting matrix by -1.

Theorem 3.2.4. *If f is a reduced form of discriminant $D \in \mathbb{Z}^+$, then there exists a*

matrix $\mathbf{U} \in \mathrm{SL}_2(\mathbb{Z})$ *such that every element in* $\mathrm{Aut}\,(f)$ *can be written uniquely in the form* $\pm\mathbf{U}^k$*, as* $k \in \mathbb{Z}$ *ranges over all the integers.*

Proof. Assume that there are exactly $m \in \mathbb{Z}^+$ forms lying in the cycle of reduced forms to which $f = f_0$ belongs, and therefore that we have the following diagram:

$$f_0 \xrightarrow{n_0} f_1 \xrightarrow{n_1} \cdots \to f_{m-1} \underset{n_{m-1}}{\longrightarrow}, \tag{3.2.12}$$

recalling that $f_0, f_1, \ldots, f_{m-1}$ are all distinct from each other if $m \geq 2$. From this diagram, we see that $\mathbf{U}_{m-1} = \prod_{j=0}^{m-1} \mathbf{S}(n_j)$ is an automorph of the reduced form f. Set $\mathbf{U} = \mathbf{U}_{m-1}$. Recalling from Theorem 2.3.2 that $\mathrm{Aut}\,(f)$ is a subgroup of $\mathrm{SL}_2(\mathbb{Z})$, we note that every power $\mathbf{U}^k$, for all $k \in \mathbb{Z}$, lies in $\mathrm{Aut}\,(f)$. Also from Section 2.3, we recall that

$$-\mathbf{I} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \in \mathrm{Aut}\,(f).$$

This means that every 2×2 matrix of the form $(\pm\mathbf{I}) \cdot \mathbf{U}^k$, for all $k \in \mathbb{Z}$, lies in $\mathrm{Aut}\,(f)$. In terms of the notation introduced above, we have $(-\mathbf{I}) \cdot \mathbf{A} = -\mathbf{A}$ for every $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$. We also note that $-\mathbf{I}$ commutes with every matrix $\mathbf{A} \in \mathrm{SL}_2(\mathbb{Z})$, namely,

$$(-\mathbf{I}) \cdot \mathbf{A} = \mathbf{A} \cdot (-\mathbf{I}). \tag{3.2.13}$$

Now that we know that $\pm\mathbf{U}^k \in \mathrm{Aut}\,(f)$ for every $k \in \mathbb{Z}$, we wish to prove the opposite inclusion, namely, that every $\mathbf{A} \in \mathrm{Aut}\,(f)$ is equal to either $\mathbf{U}^k$ or $-\mathbf{U}^k$ for some $k \in \mathbb{Z}$. This requires that we take a careful look back at the proof of Theorem 3.1.11. At the beginning of this proof, we assume that $f^* = [a^*, b^*, c^*]$ is a reduced form of discriminant D lying in the *same* equivalence class as f, and therefore that

there exists a matrix

$$\mathbf{A} = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \tag{3.2.14}$$

such that $f \cdot \mathbf{A} = f^*$. Of course, we are allowed to let $f^* = f$ in this setting, and then $\mathbf{A} \in \mathrm{Aut}\,(f)$. The proof of Theorem 3.1.11 gives a methodical analysis of the shape that $\mathbf{A}$ must take. For example, it is shown that we must have $t \neq u$. If $t < u$, there are three cases to consider. If $t = 0$ (Case I), we find that $\mathbf{A} = \mathbf{I}$. If $t < 0$ (Case II), we find that $f^* = f = f_j$ for some $j \in \mathbb{Z}^+$. A look back at the proof also shows that $\mathbf{A} = \prod_{i=0}^{j-1} \mathbf{S}(n_i)$ in this case. The only $j \in \mathbb{Z}^+$ such that $f_j = f$ are $j = m, 2m, 3m, \ldots$. If $j = m$, then $\mathbf{A} = \mathbf{U}_{m-1} = \mathbf{U}$. In general, if $j = km$ for some $k \in \mathbb{Z}^+$, then $\mathbf{A} = \mathbf{U}^k = \mathbf{U}_{km-1}$. It is of interest to note that for any given $\ell \in \mathbb{Z}^{\geq 0}$, the lower left entry of $\mathbf{U}_\ell$ is $-q_\ell$ by Lemma 3.2.1 and the lower right entry is $-q_{\ell-1}$. By Corollary 3.2.2, we have

$$q_{-1} = 0 < 1 = q_0 < q_1 < q_2 < \cdots, \tag{3.2.15}$$

and so $-q_\ell < -q_{\ell-1}$, as well as $-q_\ell < 0$, which shows the consistency with the assumptions in Case II. This also shows that $\mathbf{I}, \mathbf{U}_0, \mathbf{U}_1, \mathbf{U}_2, \ldots$ are all mutually distinct matrices, and so $\mathbf{I}, \mathbf{U}, \mathbf{U}^2, \mathbf{U}^3, \ldots$ are all mutually distinct as well. If $t > 0$ (Case III), we find that $\mathbf{A}^{-1} = \mathbf{U}^k$ for some $k \in \mathbb{Z}^+$, or that $\mathbf{A} = \mathbf{U}^{-k}$ for some $k \in \mathbb{Z}^+$. For any given $\ell \in \mathbb{Z}^{\geq 0}$, we note by Lemma 3.2.1 that

$$\mathbf{U}_\ell^{-1} = \begin{pmatrix} -q_{\ell-1} & -p_{\ell-1} \\ q_\ell & p_\ell \end{pmatrix}. \tag{3.2.16}$$

The lower left entry of $\mathbf{U}_\ell^{-1}$ is q_ℓ, which is *positive* for each $\ell \in \mathbb{Z}^{\geq 0}$ by (3.2.15), and thus no matrix $\mathbf{U}_\ell^{-1}$, with $\ell \in \mathbb{Z}^{\geq 0}$, falls into the set $\{\mathbf{I}, \mathbf{U}_0, \mathbf{U}_1, \ldots\}$. Also from

(3.2.15), we may conclude that $\mathbf{U}_0^{-1}, \mathbf{U}_1^{-1}, \mathbf{U}_2^{-1}, \dots$ are all mutually distinct among themselves. This in turn allows us to conclude that the matrices listed below

$$\{\dots, \mathbf{U}^{-2}, \mathbf{U}^{-1}, \mathbf{I}, \mathbf{U}, \mathbf{U}^2, \dots\} \tag{3.2.17}$$

are all mutually distinct from each other as well, meaning that set notation is justified here.

So far, we have assumed that the inequality $t < u$ holds in the matrix (3.2.14). If $t > u$ instead, then $-\mathbf{A} \in \mathrm{Aut}\,(f)$ has second row entries $-t$, $-u$, with $-t < -u$, which puts us back into the setting already covered above, and so $-\mathbf{A} = \mathbf{U}^k$ for some $k \in \mathbb{Z}$, or $\mathbf{A} = -\mathbf{U}^k$ for some $k \in \mathbb{Z}$. Every matrix of the form $-\mathbf{U}^k$, for $k \in \mathbb{Z}$, has a lower left entry greater than its lower right entry, and so none of these matrices fall into the set in (3.2.17). They are all distinct among themselves as well. We conclude that every $\mathbf{A} \in \mathrm{Aut}\,(f)$ is uniquely expressible either in the form $\mathbf{U}^k$, or as $-\mathbf{U}^k$, for some $k \in \mathbb{Z}$. □

Now that we have a complete description of the countably infinite set $\mathrm{Aut}\,(f)$, attached to any reduced form f of discriminant $D \in \mathbb{Z}^+$, we may derive a few important corollaries.

Corollary 3.2.5. *If f is a reduced form of discriminant $D \in \mathbb{Z}^+$, then the infinite group $\mathrm{Aut}\,(f)$ is abelian.*

Proof. There are technically four cases depending upon plus or minus signs. It suffices to consider one case. Assume that $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Aut}\,(f)$, with $\mathbf{A}_1 = \mathbf{U}^{k_1}$ and $\mathbf{A}_2 = -\mathbf{U}^{k_2}$, for integers $k_1, k_2 \in \mathbb{Z}$. Then $\mathbf{A}_1 \cdot \mathbf{A}_2 = \mathbf{U}^{k_1} \cdot \left(-\mathbf{I} \cdot \mathbf{U}^{k_2}\right) = (-\mathbf{I}) \cdot \left(\mathbf{U}^{k_1} \cdot \mathbf{U}^{k_2}\right)$, where (3.2.13) and the associative law were used in the second equality. Since we have $\mathbf{U}^{k_1} \cdot \mathbf{U}^{k_2} = \mathbf{U}^{k_1+k_2} = \mathbf{U}^{k_2} \cdot \mathbf{U}^{k_1}$, we may conclude that $\mathbf{A}_1 \cdot \mathbf{A}_2 = \mathbf{A}_2 \cdot \mathbf{A}_1$. □

Now let $f = f_0 = [a, b, c]$ be a form that is both primitive and reduced, of discriminant $D \in \mathbb{Z}^+$. Since we have an exact description of $\mathrm{Aut}\,(f)$ by Theorem 3.2.4, and f is primitive by assumption, we may directly apply Theorem 2.3.3 to obtain all solutions to the Fermat-Pell 4-Equation (2.3.1). Of particular interest is the minimal solution (t_1, u_1) of (2.3.1), as uniquely characterized in Definition 2.3.6. Looking at (2.3.3), it is apparent that in searching through all $\mathbf{A} \in \mathrm{Aut}\,(f)$, we wish to find the smallest positive lower left entry possible in $\mathbf{A}$, which we denote by q^*, in order to find u_1. Looking back through the proof of Theorem 3.2.4, and taking particular note of (3.2.15), it is evident that $q^* = q_{m-1}$, where $m \in \mathbb{Z}^+$ is the number of reduced forms in the cycle to which f belongs. These considerations furnish us with an algorithm for finding the minimal solution (t_1, u_1) of (2.3.1), which only requires having in hand the diagram (3.2.12) attached to a primitive cycle of discriminant $D \in \mathbb{Z}^+$. In Example 3.2.6 below, we present a concrete example illustrating this algorithm, and afterwards we give a more formal description of the algorithm itself.

Example 3.2.6. Consider the primitive and reduced form $f = [9, 17, 7]$ of discriminant $D = 37$. According to Table 3.1.3, the form f lies in a cycle of reduced forms of length $m = 7$. Starting with $f = f_0$, and going around the cycle exactly once, leads to the values displayed in Table 3.2.1 below.

Table 3.2.1. Sample computations illustrating Algorithm 3.2.7

j	-2	-1	0	1	2	3	4	5	6
n_j			2	2	2	3	7	3	2
p_j	0	1	2	3	4	9	59	168	277
q_j	-1	0	1	2	3	7	46	131	216

In this case, we have

$$\mathbf{U} = \mathbf{U}_6 = \begin{pmatrix} p_6 & p_5 \\ -q_6 & -q_5 \end{pmatrix} = \begin{pmatrix} 277 & 168 \\ -216 & -131 \end{pmatrix}.$$

Comparison with Theorem 2.3.3 gives $u = -q_6/a = -216/9 = -24$ as well as $t = p_6 + (-q_5) = 277 - 131 = 146$. In any such computation involving the matrix $\mathbf{U}_{m-1}$, the u–value computed in this way will always be negative, and so $u_1 = q_{m-1}/a$ is the sought-after least positive integer in the minimal solution pair (t_1, u_1) of (2.3.1). The quantity $t = p_{m-1} + (-q_{m-2})$ computed in this way will always be positive (see Section 4.1 for a proof of this fact) and so $t_1 = p_{m-1} + (-q_{m-2})$. The minimal solution pair $(t_1, u_1) = (146, 24)$ for $D = 37$ found here agrees with the answer found earlier in Table 2.3.1 by use of the brute-force algorithm described in Section 2.3.

Using Example 3.2.6 as inspiration, we may now give a concise description of an efficient and effective algorithm that can be used to calculate the minimal solution pair (t_1, u_1) of the Fermat-Pell 4-Equation (2.3.1).

Algorithm 3.2.7. *Let $f = f_0 = [a, b, c]$ be a primitive and reduced form of discriminant $D \in \mathbb{Z}^+$ lying in a cycle of reduced forms of length $m \in \mathbb{Z}^+$. If $n_0, \ldots, n_{m-1}$ are the integers (all in $\mathbb{Z}^{\geq 2}$) taking us around diagram* (3.2.12) *exactly once, we compute $p_0, \ldots, p_{m-1}$ and $q_0, \ldots, q_{m-1}$ recursively using* (3.2.3) *and* (3.2.4)*, respectively. Then $t_1 = p_{m-1} - q_{m-2}$ and $u_1 = q_{m-1}/a$.*

CHAPTER IV

REAL QUADRATIC IRRATIONALS AND MINUS CONTINUED FRACTIONS

4.1. Real Quadratic Irrationals and the Fundamental Unit

In this section, we offer a solution to Exercise 2 on page 138 of Zagier's book [8]. This exercise, as proposed in Zagier's book, deals specifically with fundamental discriminants (such discriminants are defined right below Definition 2.2.16), but our solution applies to any given discriminant $D \in \mathbb{Z}^+$, as long as we restrict ourselves to classes of forms containing only forms that are primitive.

In (2.3.20), we defined the fundamental unit $\varepsilon_1(D)$ associated to a fixed discriminant $D \in \mathbb{Z}^+$. According to this definition, we require the minimal solution (t_1, u_1) of the Fermat-Pell 4-Equation (2.3.1) in order to obtain this unit. However, there is an alternate method to compute $\varepsilon_1(D)$ involving primitive cycles of reduced forms of discriminant D, which is inspired by Zagier's Exercise 2. This method is based upon the following result.

Theorem 4.1.1. *Assume that $D \in \mathbb{Z}^+$ is a fixed discriminant and let $f_0, \ldots, f_{m-1}$ be the set of all reduced forms lying in a primitive cycle of discriminant D. Then*

$$\varepsilon_1(D) = Z(f_0) \cdots Z(f_{m-1}), \tag{4.1.1}$$

where the function Z is that given in Definition 2.4.2.

As an example, consider the forms $f_0 = [1, 5, 3]$, $f_1 = [3, 5, 1]$, and $f_2 = [3, 7, 3]$, which make up the principle cycle of discriminant 13 (clearly a primitive cycle as well). An

easy computation shows that

$$Z(f_0) \cdot Z(f_1) \cdot Z(f_2) = \frac{5+\sqrt{13}}{2} \cdot \frac{5+\sqrt{13}}{6} \cdot \frac{7+\sqrt{13}}{6} = \frac{11+3\sqrt{13}}{2} = \varepsilon_1(13).$$

In order to prove Theorem 4.1.1, we first need to establish some preliminary results that are important in their own right. Returning to the reduction theory of Section 3.1, we recall that if $f = f_0 = [a_0, b_0, c_0]$ is an arbitrary form of discriminant $D \in \mathbb{Z}^+$, then the reduction algorithm produces a uniquely determined infinite sequence of right neighboring forms $f_1 = [a_1, b_1, c_1], f_2 = [a_2, b_2, c_2], \ldots$, and the key ingredients of this sequence are encapsulated in the following diagram (the same diagram as already displayed in (3.2.1))

$$f_0 \xrightarrow{n_0} f_1 \xrightarrow{n_1} f_2 \xrightarrow{n_2} \cdots . \tag{4.1.2}$$

We have $f_0 \cdot \mathbf{S}(n_0) = f_1$, $f_0 \cdot \mathbf{S}(n_0) \cdot \mathbf{S}(n_1) = f_2$, and in general $f_0 \cdot \mathbf{U}_j = f_{j+1}$ for $j = 0, 1, 2, \ldots$, making use of the notation introduced in Lemma 3.2.1. Associated to each form f_j, for $j = 0, 1, 2, \ldots$, is a uniquely defined real quadratic irrational number β_j defined by

$$\beta_j = \frac{b_j + \sqrt{D}}{2a_j}. \tag{4.1.3}$$

We also use the notation $\beta_j = Z(f_j)$, introduced in Definition 2.4.2. With regard to diagram (4.1.2), we recall from (3.1.8) that $n_j = \lceil Z(f_j) \rceil$ for $j = 0, 1, 2, \ldots$. This leads us to introduce a new diagram involving the β's which matches step by step with diagram (4.1.2):

$$\beta_0 \xrightarrow{n_0} \beta_1 \xrightarrow{n_1} \beta_2 \xrightarrow{n_2} \cdots . \tag{4.1.4}$$

We now show that if we start with the real quadratic irrational number

$$\beta = \beta_0 = \frac{b_0 + \sqrt{D}}{2a_0}, \tag{4.1.5}$$

then diagram (4.1.4) may be constructed from scratch (with respect to both the β's and the n's) using a well-defined algorithm which is identical to the generation of the minus continued fraction expansion of the real number β. The first step of the minus continued fraction algorithm, as applied to β (see Section 4.2), is to compute the number $n_0^* = \lfloor \beta_0 \rfloor + 1$. Since β_0, as an irrational number, lies strictly between two integers, we have $n_0^* = \lceil \beta_0 \rceil$, and so n_0^* has the same value as n_0 in (4.1.2). Since $f_0 \cdot \mathbf{S}(n_0) = f_1$, we have $\beta_1 = \mathbf{S}(n_0)^w \cdot \beta_0$ by (2.4.12). Writing this out explicitly gives

$$\beta_1 = \begin{pmatrix} 0 & 1 \\ -1 & n_0 \end{pmatrix} \cdot \beta_0 = \frac{1}{-\beta_0 + n_0}, \tag{4.1.6}$$

which is an exact match with (4.2.1). Since β_1 is also an irrational number, the same argument gives $n_1 = \lceil \beta_1 \rceil = \lfloor \beta_1 \rfloor + 1$, as well as $\beta_2 = 1/(n_1 - \beta_1)$. Because *all* of the β's corresponding to the f's in (4.1.2) are irrational numbers, continuing in this way reproduces precisely the same infinite sequences of β's and n's given by the minus continued fraction algorithm defined by (4.2.1) and (4.2.2). This shows that diagram (4.1.4) is generated directly upon application of the minus continued fraction algorithm to the real number β_0 given in (4.1.5). It is worth noting that the minus continued fraction algorithm is applicable to *any* real number β, so that the present section concerns a special (though highly important) case of a much more general algorithm. The following lemma is crucial to the proof of Theorem 4.1.1.

Lemma 4.1.2. *Given the diagrams* (4.1.2) *and* (4.1.4)*, and* (4.1.5)*, we have*

$$\beta_j = \frac{-q_{j-2}\beta + p_{j-2}}{-q_{j-1}\beta + p_{j-1}} \quad \textit{for } j = 0, 1, 2, \ldots, \tag{4.1.7}$$

where the p's and q's are exactly those defined within the statement of Lemma 3.2.1.

Proof. For $j = 0$, (4.1.7) translates into the statement $\beta_0 = \beta$, which is correct. For $j \in \mathbb{Z}^+$, we make use of Corollary 2.4.8, which shows that since $f_0 \cdot \mathbf{U}_{j-1} = f_j$, we then have

$$Z(f_j) = \beta_j = (\mathbf{U}_{j-1})^w \cdot \beta.$$

By (3.2.5), we have

$$(\mathbf{U}_{j-1})^w = \begin{pmatrix} -q_{j-2} & p_{j-2} \\ -q_{j-1} & p_{j-1} \end{pmatrix},$$

from which (4.1.7) follows. □

Before moving on to the proof of Theorem 4.1.1, we quickly derive two relations which are special cases of results which are found to appear in Section 4.2. Going back to (3.2.5), since $\mathbf{U}_j \in \mathrm{SL}_2(\mathbb{Z})$ for $j = 0, 1, 2 \ldots$, we note that (see (4.2.14))

$$p_{j-1}q_j - p_j q_{j-1} = 1 \quad \text{for } j = 0, 1, 2, \ldots. \tag{4.1.8}$$

It is easy to check that (4.1.8) also holds for $j = -1$. We next show that (4.1.7) leads directly to (4.2.17). Fixing $j \in \mathbb{Z}^+$, (4.1.7) gives $-q_{j-1}\beta_j\beta + p_{j-1}\beta_j = -q_{j-2}\beta + p_{j-2}$, and solving for β gives (see (4.2.17))

$$\beta = \frac{p_{j-1}\beta_j - p_{j-2}}{q_{j-1}\beta_j - q_{j-2}}. \tag{4.1.9}$$

Proof of Theorem 4.1.1. In this proof, it is beneficial to work instead with the product $\beta_1 \cdots \beta_m$, which is equal to the product on the right hand side of (4.1.1) since $\beta_0 = \beta_m$.

If we write down explicitly the equations in (4.1.7) for $j = 1, 2, 3$, we see something of immediate interest:

$$\beta_1 = \frac{1}{-q_0\beta + p_0}, \qquad \beta_2 = \frac{-q_0\beta + p_0}{-q_1\beta + p_1}, \qquad \beta_3 = \frac{-q_1\beta + p_1}{-q_2\beta + p_2}, \tag{4.1.10}$$

which shows that the denominator of β_j is equal to the numerator of β_{j+1} for $j = 1, 2, 3, \ldots$. We readily conclude from (4.1.10) that

$$\beta_1 \cdots \beta_m = \frac{1}{p_{m-1} - q_{m-1}\beta}\ , \tag{4.1.11}$$

so we just need to rework the expression on the right hand side of (4.1.11). Our goal, to complete the proof, is to show that

$$\frac{1}{p_{m-1} - q_{m-1}\beta} = \frac{t_1 + u_1\sqrt{D}}{2}, \tag{4.1.12}$$

where (t_1, u_1) is the minimal solution of (2.3.1). Recall that t_1 and u_1 are both positive integers by definition. We are assuming here that $f = f_0 = [a, b, c]$ is a primitive and reduced form of discriminant $D \in \mathbb{Z}^+$ and that

$$\beta = \beta_0 = \frac{b + \sqrt{D}}{2a}. \tag{4.1.13}$$

Recall from Section 3.2 that

$$au = -q_{m-1}, \tag{4.1.14}$$

where u is a negative integer and that

$$u_1 = -u. \tag{4.1.15}$$

We also have

$$p_{m-1} = \frac{t - bu}{2}, \tag{4.1.16}$$

as well as

$$p_{m-1} - q_{m-2} = t. \tag{4.1.17}$$

As of right now, we do not know whether the value of t in (4.1.17) is positive or negative, but we do know that

$$t_1 = |t| \neq 0. \tag{4.1.18}$$

Plugging (4.1.13), (4.1.14), and (4.1.16) into the expression on the left hand side of (4.1.12) gives

$$\begin{aligned}
\frac{1}{p_{m-1} - q_{m-1}\beta} &= \frac{1}{\frac{(t-bu)}{2} + au\left[\frac{b+\sqrt{D}}{2a}\right]} \\
&= \frac{2}{t - bu + bu + u\sqrt{D}} \\
&= \frac{2}{(t + u\sqrt{D})} \cdot \frac{(t - u\sqrt{D})}{(t - u\sqrt{D})} \\
&= \frac{t - u\sqrt{D}}{2},
\end{aligned}$$

with the last equality holding since (t, u) is an integer pair solution to (2.3.1). So far, we have

$$\beta_1 \cdots \beta_m = \frac{t - u\sqrt{D}}{2}, \tag{4.1.19}$$

and because of (4.1.15) it only remains to prove that t is a positive integer in order to finally establish (4.1.12), in light of (4.1.18). By Lemma 3.1.5, we note that $1 < \beta_j$ for $j = 1, \ldots, m$, and so by (4.1.19) we have

$$1 < \frac{t + u_1\sqrt{D}}{2}. \tag{4.1.20}$$

Since $4 = t_1^2 - Du_1^2$, we have $0 < t_1^2 - Du_1^2$, which implies that $u_1\sqrt{D} < t_1$, or $-t_1 + u_1\sqrt{D} < 0$. We have either $t = t_1$ or $t = -t_1$. If $t = -t_1$ held, we would conclude

that $t + u_1\sqrt{D} < 0$, in contradiction to (4.1.20). We therefore must have $t = t_1 \in \mathbb{Z}^+$, which completes the proof of Theorem 4.1.1. Knowing that t, as defined by (4.1.17), is positive also consolidates the validity of Algorithm 3.2.7, clearing up the one point still in question regarding this algorithm. □

4.2. The Theory of Minus Continued Fractions

The theory of positive continued fractions assumed its classic form long ago and expositions of this theory may be found in most books that cover Elementary Number Theory. The theory of minus continued fractions is of more recent vintage and is definitely less well known; however, this theory has all of the same remarkable features as its classical counterpart and it is the theory of continued fractions directly connected to Zagier's reduction theory of indefinite binary quadratic forms, as already noted in Section 4.1. Given that the theory of minus continued fractions is not as widely known as it deserves, we give a detailed exposition of it here following the very nice presentation of Katok [4].

In the following, let $\beta = \beta_0 \in \mathbb{R}$ be any given fixed real number. We may define an infinite sequence of integers $n_0, n_1, n_2, \ldots$, and an infinite sequence of real numbers $\beta_0, \beta_1, \beta_2, \ldots$ inductively and uniquely by use of the following system of equations:

$$n_0 = \lfloor \beta_0 \rfloor + 1, \quad \beta_1 = \frac{1}{n_0 - \beta_0}, \tag{4.2.1}$$

$$n_j = \lfloor \beta_j \rfloor + 1, \quad \beta_{j+1} = \frac{1}{n_j - \beta_j} \quad \text{for } j = 1, 2, 3, \ldots. \tag{4.2.2}$$

Note that $n_0 > \beta_0$ and $n_j > \beta_j$ for $j = 1, 2, 3, \ldots$, and so each quotient in (4.2.1) and (4.2.2) is well-defined and strictly positive so that $\beta_1, \beta_2, \ldots \in \mathbb{R}^{>0}$. Also, note that $0 < n_0 - \beta_0 \leq 1$, which implies that $1 \leq \beta_1$; we have $\beta_1 = 1$ if and only if $\beta \in \mathbb{Z}$. This implies in turn that $n_1 \geq 2$. Now, assume that $j \in \mathbb{Z}^+$ is fixed. We

have $0 < n_j - \beta_j \leq 1$, which implies that $1 \leq \beta_{j+1}$, again with $\beta_{j+1} = 1$ if and only if $\beta_j \in \mathbb{Z}$. This in turn implies that $n_{j+1} \geq 2$. The following general fact has now been established:

$$n_j \in \mathbb{Z}^{\geq 2} \quad \text{for } j = 1, 2, 3, \ldots. \tag{4.2.3}$$

We next define a corresponding infinite sequence of rational numbers as follows:

$$r_0 = n_0; \quad r_1 = n_0 - \frac{1}{n_1}; \quad r_2 = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2}}; \tag{4.2.4}$$

and in general,

$$r_j = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots \; n_{j-2} - \cfrac{1}{n_{j-1} - \cfrac{1}{n_j}}}}}. \tag{4.2.5}$$

To verify that we completely avoid division by zero when forming such complicated quotients, we first take a more careful look at the formation of these expressions. Consider the value r_3. By (4.2.3), we see that

$$1 < n_2 - \frac{1}{n_3},$$

and so

$$0 < \cfrac{1}{n_2 - \cfrac{1}{n_3}} < 1.$$

Again, by (4.2.3), we have

$$1 < n_1 - \cfrac{1}{n_2 - \cfrac{1}{n_3}},$$

and so

$$0 < \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{n_3}}} < 1.$$

For r_j with $j \geq 4$ fixed, we have the following inequalities for the expression at the "bottom" :

$$0 < \cfrac{1}{n_{j-2} - \cfrac{1}{n_{j-1} - \cfrac{1}{n_j}}} < 1,$$

which holds by the same reasoning as above. We then subtract this number from $n_{j-3} \geq 2$ to obtain a number that is greater than 1, take the reciprocal to obtain a number strictly between 0 and 1, and continue in this way, avoiding division by zero in every case.

Given an infinite sequence of integers $n_0, n_1, n_2, \ldots$ with $n_0 \in \mathbb{Z}$ arbitrary and $n_j \in \mathbb{Z}^{\geq 2}$ for $j = 1, 2, 3, \ldots$, we introduce the following notation based upon (4.2.4) and (4.2.5) above:

$$r_0 = (n_0) = n_0; \quad r_1 = (n_0, n_1) = n_0 - \frac{1}{n_1}; \quad r_2 = (n_0, n_1, n_2) = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2}};$$

and in general,

$$r_j = (n_0, n_1, n_2, \ldots, n_j) = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots \; n_{j-2} - \cfrac{1}{n_{j-1} - \cfrac{1}{n_j}}}}} \quad \text{for } j = 3, 4, 5, \ldots.$$

It is clear that $r_0, r_1, r_2, \ldots$ are all rational numbers since $\mathbb{Q}$ forms a field. Our primary goal in this section is the proof of the following theorem.

Theorem 4.2.1. *Given an arbitrary real number $\beta \in \mathbb{R}$, along with the corresponding uniquely defined infinite sequence of integers $n_0, n_1, n_2, \ldots$ and rational numbers $r_0, r_1, r_2, \ldots$ as defined above, we have $\lim_{j\to\infty} r_j = \beta$.*

Proof. We first define two new infinite sequences of integers: $p_{-2}, p_{-1}, p_0, p_1, \ldots$ and $q_{-2}, q_{-1}, q_0, q_1, \ldots$ inductively from the n's as follows:

$$p_{-2} = 0, \quad p_{-1} = 1, \quad p_j = n_j p_{j-1} - p_{j-2} \quad \text{for } j = 0, 1, 2, \ldots \tag{4.2.6}$$

$$q_{-2} = -1, \quad q_{-1} = 0, \quad q_j = n_j q_{j-1} - q_{j-2} \quad \text{for } j = 0, 1, 2, \ldots. \tag{4.2.7}$$

We will eventually prove that $r_j = p_j/q_j$ for $j = 0, 1, 2, \ldots$, but we first wish to show that

$$1 = q_0 < q_1 < q_2 < \cdots, \quad \text{which implies in turn that } \lim_{j\to\infty} q_j = \infty. \tag{4.2.7a}$$

We first note that

$$q_0 = n_0 q_{-1} - q_{-2} = 1 \quad \text{and} \quad q_1 = n_1 q_0 - q_{-1} = n_1,$$

and since $2 \leq n_1$, we see that $q_0 = 1 < q_1$. We now proceed by induction. Assume $j \geq 2$ is given, and that $0 < q_{j-2} < q_{j-1}$ holds. We need to establish that $q_{j-1} < q_j$ follows. From (4.2.3), we know that $2 \leq n_j$, and so $2q_{j-1} \leq n_j q_{j-1}$. By the inductive hypothesis, we have

$$q_{j-2} + q_{j-1} < q_{j-1} + q_{j-1} = 2q_{j-1},$$

and so

$$q_{j-2} + q_{j-1} < n_j q_{j-1}.$$

This implies that

$$q_{j-1} < n_j q_{j-1} - q_{j-2} = q_j,$$

or that $q_{j-1} < q_j$, which completes the proof of (4.2.7a).

In order to mimic the usual theory of positive continued fractions, we now introduce polynomials of several variables having similar properties to the usual "continuant polynomials." These are defined as follows:

$$L_0(\ ) = 1; \tag{4.2.8a}$$

$$L_1(x_1) = x_1; \quad \text{and in general for } j = 2, 3, 4, \ldots, \tag{4.2.8b}$$

$$L_j(x_1, \ldots, x_j) = x_j L_{j-1}(x_1, \ldots, x_{j-1}) - L_{j-2}(x_1, \ldots, x_{j-2}). \tag{4.2.8}$$

For example, when $j = 2$, we have

$$L_2(x_1, x_2) = x_2 L_1(x_1) - L_0(\ ) = x_2 x_1 - 1$$

(note that $L_0(\ ) = 1$ is a polynomial dependent upon zero variables). When $j = 3$, we have

$$\begin{aligned} L_3(x_1, x_2, x_3) &= x_3 L_2(x_1, x_2) - L_1(x_1) \\ &= x_3(x_2 x_1 - 1) - x_1 \\ &= x_1 x_2 x_3 - x_1 - x_3, \end{aligned}$$

and when $j = 4$, we have

$$\begin{aligned} L_4(x_1, x_2, x_3, x_4) &= x_4 L_3(x_1, x_2, x_3) - L_2(x_1, x_2) \\ &= x_4[x_1 x_2 x_3 - x_1 - x_3] - (x_1 x_2 - 1) \\ &= x_1 x_2 x_3 x_4 - x_1 x_4 - x_3 x_4 - x_1 x_2 + 1. \end{aligned}$$

Note that $q_0 = 1 = L_0(\,)$, and that $q_1 = n_1 = L_1(n_1)$; we claim in general that

$$q_j = L_j(n_1, \ldots, n_j) \quad \text{for } j = 2, 3, 4, \ldots. \tag{4.2.9}$$

We have

$$q_2 = n_2 q_1 - q_0 = n_2 n_1 - 1 = L_2(n_1, n_2),$$

and we now proceed by induction. Assume that $j \geq 3$ is given, and that both $q_{j-2} = L_{j-2}(n_1, \ldots, n_{j-2})$ and $q_{j-1} = L_{j-1}(n_1, \ldots, n_{j-1})$ hold. By (4.2.8), we have

$$\begin{aligned} L_j(n_1, \ldots, n_j) &= n_j L_{j-1}(n_1, \ldots, n_{j-1}) - L_{j-2}(n_1, \ldots, n_{j-2}) \\ &= n_j q_{j-1} - q_{j-2} = q_j, \end{aligned}$$

with the last equality holding by (4.2.7), which completes the inductive step.

In a similar way, we now show that $p_0 = L_1(n_0)$, $p_1 = L_2(n_0, n_1)$, and, in general,

$$p_j = L_{j+1}(n_0, \ldots, n_j) \quad \text{for } j = 2, 3, 4, \ldots. \tag{4.2.10}$$

We first note that

$$p_0 = n_0 p_{-1} - p_{-2} = n_0 = L_1(n_0),$$

and

$$p_1 = n_1 p_0 - p_{-1} = n_1 n_0 - 1 = L_2(n_0, n_1).$$

Now, assume that $j \geq 2$ is given, and that both $p_{j-2} = L_{j-1}(n_0, \ldots, n_{j-2})$ and $p_{j-1} = L_j(n_0, \ldots, n_{j-1})$ hold. By (4.2.8), we have

$$\begin{aligned} L_{j+1}(n_0, \ldots, n_j) &= n_j L_j(n_0, \ldots, n_{j-1}) - L_{j-1}(n_0, \ldots, n_{j-2}) \\ &= n_j p_{j-1} - p_{j-2} = p_j, \end{aligned}$$

with the last equality holding by (4.2.6), which completes the inductive step.

Next, we move towards proving that

$$r_j = \frac{p_j}{q_j} \quad \text{for } j = 0, 1, 2, \ldots. \tag{4.2.11}$$

The first few cases are easy to verify:

$$r_0 = n_0 = \frac{n_0}{1} = \frac{p_0}{q_0};$$

$$r_1 = \frac{n_0 n_1 - 1}{n_1} = \frac{p_1}{q_1};$$

and using (4.2.9) and (4.2.10), we obtain:

$$\begin{aligned} r_2 = n_0 - \frac{1}{\dfrac{n_1 n_2 - 1}{n_2}} &= n_0 - \frac{n_2}{n_1 n_2 - 1} \\ = \frac{n_0 (n_1 n_2 - 1) - n_2}{n_1 n_2 - 1} &= \frac{n_0 n_1 n_2 - n_0 - n_2}{n_1 n_2 - 1} \\ = \frac{L_3 (n_0, n_1, n_2)}{L_2 (n_1, n_2)} &= \frac{p_2}{q_2}. \end{aligned}$$

Our goal is to prove, in general, that

$$r_j = \frac{L_{j+1} (n_0, \ldots, n_j)}{L_j (n_1, \ldots, n_j)} \quad \text{for } j = 1, 2, 3, \ldots. \tag{4.2.12}$$

In combination with (4.2.9) and (4.2.10), we see that (4.2.11) is an immediate consequence of (4.2.12). We have already verified (4.2.12) for $j = 1$ and $j = 2$. We will prove the general case by induction. Assume that $j \geq 3$ is given, and that (4.2.12) holds for $j - 1$; that is,

$$r_{j-1} = (n_0, \ldots, n_{j-1}) = \frac{L_j (n_0, \ldots, n_{j-1})}{L_{j-1} (n_1, \ldots, n_{j-1})}. \tag{4.2.13}$$

Looking at the expression of r_j in (4.2.5), we first note that r_j may be rewritten as a

continued fraction which depends only upon j terms instead of the usual $j+1$ terms:

$$r_j = (n_0, \ldots, n_{j-2}, y), \quad \text{where } \; y = n_{j-1} - \frac{1}{n_j}.$$

By the inductive hypothesis (4.2.13), which applies to an expression with j terms in its continued fraction, we have

$$\begin{aligned} r_j &= \frac{L_j(n_0, \ldots, n_{j-2}, y)}{L_{j-1}(n_1, \ldots, n_{j-2}, y)} \\ &= \frac{yL_{j-1}(n_0, \ldots, n_{j-2}) - L_{j-2}(n_0, \ldots, n_{j-3})}{yL_{j-2}(n_1, \ldots, n_{j-2}) - L_{j-3}(n_1, \ldots, n_{j-3})}, \end{aligned}$$

with the last equality holding by (4.2.8). We note that if $j = 3$, then the polynomial $L_{j-3}(n_1, \ldots, n_{j-3})$ is equal to $L_0(\,) = 1$. Using our expression for $y = n_{j-1} - (1/n_j)$, we have

$$r_j = \frac{\left(n_{j-1} - \frac{1}{n_j}\right) L_{j-1}(n_0, \ldots, n_{j-2}) - L_{j-2}(n_0, \ldots, n_{j-3})}{\left(n_{j-1} - \frac{1}{n_j}\right) L_{j-2}(n_1, \ldots, n_{j-2}) - L_{j-3}(n_1, \ldots, n_{j-3})},$$

or

$$\begin{aligned} r_j &= \frac{n_{j-1}L_{j-1}(n_0, \ldots, n_{j-2}) - L_{j-2}(n_0, \ldots, n_{j-3}) - \frac{1}{n_j}L_{j-1}(n_0, \ldots, n_{j-2})}{n_{j-1}L_{j-2}(n_1, \ldots, n_{j-2}) - L_{j-3}(n_1, \ldots, n_{j-3}) - \frac{1}{n_j}L_{j-2}(n_1, \ldots, n_{j-2})} \\ &= \frac{L_j(n_0, \ldots, n_{j-1}) - \frac{1}{n_j}L_{j-1}(n_0, \ldots, n_{j-2})}{L_{j-1}(n_1, \ldots, n_{j-1}) - \frac{1}{n_j}L_{j-2}(n_1, \ldots, n_{j-2})}, \end{aligned}$$

with the last equality holding again by (4.2.8). Multiplying the top and bottom of this last quotient by n_j gives

$$r_j = \frac{n_jL_j(n_0, \ldots, n_{j-1}) - L_{j-1}(n_0, \ldots, n_{j-2})}{n_jL_{j-1}(n_1, \ldots, n_{j-1}) - L_{j-2}(n_1, \ldots, n_{j-2})},$$

or finally, by (4.2.8),

$$r_j = \frac{L_{j+1}(n_0, \ldots, n_j)}{L_j(n_1, \ldots, n_j)}.$$

This completes the induction process, and establishes (4.2.12) for $j = 3, 4, 5, \ldots$.

We now wish to show that

$$p_{j-1}q_j - p_j q_{j-1} = 1 \quad \text{for } j = -1, 0, 1, 2, \ldots. \tag{4.2.14}$$

We begin by checking a few base cases. For $j = -1$, we have

$$p_{-2}q_{-1} - p_{-1}q_{-2} = 0 \cdot 0 - (1)(-1) = 1,$$

and for $j = 0$, we have

$$p_{-1}q_0 - p_0 q_{-1} = 1 \cdot 1 - p_0 \cdot 0 = 1.$$

Now, assume that $j \in \mathbb{Z}^+$ is fixed. From (4.2.6) and (4.2.7), we obtain

$$\begin{aligned} p_{j-1}q_j - p_j q_{j-1} &= p_{j-1}\left(n_j q_{j-1} - q_{j-2}\right) - \left(n_j p_{j-1} - p_{j-2}\right) q_{j-1} \\ &= -p_{j-1}q_{j-2} + p_{j-2}q_{j-1} \\ &= p_{j-2}q_{j-1} - p_{j-1}q_{j-2}. \end{aligned}$$

This shows that if $p_{j-2}q_{j-1} - p_{j-1}q_{j-2} = 1$, then we also obtain $p_{j-1}q_j - p_j q_{j-1} = 1$. Having already established the base cases above, we see that (4.2.14) holds for all $j = -1, 0, 1, 2, \ldots$.

From (4.2.14), we may now prove with ease that the infinite sequence of rational numbers $r_0, r_1, r_2, \ldots$ is *strictly* decreasing. Assuming that $j \in \mathbb{Z}^+$ is fixed, we already proved in (4.2.7a) that the product $q_j \cdot q_{j-1}$ is a *positive* integer. Dividing both sides of (4.2.14) by $q_j q_{j-1}$ gives

$$\frac{p_{j-1}}{q_{j-1}} - \frac{p_j}{q_j} = \frac{1}{q_j q_{j-1}} > 0.$$

When $j = 1$, we have $r_0 - r_1 > 0$ by (4.2.11); when $j = 2$, we have $r_1 - r_2 > 0$; and in

general, we have $r_k > r_{k+1}$ for $k = 0, 1, 2, \ldots$. We now wish to show that the strictly decreasing sequence $r_0 > r_1 > r_2 > r_3 > \cdots$ is bounded below by $n_0 - 1$. Since $r_0 = n_0 > n_0 - 1$ clearly holds, we just need to show that $n_0 - 1 < r_j$ for all $j \in \mathbb{Z}^+$. By (4.2.5), we may write r_j for each $j \in \mathbb{Z}^+$ in the form $r_j = n_0 - x_j$, where

$$x_j = \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots \; n_{j-2} - \cfrac{1}{n_{j-1} - \cfrac{1}{n_j}}}}}.$$

The proof that $n_0 - 1 < r_j = n_0 - x_j$ for each $j \in \mathbb{Z}^+$ is equivalent to showing that $x_j < 1$. We will prove even more strictly that $0 < x_j < 1$ for $j = 1, 2, 3, \ldots$. We begin by examining a few cases directly. For $j = 1$, we have $x_1 = 1/n_1$, and since $n_1 \in \mathbb{Z}^{\geq 2}$, we have $0 < x_1 < 1$. If $j = 2$, then

$$x_2 = \cfrac{1}{n_1 - \cfrac{1}{n_2}},$$

and since $n_1, n_2 \in \mathbb{Z}^{\geq 2}$, we have $1 < n_1 - (1/n_2)$, and so $0 < x_2 < 1$. The case $j = 3$ was already covered just below equation (4.2.5), and by the same reasoning, we may conclude that $0 < x_j < 1$ for all $j \in \mathbb{Z}^{\geq 4}$ as well, proving that $n_0 - 1 < r_j$ for all $j = 0, 1, 2, 3, \ldots$.

The completeness axiom for the real number system guarantees that a strictly decreasing infinite sequence of real numbers that is bounded below converges to a unique real number. In order to complete the proof of Theorem 4.2.1, it only remains to show that the limit value of the infinite sequence of rational numbers r_0, r_1, $r_2, \ldots$ is exactly equal to the real number β with which we started.

We recall from the beginning of this section that $1 \leq \beta_j$ for all $j = 1, 2, 3, \ldots$. We claim that

$$1 \leq \beta_j q_{j-1} - q_{j-2} \quad \text{for } j = 1, 2, 3, \ldots \tag{4.2.15}$$

as well. We first verify (4.2.15) directly for the two smallest values of j. For $j = 1$, we have $\beta_1 q_0 - q_{-1} = \beta_1$, and since $1 \leq \beta_1$, this case is done. For $j = 2$, since $1 \leq \beta_2$ and $2 \leq n_1 = q_1$, we have $2 \leq \beta_2 q_1$, which implies that $1 \leq \beta_2 q_1 - 1 = \beta_2 q_1 - q_0$, and so this case is verified. Now, assume that $j \in \mathbb{Z}^{\geq 3}$ is fixed. From the strict inequalities in (4.2.7a), and the fact that all of the q's are *integers*, we may conclude that $q_{j-2} + 1 \leq q_{j-1}$. Since $q_{j-1} \in \mathbb{Z}^+$ and $1 \leq \beta_j$, we have $q_{j-1} \leq \beta_j q_{j-1}$. Combining these inequalities gives $q_{j-2} + 1 \leq \beta_j q_{j-1}$, or

$$1 \leq \beta_j q_{j-1} - q_{j-2},$$

which establishes the claim in (4.2.15) for $j = 1, 2, 3, \ldots$.

From (4.2.1), we have $n_0 - \beta = 1/\beta_1$, or $\beta - n_0 = -(1/\beta_1)$, and so

$$\beta = n_0 - \frac{1}{\beta_1} = (n_0, \beta_1)\,,$$

where (n_0, β_1) refers to the continued fraction notation that was introduced right before the statement of Theorem 4.2.1. From (4.2.2), we have $n_1 - \beta_1 = 1/\beta_2$, or $\beta_1 - n_1 = -(1/\beta_2)$, and so

$$\beta_1 = n_1 - \frac{1}{\beta_2}.$$

Combining this equation with the equation above for β gives

$$\beta = n_0 - \cfrac{1}{n_1 - \cfrac{1}{\beta_2}} = (n_0, n_1, \beta_2)\,.$$

Similarly, we have $\beta_2 = n_2 - (1/\beta_3)$, and so $\beta = (n_0, n_1, n_2, \beta_3)$. In general we have

$\beta_j = n_j - (1/\beta_{j+1})$ for $j = 1, 2, 3, \ldots$, and

$$\beta = (n_0, \ldots, n_{j-1}, \beta_j) \quad \text{for } j = 1, 2, 3, \ldots. \tag{4.2.16}$$

Now, fix $k \in \mathbb{Z}^+$. By (4.2.11), we have

$$r_{k-1} = (n_0, \ldots, n_{k-1}) = \frac{p_{k-1}}{q_{k-1}},$$

which shows that p_{k-1} and q_{k-1} depend strictly upon the first k integers $n_0, \ldots, n_{k-1}$ in the infinite sequence of n's. For $r_k = (n_0, \ldots, n_{k-1}, n_k)$, we have

$$r_k = \frac{p_k}{q_k} = \frac{n_k p_{k-1} - p_{k-2}}{n_k q_{k-1} - q_{k-2}},$$

by (4.2.6) and (4.2.7). We found in (4.2.16) that $\beta = (n_0, \ldots, n_{k-1}, \beta_k)$, and if we compare to the expression for r_k above and replace the n_k there by the β_k in the continued fraction expression for β, we deduce that

$$\beta = \frac{\beta_k p_{k-1} - p_{k-2}}{\beta_k q_{k-1} - q_{k-2}}. \tag{4.2.17}$$

For example, if $k = 1$, then (4.2.17) becomes

$$\beta = \frac{\beta_1 p_0 - p_{-1}}{\beta_1 q_0 - q_{-1}} = \frac{\beta_1 n_0 - 1}{\beta_1},$$

which is correct.

We are finally ready to prove that $\lim_{j\to\infty} r_j = \beta$. For any fixed $k \in \mathbb{Z}^+$, by (4.2.11) and (4.2.17), we have

$$r_{k-1} - \beta = \frac{p_{k-1}}{q_{k-1}} - \beta = \frac{p_{k-1}}{q_{k-1}} - \frac{\beta_k p_{k-1} - p_{k-2}}{\beta_k q_{k-1} - q_{k-2}}.$$

Forming a common denominator gives

$$\begin{aligned} r_{k-1} - \beta &= \frac{p_{k-1}\left(\beta_k q_{k-1} - q_{k-2}\right) - q_{k-1}\left(\beta_k p_{k-1} - p_{k-2}\right)}{q_{k-1}\left(\beta_k q_{k-1} - q_{k-2}\right)} \\ &= \frac{p_{k-2}q_{k-1} - p_{k-1}q_{k-2}}{q_{k-1}\left(\beta_k q_{k-1} - q_{k-2}\right)}, \end{aligned}$$

or finally

$$r_{k-1} - \beta = \frac{1}{q_{k-1}\left(\beta_k q_{k-1} - q_{k-2}\right)}, \tag{4.2.18}$$

with this last equality following from (4.2.14). Given the inequalities in (4.2.15), we know that the number $\beta_k q_{k-1} - q_{k-2}$ in the denominator on the right side of (4.2.18) is positive. Since q_j is positive for every $j \in \mathbb{Z}^{\geq 0}$, we see that the number $r_{k-1} - \beta$ in (4.2.18) is also positive. By (4.2.15), we have

$$0 < \frac{1}{\beta_k q_{k-1} - q_{k-2}} \leq 1,$$

and the preceding comments combined with (4.2.18) allow us to deduce that

$$|r_{k-1} - \beta| = r_{k-1} - \beta \leq \frac{1}{q_{k-1}},$$

or

$$|r_j - \beta| = r_j - \beta \leq \frac{1}{q_j} \quad \text{for } j = 0, 1, 2, \ldots. \tag{4.2.19}$$

This shows that $\beta < \cdots < r_j < \cdots < r_1 < r_0$, since we already know that the infinite sequence of r's is strictly decreasing. Combined with the limit $\lim_{j\to\infty} q_j = \infty$ in (4.2.7a), we conclude from (4.2.19) that $\lim_{j\to\infty} r_j = \beta$. This completes the proof of Theorem 4.2.1. □

Given the notation introduced earlier in this section that $r_j = (n_0, \ldots, n_j)$ for $j = 0, 1, 2, \ldots$, and the result just proven that $\lim_{j\to\infty} r_j = \beta$, it makes sense to

introduce one last piece of notation, declaring that

$$\beta = (n_0, n_1, n_2, \dots)\,. \tag{4.2.20}$$

Given the behavior of the rational numbers $r_0, r_1, r_2, \dots$ with respect to the limit value β, it is natural to call these rational numbers "the convergents to β". A concrete example illustrating the numerics of the convergence process in the employ of minus continued fractions is provided at the end of Section 5.2.

CHAPTER V

MINUS CONTINUED FRACTIONS AND THE FERMAT-PELL 1-EQUATION

5.1. A Modified Version of the English Method

This section is motivated by the presentation of the "English method", due to Brouncker and Wallis, found in the exercises to §1.9 of the book by Edwards [2]. Their method is directly connected to the classical theory of continued fractions, and our modification consists in revamping their algorithm by connecting it with the theory of minus continued fractions instead.

In the following, let $d \in \mathbb{Z}^+$ denote a fixed positive integer that is not a perfect square. We wish to find nontrivial integer pair solutions, with $y > 0$, to the corresponding Fermat-Pell 1-Equation

$$x^2 - dy^2 = 1 \tag{5.1.1}$$

through the use of the English method. The connection of this algorithmic method to Zagier's reduction theory for indefinite binary quadratic forms as well as to the theory of minus continued fractions is made in Section 5.2.

We initiate our algorithmic approach by first setting

$$p_{-2} = 0;\ \ q_{-2} = -1;\ \ k_{-2} = -d;\ \ p_{-1} = 1;\ \ q_{-1} = 0;\ \ k_{-1} = 1;\ \ r_{-1} = 0. \tag{5.1.2}$$

By construction, we have

$$p_{-2}^2 - dq_{-2}^2 = k_{-2}, \tag{5.1.3a}$$

and

$$p_{-1}^2 - dq_{-1}^2 = k_{-1}. \tag{5.1.3}$$

As our first step, we set

$$r_0 = n_0 = \left\lceil \sqrt{d} \right\rceil, \tag{5.1.4}$$

and we note that $n_0 \geq 2$. Also note that r_0 satisfies the congruence

$$r_{-1} + r_0 \equiv 0 \pmod{k_{-1}}$$

automatically since $k_{-1} = 1$, and r_0 is the smallest positive integer satisfying this congruence such that $\left\lceil \sqrt{d} \right\rceil \leq r_0$. We set $C = \left\lceil \sqrt{d} \right\rceil$ from this point forward. Our goal is to give a well-defined algorithm by which six infinite sequences of integers are generated; these sequences are as follows:

$$k_{-2}, k_{-1}, k_0, k_1, k_2, \ldots$$

$$r_{-1}, r_0, r_1, r_2, \ldots$$

$$n_0, n_1, n_2, \ldots$$

$$p_{-2}, p_{-1}, p_0, p_1, p_2, \ldots$$

$$q_{-2}, q_{-1}, q_0, q_1, q_2, \ldots$$

$$s_{-1}, s_0, s_1, s_2, \ldots.$$

We set $s_{-1} = -d$, and note that $s_{-1} = r_{-1}^2 - d$ holds by construction. In general, we set $s_j = r_j^2 - d$ for $j = -1, 0, 1, 2, \ldots$, and we note that $s_{-1}, s_0, s_1, s_2, \ldots$ are *all* nonzero since $d > 0$ is not a perfect square. In this section, we focus on the algorithm that allows us to generate these six infinite sequences in a uniquely determined way. In the next section, we show in detail how these sequences of integers are related to

the minus continued fraction expansion of $\sqrt{d}$, as well as how they provide nontrivial integer pair solutions to the Fermat-Pell 1-Equation $x^2 - dy^2 = 1$.

For future reference, we note that

$$\frac{r_{-1} + r_0}{k_{-1}} = r_0 = n_0, \tag{5.1.5}$$

and also that

$$n_0 = \left\lceil \frac{r_{-1} + \sqrt{d}}{k_{-1}} \right\rceil = \left\lceil \sqrt{d} \right\rceil. \tag{5.1.6}$$

Both (5.1.5) and (5.1.6) lay down a pattern that is consistent throughout the algorithm. We may now set

$$p_0 = n_0 p_{-1} - p_{-2} = n_0 \cdot 1 - 0 = n_0 \tag{5.1.7}$$

and

$$q_0 = n_0 q_{-1} - q_{-2} = n_0 \cdot 0 - (-1) = 1, \tag{5.1.8}$$

and these relations again form part of a pattern that is consistent throughout. By construction, we have $s_{-1} = k_{-2} \cdot k_{-1}$. The next step is to compute k_0, which is determined by the following relationship:

$$r_0^2 - d = s_0 = k_{-1} \cdot k_0 = k_0, \tag{5.1.9}$$

with the last equality holding since $k_{-1} = 1$ by definition. Since $r_0^2 > d$ by (5.1.4), we see that both s_0 and k_0 are positive integers. We also note that

$$k_0 = r_0^2 - d = p_0^2 - d \cdot q_0^2 \tag{5.1.10}$$

by (5.1.5), (5.1.7), and (5.1.8) above. Equations (5.1.3a), (5.1.3), and (5.1.10) are the

first three in an infinite sequence of equations:

$$p_j^2 - dq_j^2 = k_j \quad \text{holding for } j = -2, -1, 0, 1, 2, \ldots, \tag{5.1.11}$$

that we wish to establish.

Our next step is to compute r_1. Before we move on to this next step, we first make some general comments about our algorithm. Recall that $r_0 = C \in \mathbb{Z}^{\geq 2}$, and in general we will always choose each r_j for $j = 0, 1, 2, \ldots$ in our algorithm so that $r_j \geq C$ for each $j = 0, 1, 2, \ldots$, which implies that $r_j \in \mathbb{Z}^{\geq 2}$ for all $j \geq 0$. This implies in turn that $r_j^2 > d$ for all $j \geq 0$, and so $r_j^2 - d = s_j > 0$ for all $j \geq 0$, so that $s_0, s_1, s_2, \ldots$ are all positive integers. In (5.1.9), we computed k_0 by using the formula $s_0 = k_{-1} \cdot k_0$, and in general we will prove that $k_{j-1} \mid s_j$ for $j = 0, 1, 2, \ldots$, and based on this, we will define

$$k_j = \frac{s_j}{k_{j-1}} \quad \text{for } j = 0, 1, 2, \ldots. \tag{5.1.12}$$

Since each s_j is a positive integer and $k_{-1} = 1$, this recursive process will guarantee that $k_0, k_1, k_2, \ldots$ are all positive integers as well. By definition, we have

$$p_{-2}q_{-1} - p_{-1}q_{-2} = 0 \cdot 0 - 1(-1) = 1,$$

and from (5.1.7) and (5.1.8), we have

$$p_{-1}q_0 - p_0q_{-1} = 1 \cdot 1 - n_0 \cdot 0 = 1.$$

We will show in general that

$$p_{j-1}q_j - p_jq_{j-1} = 1 \quad \text{for } j = -1, 0, 1, 2, \ldots. \tag{5.1.13}$$

This implies that $\gcd(p_j, q_j) = 1$ for $j = -1, 0, 1, 2, \ldots$ by Corollary 1.1.6. Equation

(5.1.6) gives us the formula for n_0, and once we have $n_1, n_2, n_3, \ldots$ in hand, we will find that

$$p_j = n_j p_{j-1} - p_{j-2} \quad \text{for } j = 0, 1, 2, \ldots \tag{5.1.14}$$

and

$$q_j = n_j q_{j-1} - q_{j-2} \quad \text{for } j = 0, 1, 2, \ldots. \tag{5.1.15}$$

We have already seen the use of these recursive formulas in (5.1.7) and (5.1.8) for $j = 0$. Note that (5.1.14) and (5.1.15) allow us to compute p_j and q_j in terms of the previous *two* values for p and q, respectively. However, it is of interest to note that p_j and q_j may be computed in terms of p_{j-1} and q_{j-1} alone. The formulas that apply in this context are

$$p_j = \frac{p_{j-1} r_j + d q_{j-1}}{k_{j-1}} \quad \text{for } j = -1, 0, 1, 2, \ldots \tag{5.1.16}$$

and

$$q_j = \frac{q_{j-1} r_j + p_{j-1}}{k_{j-1}} \quad \text{for } j = -1, 0, 1, 2, \ldots. \tag{5.1.17}$$

We first check that these two formulas are consistent with the initialization data set up in (5.1.2). When $j = -1$, the right side of (5.1.16) reads as follows:

$$\frac{p_{-2} r_{-1} + d q_{-2}}{k_{-2}} = \frac{0 \cdot 0 + d \cdot (-1)}{-d} = 1,$$

in agreement with the value $p_{-1} = 1$. When $j = -1$, the right side of (5.1.17) reads as follows:

$$\frac{q_{-2} r_{-1} + p_{-2}}{k_{-2}} = \frac{(-1) \cdot 0 + 0}{-d} = 0,$$

in agreement with the value $q_{-1} = 0$. We now check (5.1.16) gives the same answer as in (5.1.7) when $j = 0$. When $j = 0$, by use of (5.1.4), the right side of (5.1.16) reads

as:

$$\frac{p_{-1}r_0 + dq_{-1}}{k_{-1}} = \frac{1 \cdot n_0 + d \cdot 0}{1} = n_0,$$

which agrees with the value $p_0 = n_0$ in (5.1.7). We also check (5.1.17) gives the same answer as in (5.1.8) when $j = 0$. When $j = 0$, by use of (5.1.4), the right side of (5.1.17) reads as:

$$\frac{q_{-1}r_0 + p_{-1}}{k_{-1}} = \frac{0 \cdot n_0 + 1}{1} = 1,$$

which agrees with the value $q_0 = 1$ in (5.1.8). In the implementation of our algorithm, we use (5.1.16) and (5.1.17) instead of (5.1.14) and (5.1.15) to generate the integers $p_0, p_1, p_2, \ldots$ and $q_0, q_1, q_2, \ldots$.

There are two (equivalent) ways to compute r_j. The first method requires only r_{j-1} and k_{j-1}, whereas the second method requires p_{j-1}, q_{j-1}, and k_{j-1}. Even though the second method appears to be more complicated, it is the method used in the implementation of our algorithm. We now demonstrate the first method by showing how to use it to compute r_1. The integer r_1 is uniquely determined as the smallest positive integer satisfying the congruence $r_0 + r_1 \equiv 0 \pmod{k_0}$ such that $C \leq r_1$. This congruence is satisfied if there exists an integer n such that $r_0 + r_1 = nk_0$, or $r_1 = -r_0 + nk_0$. Indeed, our task is to find the smallest positive integer n such that $\sqrt{d} < C \leq -r_0 + nk_0 = r_1$. Our claim is that this uniquely defined integer n is given by the formula

$$n = \left\lceil \frac{r_0 + \sqrt{d}}{k_0} \right\rceil. \tag{5.1.18}$$

To prove this claim, we recall that the integer n determined by (5.1.18) is the unique integer n such that

$$n - 1 < \frac{r_0 + \sqrt{d}}{k_0} < n, \tag{5.1.19}$$

where, since the quantity $(r_0 + \sqrt{d})/k_0$ is an irrational number, both inequalities are

strict. If we multiply the inequalities in (5.1.19) through by the positive integer k_0, we have $(n-1)k_0 < r_0 + \sqrt{d} < nk_0$, or $-r_0 + (n-1)k_0 < \sqrt{d} < -r_0 + nk_0$. These last inequalities show indeed that the integer n determined by (5.1.18) is the least positive integer such that the expression $-r_0 + nk_0$ is greater than $\sqrt{d}$. If we set $n_1 = n$ here, then (5.1.18) reads as

$$n_1 = \left\lceil \frac{r_0 + \sqrt{d}}{k_0} \right\rceil, \tag{5.1.20}$$

and we immediately see the similarity of this to (5.1.6). We also see from above that

$$n_1 = \frac{r_0 + r_1}{k_0}, \tag{5.1.21}$$

which is the next step in the pattern that begins with (5.1.5).

We now demonstrate the second method mentioned above by showing how to use it to compute r_1. The integer r_1 is the uniquely determined smallest positive integer satisfying the congruence $q_0 r_1 + p_0 \equiv 0 \pmod{k_0}$ such that $C \leq r_1$. Since $q_0 = 1$ and $p_0 = r_0$, we see that this congruence simply reads as $r_0 + r_1 \equiv 0 \pmod{k_0}$, so that the integer r_1 obtained here is exactly the same one obtained through the use of the first method discussed just above.

Using the first method, we can actually compute n_1 right away from the already known values of r_0 and k_0 by use of (5.1.20), and then we could set $r_1 = -r_0 + k_0 n_1$. Using the second method, however, we first compute r_1, and then we use (5.1.21) to obtain $n_1 \in \mathbb{Z}^+$.

Using the second method, the positive integer r_1 has been chosen in such a way that the congruence $q_0 r_1 + p_0 \equiv 0 \pmod{k_0}$ is satisfied. This means that the number $(q_0 r_1 + p_0)/k_0$ is actually an integer, and it is a positive integer since q_0, r_1, p_0, and k_0

are all positive integers. In light of this, we set

$$q_1 = \frac{q_0 r_1 + p_0}{k_0}, \tag{5.1.22}$$

and note that $q_1 \in \mathbb{Z}^+$. We claim that the positive integer $p_0 r_1 + d q_0$ is also divisible by k_0. To see this, we note that

$$q_0 \left(p_0 r_1 + d q_0\right) = q_0 p_0 r_1 + d q_0^2 = q_0 p_0 r_1 + p_0^2 - k_0$$

by (5.1.10), or

$$q_0 \left(p_0 r_1 + d q_0\right) = p_0 \left(q_0 r_1 + p_0\right) - k_0. \tag{5.1.23}$$

Since we saw above that $k_0 \mid (q_0 r_1 + p_0)$, we see from (5.1.23) that $k_0 \mid q_0 \left(p_0 r_1 + d q_0\right)$. Since $\gcd\left(k_0, q_0\right) = 1$, which is easy to see since $q_0 = 1$, we may conclude from Theorem 1.1.8 that $k_0 \mid (p_0 r_1 + d q_0)$, proving the claim above. Based on this, we set

$$p_1 = \frac{p_0 r_1 + d q_0}{k_0}, \tag{5.1.24}$$

and note that $p_1 \in \mathbb{Z}^+$. We see that (5.1.24) is just (5.1.16) with $j = 1$, and that (5.1.22) is just (5.1.17) with $j = 1$. We now wish to verify (5.1.13) for $j = 1$, noting that we have already verified (5.1.13) with $j = -1$ and $j = 0$. We have

$$p_0 q_1 - q_0 p_1 = \frac{p_0 \left(q_0 r_1 + p_0\right)}{k_0} - \frac{q_0 \left(p_0 r_1 + d q_0\right)}{k_0} = \frac{p_0^2 - d q_0^2}{k_0} = \frac{k_0}{k_0},$$

with the first equality holding by (5.1.22) and (5.1.24), and the last equality holding by (5.1.10). Thus, we have

$$p_0 q_1 - q_0 p_1 = 1, \tag{5.1.25}$$

which confirms (5.1.13) for $j = 1$. This implies in turn that $\gcd\left(p_1, q_1\right) = 1$ by Corollary 1.1.6.

By definition, we have $s_1 = r_1^2 - d$, and $s_1 \in \mathbb{Z}^+$ since $C \leq r_1$. We now wish to confirm that s_1 is divisible by k_0. For this, we require the following algebraic identity:

$$\left(p^2 - dq^2\right)\left(r^2 - d \cdot 1^2\right) = (pr + dq \cdot 1)^2 - d\,(p \cdot 1 + qr)^2\,, \tag{5.1.26}$$

which is easily verified by expanding out both sides and comparing. Any identity of this shape is known as a "composition formula." Using (5.1.10) and the definition of s_1 in conjunction with (5.1.26) gives:

$$k_0 \cdot s_1 = \left(p_0^2 - dq_0^2\right)\left(r_1^2 - d \cdot 1^2\right) = (p_0 r_1 + dq_0 \cdot 1)^2 - d\,(p_0 \cdot 1 + q_0 \cdot r_1)^2\,. \tag{5.1.27}$$

Recall that both k_0 and s_1 are positive integers. We also have $k_0 \mid (p_0 r_1 + dq_0)$, as well as $k_0 \mid (q_0 r_1 + p_0)$. Thus, by (5.1.27) there exists a positive integer $u \in \mathbb{Z}^+$ such that

$$k_0 s_1 = k_0^2 u,$$

which implies that $k_0 u = s_1$, and so $k_0 \mid s_1$, confirming that s_1 is divisible by k_0. We set

$$k_1 = \frac{s_1}{k_0}, \tag{5.1.28}$$

and note that $k_1 \in \mathbb{Z}^+$. We note that (5.1.28) is consistent with (5.1.12) when $j = 1$. We have $s_1 = k_0 k_1$ from (5.1.28), and plugging into (5.1.27) gives

$$k_0^2 k_1 = (p_0 r_1 + dq_0)^2 - d\,(q_0 r_1 + p_0)^2\,,$$

or

$$k_1 = \left[\frac{p_0 r_1 + dq_0}{k_0}\right]^2 - d\left[\frac{q_0 r_1 + p_0}{k_0}\right]^2, \tag{5.1.29}$$

which gives

$$p_1^2 - dq_1^2 = k_1 \tag{5.1.30}$$

by (5.1.22) and (5.1.24). This last equation is consistent with (5.1.11) when $j = 1$.

Another general pattern that we wish to establish is the following:

$$\gcd(k_j, q_j) = 1 \quad \text{for } j = -2, -1, 0, 1, 2, \ldots. \tag{5.1.31}$$

This clearly holds for $j = -2$ since $k_{-2} = -d$ and $q_{-2} = -1$ and for $j = -1$ as well since $k_{-1} = 1$ and $q_{-1} = 0$. It also holds for $j = 0$ since $q_0 = 1$. We now verify it for $j = 1$. We already know from (5.1.25) that

$$\gcd(p_1, q_1) = 1, \tag{5.1.32}$$

and from (5.1.30) we have $p_1^2 = k_1 + dq_1^2$. If we had $\gcd(k_1, q_1) > 1$, there would be a prime number $t > 1$ such that $t \mid k_1$ and $t \mid q_1$. But then, by (5.1.30), we would have $t \mid p_1$ as well, which contradicts (5.1.32), so we must have

$$\gcd(k_1, q_1) = 1. \tag{5.1.33}$$

The general case of this for $j \in \mathbb{Z}^{\geq 2}$ will be proven later.

We also wish to establish the following:

$$k_j \mid (q_j r_j - p_j) \quad \text{for } j = -1, 0, 1, 2, \ldots. \tag{5.1.34}$$

We already know that k_{-1}, k_0, and k_1 are all positive integers, and as a complement to (5.1.34), we will later show that

$$k_j \in \mathbb{Z}^+ \quad \text{for } j = -1, 0, 1, 2, \ldots. \tag{5.1.35}$$

We note that (5.1.34) clearly holds when $j = -1$ since $k_{-1} = 1$, and when $j = 0$ this is equally clear since $q_0 r_0 - p_0 = 1 \cdot r_0 - r_0 = 0$ by (5.1.5), (5.1.7), and (5.1.8). For

$j = 1$, we have, by (5.1.22), (5.1.24), and (5.1.28),

$$\begin{aligned} q_1 r_1 - p_1 &= \frac{q_0 r_1^2 + p_0 r_1}{k_0} - \frac{p_0 r_1 + d q_0}{k_0} = \frac{q_0 \left(r_1^2 - d\right)}{k_0} \\ &= \frac{q_0 s_1}{k_0} = q_0 k_1 \end{aligned}$$

and so $k_1 \mid (q_1 r_1 - p_1)$.

Yet another general pattern we wish to establish is the following:

$$q_j^{-1} \cdot p_j \equiv r_j \pmod{k_j} \quad \text{for } j = 0, 1, 2, \ldots. \tag{5.1.36}$$

Once we know that $k_j \in \mathbb{Z}^+$ for $j = 0, 1, 2, \ldots$ (see (5.1.35)), then working modulo k_j for any fixed $j \in \mathbb{Z}^{\geq 0}$ makes sense. We will also later show that

$$q_j \in \mathbb{Z}^+ \quad \text{for } j = 0, 1, 2, \ldots \tag{5.1.37}$$

and

$$p_j \in \mathbb{Z}^+ \quad \text{for } j = -1, 0, 1, 2, \ldots. \tag{5.1.38}$$

Once we know that $\gcd(k_j, q_j) = 1$ for $j = 0, 1, 2, \ldots$ (see (5.1.31)), then with (5.1.37) we see that q_j is invertible modulo k_j, and (5.1.36) makes sense. We easily verify (5.1.36) when $j = 0$ since $q_0 = 1$ and $p_0 = n_0 = r_0$. For $j = 1$, we recall that $k_1 \mid (q_1 r_1 - p_1)$, and so

$$p_1 \equiv q_1 r_1 \pmod{k_1}. \tag{5.1.39}$$

Since $q_1 \in \mathbb{Z}^+$ by (5.1.22) and $\gcd(k_1, q_1) = 1$ by (5.1.33), the integer q_1 is invertible modulo k_1, and if we multiply both sides of (5.1.39) by the multiplicative inverse of q_1 modulo k_1, we obtain $q_1^{-1} \cdot p_1 \equiv r_1 \pmod{k_1}$, which confirms (5.1.36) for $j = 1$.

Going back to (5.1.34), we will later establish the following identity which actually implies (5.1.34) in general, so it really just suffices to prove the following

identity:

$$q_j r_j - p_j = q_{j-1} k_j \quad \text{for } j = -1, 0, 1, 2, \ldots. \tag{5.1.40}$$

For $j = -1$, we have $q_{-1} r_{-1} - p_{-1} = 0 \cdot 0 - 1 = -1 = (-1) \cdot 1 = q_{-2} \cdot k_{-1}$. For $j = 0$, by (5.1.5), (5.1.7), and (5.1.8), we have $q_0 r_0 - p_0 = 1 \cdot r_0 - r_0 = 0 = 0 \cdot k_0 = q_{-1} \cdot k_0$, which may be rewritten for later use as

$$q_0 r_0 - q_{-1} k_0 = p_0. \tag{5.1.40a}$$

We already proved just above (5.1.36) that (5.1.40) holds when $j = 1$.

We will also later establish the following identity, which is quite similar to (5.1.40):

$$p_j r_j - d q_j = p_{j-1} k_j \quad \text{for } j = -1, 0, 1, 2, \ldots. \tag{5.1.41}$$

For $j = -1$, we have $p_{-1} r_{-1} - d q_{-1} = 1 \cdot 0 - d \cdot 0 = 0 = 0 \cdot 1 = p_{-2} k_{-1}$. For $j = 0$, by (5.1.5), (5.1.7), (5.1.8), and (5.1.10), we have $p_0 r_0 - d q_0 = r_0^2 - d = 1 \cdot (r_0^2 - d) = p_{-1} k_0$, which may be rewritten for later use as

$$p_0 r_0 - p_{-1} k_0 = d q_0. \tag{5.1.42}$$

For $j = 1$, by (5.1.22), (5.1.24), and (5.1.28), we have

$$\begin{aligned} p_1 r_1 - d q_1 &= \frac{p_0 r_1^2 + d q_0 r_1}{k_0} - \frac{d q_0 r_1 + d p_0}{k_0} = \frac{p_0 \, (r_1^2 - d)}{k_0} \\ &= \frac{p_0 s_1}{k_0} = p_0 k_1. \end{aligned}$$

There is yet one more divisibility statement that we would like to establish, namely:

$$k_{j-1} \mid (r_{j-1} + r_j) \quad \text{for } j = 0, 1, 2, \ldots, \tag{5.1.43}$$

and based on this, we define

$$n_j = \frac{r_{j-1} + r_j}{k_{j-1}} \quad \text{for } j = 0, 1, 2, \ldots. \tag{5.1.44}$$

Note that (5.1.43) is automatic when $j = 0$ since $k_{-1} = 1$. We previously saw that our definitions led to $n_0 = (r_{-1} + r_0)/k_{-1}$ in (5.1.5), which is identical to (5.1.44) when $j = 0$. Using either the first or the second method mentioned earlier, we chose r_1 to be the smallest positive integer satisfying the congruence $q_0 r_1 + p_0 = r_1 + r_0 \equiv 0 \pmod{k_0}$ such that $C \leq r_1$. As an immediate consequence, we have $k_0 \mid (r_0 + r_1)$, which is (5.1.43) with $j = 1$. In this case, we set (see (5.1.21)) $n_1 = (r_0 + r_1)/k_0$, which is identical to (5.1.44) when $j = 1$. We will later establish (5.1.43) in general, and based on this and (5.1.35), we use (5.1.44) as our *definition* of n_j for $j = 0, 1, 2, \ldots$.

We will later prove in general that we have

$$n_j = \left\lceil \frac{r_{j-1} + \sqrt{d}}{k_{j-1}} \right\rceil \quad \text{for } j = 0, 1, 2, \ldots. \tag{5.1.45}$$

By (5.1.6) and (5.1.20), we already know that (5.1.45) holds when $j = 0$ and $j = 1$, respectively.

We now finally wish to establish (5.1.14) and (5.1.15) for $j = 1$. To confirm (5.1.14) for $j = 1$, we note that, by (5.1.21), (5.1.42), and (5.1.24), in that order, we have

$$\begin{aligned} n_1 p_0 - p_{-1} &= \frac{p_0 (r_0 + r_1)}{k_0} - \frac{p_{-1} k_0}{k_0} = \frac{p_0 r_1 + (p_0 r_0 - p_{-1} k_0)}{k_0} \\ &= \frac{p_0 r_1 + d q_0}{k_0} = p_1. \end{aligned}$$

To confirm (5.1.15) for $j = 1$, we note that, by (5.1.21), (5.1.40a), and (5.1.22), in

that order, we have

$$n_1 q_0 - q_{-1} = \frac{q_0 (r_0 + r_1)}{k_0} - \frac{q_{-1} k_0}{k_0} = \frac{q_0 r_1 + (q_0 r_0 - q_{-1} k_0)}{k_0}$$
$$= \frac{q_0 r_1 + p_0}{k_0} = q_1.$$

Let us now take stock of what we have already proven, and what we still need to establish. We summarize our situation with the following sequence of statements:

(A) $k_j \in \mathbb{Z}^+$ for $j = -1, 0, 1, 2, \ldots$. This is already known for $j = -1, 0, 1$.

(B) $\left\lceil \sqrt{d} \right\rceil = C \leq r_j$ for $j = 0, 1, 2, \ldots$. These inequalities hold by the very construction of our algorithm. This is already established for $j = 0, 1$.

(C) $s_j := r_j^2 - d \in \mathbb{Z}^+$ for $j = 0, 1, 2, \ldots$. This positivity statement is an immediate corollary of (B).

(D) $q_j \in \mathbb{Z}^+$ for $j = 0, 1, 2, \ldots$. This is already known for $j = 0, 1$.

(E) $p_j \in \mathbb{Z}^+$ for $j = -1, 0, 1, 2, \ldots$. This is already known for $j = -1, 0, 1$.

(F) $n_j \in \mathbb{Z}^+$ for $j = 0, 1, 2, \ldots$. This is already known for $j = 0, 1$.

(G) $p_j^2 - d q_j^2 = k_j$ for $j = -2, -1, 0, 1, 2, \ldots$. This is already known for $j = -2, -1, 0, 1$.

(H) $k_{j-1} \mid s_j$ for $j = -1, 0, 1, 2, \ldots$. This is already known for $j = -1, 0, 1$.

(I) Based upon (H), we define $k_j := s_j / k_{j-1}$ for $j = -1, 0, 1, 2, \ldots$. We then automatically have $s_j = k_{j-1} \cdot k_j$ for $j = -1, 0, 1, 2, \ldots$.

(J) $p_{j-1} q_j - p_j q_{j-1} = 1$ for $j = -1, 0, 1, 2, \ldots$. This is already known for $j = -1, 0, 1$.

(K) $\gcd(p_j, q_j) = 1$ for $j = -1, 0, 1, 2, \ldots$. This is an immediate corollary of (J) by Corollary 1.1.6.

(L) $k_{j-1} \mid (r_{j-1} + r_j)$ for $j = 0, 1, 2, \ldots$. This is already known for $j = 0, 1$.

(M) Based upon (L), we define $n_j := (r_{j-1} + r_j)/k_{j-1}$ for $j = 0, 1, 2, \ldots$.

(N) $n_j = \left\lceil \frac{r_{j-1}+\sqrt{d}}{k_{j-1}} \right\rceil$ for $j = 0, 1, 2, \ldots$. This is already known for $j = 0, 1$.

(O) $p_j = n_j p_{j-1} - p_{j-2}$ for $j = 0, 1, 2, \ldots$. This is already known for $j = 0, 1$.

(P) $q_j = n_j q_{j-1} - q_{j-2}$ for $j = 0, 1, 2, \ldots$. This is already known for $j = 0, 1$.

(Q) $k_{j-1} \mid (p_{j-1} r_j + d q_{j-1})$ for $j = -1, 0, 1, 2, \ldots$. This is already known for $j = -1, 0, 1$.

(R) Based upon (Q), we define $p_j = (p_{j-1} r_j + d q_{j-1})/k_{j-1}$ for $j = -1, 0, 1, 2, \ldots$.

(S) $k_{j-1} \mid (q_{j-1} r_j + p_{j-1})$ for $j = -1, 0, 1, 2, \ldots$. This is already known for $j = -1, 0, 1$.

(T) Based upon (S), we define $q_j = (q_{j-1} r_j + p_{j-1})/k_{j-1}$ for $j = -1, 0, 1, 2, \ldots$.

(U) $\gcd(k_j, q_j) = 1$ for $j = -2, -1, 0, 1, 2, \ldots$. This is already known for $j = -2, -1, 0, 1$.

(V) The crucial step in the algorithm is the following: Given a fixed $j \in \mathbb{Z}^{\geq 0}$ so that k_j, q_j, and p_j are all positive integers (by (A), (D), and (E)), and $\gcd(k_j, q_j) = 1$ (by (U)), we choose r_{j+1} to be the smallest positive integer satisfying the congruence

$$r_{j+1} + q_j^{-1} \cdot p_j \equiv 0 \pmod{k_j} \tag{5.1.46}$$

such that $C \leq r_{j+1}$.

(W) $q_j r_j - p_j = q_{j-1} k_j$ for $j = -1, 0, 1, 2, \ldots$. This is already known for $j = -1, 0, 1$.

(X) $p_j r_j - dq_j = p_{j-1} k_j$ for $j = -1, 0, 1, 2, \ldots$. This is already known for $j = -1, 0, 1$.

(Y) $q_j^{-1} \cdot p_j \equiv r_j \pmod{k_j}$ for $j = 0, 1, 2, \ldots$. This is already known for $j = 0, 1$.

Based upon what we already have in place, we now wish to obtain r_j, n_j, q_j, p_j, s_j, and k_j for $j = 2$, and to establish all of the statements (A) – (Y) above with respect to this next level in the algorithm. Each new level in the algorithm is obtained from the previous ones in exactly the same way, and so if we go through the details for this next $j = 2$ level, then the same arguments take us forward to $j = 3, 4, 5, \ldots$.

We first choose r_2 to be the unique smallest positive integer satisfying the congruence

$$r_2 + q_1^{-1} \cdot p_1 \equiv 0 \pmod{k_1} \tag{5.1.47}$$

such that $C \leq r_2$. By construction, (B) holds for $j = 2$. We already know by (Y) for $j = 1$ that $q_1^{-1} \cdot p_1 \equiv r_1 \pmod{k_1}$, and thus in (5.1.47) we have chosen r_2 to be the smallest positive integer satisfying the congruence

$$r_1 + r_2 \equiv 0 \pmod{k_1} \tag{5.1.48}$$

such that $C \leq r_2$. The congruence in (5.1.48) implies in turn that $k_1 \mid (r_1 + r_2)$ (thus (L) is satisfied for $j = 2$), and we set

$$n_2 = \frac{r_1 + r_2}{k_1}. \tag{5.1.49}$$

Note that $n_2 \in \mathbb{Z}^+$ since k_1, r_1, and r_2 are all positive integers, and so (F) is satisfied for $j = 2$. By the same argument used in proving the claim in connection with (5.1.18),

we see that

$$n_2 = \left\lceil \frac{r_1 + \sqrt{d}}{k_1} \right\rceil,$$

and so (N) holds for $j = 2$. From (5.1.47), we have

$$q_1 r_2 + p_1 \equiv 0 \pmod{k_1}, \tag{5.1.50}$$

and thus (S) is satisfied for $j = 2$. We set

$$q_2 = \frac{q_1 r_2 + p_1}{k_1} \tag{5.1.51}$$

and note that $q_2 \in \mathbb{Z}^+$ since k_1, q_1, p_1, and r_2 are all positive integers, and thus (D) is satisfied for $j = 2$. To confirm that $k_1 \mid (p_1 r_2 + d q_1)$ (see (Q)), we use (G) for $j = 1$, to see that

$$q_1 (p_1 r_2 + d q_1) = q_1 p_1 r_2 + d q_1^2 = q_1 p_1 r_2 + p_1^2 - k_1,$$

or

$$q_1 (p_1 r_2 + d q_1) = p_1 (q_1 r_2 + p_1) - k_1.$$

By (5.1.50), we see that $k_1 \mid q_1 (p_1 r_2 + d q_1)$, and since $\gcd(k_1, q_1) = 1$ (by (U) with $j = 1$), we conclude by Theorem 1.1.8 that

$$k_1 \mid (p_1 r_2 + d q_1), \tag{5.1.52}$$

and thus (Q) is satisfied for $j = 2$. We now set

$$p_2 = \frac{p_1 r_2 + d q_1}{k_1}, \tag{5.1.53}$$

and note that $p_2 \in \mathbb{Z}^+$ since k_1, q_1, p_1, d, and r_2 are all positive integers. Thus, (E) is

satisfied for $j = 2$. We now confirm (J) for $j = 2$. By (5.1.51) and (5.1.53), we have

$$\begin{aligned} p_1 q_2 - q_1 p_2 &= \frac{p_1 \left(q_1 r_2 + p_1\right)}{k_1} - \frac{q_1 \left(p_1 r_2 + d q_1\right)}{k_1} \\ &= \frac{p_1^2 - d q_1^2}{k_1} = \frac{k_1}{k_1}, \end{aligned}$$

with the last equality holding by (G) for $j = 1$, or

$$p_1 q_2 - p_2 q_1 = 1. \tag{5.1.54}$$

This implies (K) as well for $j = 2$, namely

$$\gcd\left(p_2, q_2\right) = 1. \tag{5.1.55}$$

By definition, $s_2 = r_2^2 - d$, and $s_2 \in \mathbb{Z}^+$ since $C \leq r_2$. Using this definition, (G) for $j = 1$, and the algebraic identity in (5.1.26), we obtain

$$k_1 \cdot s_2 = \left(p_1^2 - d q_1^2\right)\left(r_2^2 - d \cdot 1^2\right) = \left(p_1 r_2 + d q_1 \cdot 1\right)^2 - d\left(p_1 \cdot 1 + q_1 r_2\right)^2. \tag{5.1.56}$$

Recall that k_1 and s_2 are both positive integers, that $k_1 \mid \left(p_1 r_2 + d q_1\right)$ by (5.1.52), and that $k_1 \mid \left(q_1 r_2 + p_1\right)$ by (5.1.50), so that by (5.1.56) there exists a positive integer $u \in \mathbb{Z}^+$ such that $k_1 s_2 = k_1^2 u$. This implies that $s_2 = k_1 u$, and so $k_1 \mid s_2$, which confirms (H) for $j = 2$. We set $k_2 = s_2 / k_1$, which implies that

$$s_2 = k_1 k_2. \tag{5.1.57}$$

We note that $k_2 = u \in \mathbb{Z}^+$, and so (A) is satisfied for $j = 2$. Plugging (5.1.57) into the left hand side of (5.1.56) gives $k_1^2 k_2 = \left(p_1 r_2 + d q_1\right)^2 - d\left(q_1 r_2 + p_1\right)^2$, or

$$k_2 = \left[\frac{p_1 r_2 + d q_1}{k_1}\right]^2 - d\left[\frac{q_1 r_2 + p_1}{k_1}\right]^2, \tag{5.1.58}$$

which gives

$$p_2^2 - dq_2^2 = k_2 \tag{5.1.59}$$

by (5.1.51) and (5.1.53). This confirms (G) for $j = 2$.

We now wish to verify (U) for $j = 2$. By (5.1.55), we have $\gcd(p_2, q_2) = 1$, and by (5.1.59), we have $p_2^2 = k_2 + dq_2^2$. If we had $\gcd(k_2, q_2) > 1$ (we know already that k_2 and q_2 are positive integers), there would be a prime number $t > 1$ such that $t \mid k_2$ and $t \mid q_2$. But then, by (5.1.59), we would have $t \mid p_2$ as well, which contradicts (5.1.55), so we must have

$$\gcd(k_2, q_2) = 1. \tag{5.1.60}$$

This confirms (U) for $j = 2$. We next wish to verify (W) for $j = 2$. By (5.1.51), (5.1.53), and (5.1.57), we have

$$q_2 r_2 - p_2 = \frac{q_1 r_2^2 + p_1 r_2}{k_1} - \frac{p_1 r_2 + dq_1}{k_1} = \frac{q_1 (r_2^2 - d)}{k_1} = \frac{q_1 s_2}{k_1} = q_1 k_2. \tag{5.1.61}$$

This confirms (W) for $j = 2$. We also have

$$p_2 r_2 - dq_2 = \frac{p_1 r_2^2 + dq_1 r_2}{k_1} - \frac{dq_1 r_2 + dp_1}{k_1} = \frac{p_1 (r_2^2 - d)}{k_1} = \frac{p_1 s_2}{k_1} = p_1 k_2, \tag{5.1.62}$$

which confirms (X) for $j = 2$. To confirm (Y) for $j = 2$, we note from (5.1.61) that $k_2 \mid (q_2 r_2 - p_2)$, and so

$$p_2 \equiv q_2 r_2 \pmod{k_2}. \tag{5.1.63}$$

Since $q_2 \in \mathbb{Z}^+$, and $\gcd(k_2, q_2) = 1$ by (5.1.60), the integer q_2 is invertible modulo k_2, and if we multiply both sides of (5.1.63) by the multiplicative inverse of q_2 modulo k_2, then $q_2^{-1} \cdot p_2 \equiv r_2 \pmod{k_2}$, which confirms (Y) for $j = 2$. To confirm (O) for $j = 2$,

we note that by (5.1.49), and by (X) for $j = 1$, that

$$n_2 p_1 - p_0 = \frac{p_1 (r_1 + r_2)}{k_1} - \frac{p_0 k_1}{k_1} = \frac{p_1 r_2 + (p_1 r_1 - p_0 k_1)}{k_1} = \frac{p_1 r_2 + d q_1}{k_1},$$

and so, by (5.1.53), we have $p_2 = n_2 p_1 - p_0$, which confirms (O) for $j = 2$. To confirm (P) for $j = 2$, we note that by (5.1.49), and by (W) for $j = 1$, that

$$n_2 q_1 - q_0 = \frac{q_1 (r_1 + r_2)}{k_1} - \frac{q_0 k_1}{k_1} = \frac{q_1 r_2 + (q_1 r_1 - q_0 k_1)}{k_1} = \frac{q_1 r_2 + p_1}{k_1},$$

and so, by (5.1.51), we have $q_2 = n_2 q_1 - q_0$, which confirms (P) for $j = 2$.

5.2. Finding Solutions to the Fermat-Pell 1-Equation

In the following, let $d \in \mathbb{Z}^+$ denote a fixed positive integer that is not a perfect square. In this section, we connect the algorithm described in Section 5.1 to Zagier's reduction theory for indefinite binary quadratic forms, to the Fermat-Pell 1-Equation

$$x^2 - d y^2 = 1, \tag{5.2.1}$$

and to the minus continued fraction expansion of $\sqrt{d}$ as well. We consistently use the notations and numbering laid down in Section 5.1, and a capital letter reference such as (A) refers to statement (A) in Section 5.1.

First, set $f_0 = [k_{-1}, 2r_{-1}, k_{-2}] = [1, 0, -d\,]$, noting that f_0 is a form of discriminant $D = 4d \in \mathbb{Z}^+$. We claim that in this case, D is a discriminant satisfying Assumption 2.2.1. Clearly, $D \equiv 0 \pmod 4$, and we just need to show that D is not a perfect square. Since $d > 1$ is not a perfect square, we know by Theorem 1.1.13 that there exists a prime p such that the exponent $e_p(d)$ appearing in the prime factorization of d is odd. If p is an odd prime, then $e_p(D) = e_p(d)$ is still odd, and D is not a perfect square, again by Theorem 1.1.13. If $p = 2$, then $e_2(D) = e_2(d) + 2$ is still odd, and again D is not a perfect square by Theorem 1.1.13, proving our claim. Furthermore,

note that f_0 is the principal form of discriminant D by Definition 2.2.2. Our goal is to apply the Zagier reduction algorithm to f_0 and make a careful comparison to the algorithm presented in Section 5.1. After the proper correspondences are made, an exact agreement is seen to hold (see Theorem 5.2.1 below).

In terms of the notation introduced in Section 5.1, we define the following infinite sequence of forms:

$$f_j = [k_{j-1}, 2r_{j-1}, k_{j-2}] \quad \text{for } j = 0, 1, 2, \ldots \tag{5.2.2}$$

(note that when $j = 0$ we obtain the same f_0 as in the paragraph above). Recall that $s_{j-1} = k_{j-1}k_{j-2}$ for $j = 0, 1, 2, \ldots$ by (I), and thus

$$4r_{j-1}^2 - 4k_{j-1}k_{j-2} = 4r_{j-1}^2 - 4s_{j-1} \tag{5.2.3}$$

for $j = 0, 1, 2, \ldots$. Remembering that $s_{j-1} = r_{j-1}^2 - d$, or that $r_{j-1}^2 - s_{j-1} = d$ for $j = 0, 1, 2, \ldots$, the expression in (5.2.3) is seen to be equal to $D = 4d$ for $j = 0, 1, 2, \ldots$. This shows that each form f_j defined in the sequence in (5.2.2) has discriminant equal to D.

Theorem 5.2.1. *In terms of the sequence $n_0, n_1, n_2, \ldots$ of integers defined in Section 5.1, the diagram*

$$f_0 \xrightarrow{n_0} f_1 \xrightarrow{n_1} f_2 \xrightarrow{n_2} \cdots$$

coincides exactly with what is obtained when we apply the Zagier reduction algorithm to the starting form f_0.

Proof. When we apply the reduction algorithm to the form $f_0 = [1, 0, -d\,]$, we first compute

$$n_0 = \left\lceil \frac{b + \sqrt{D}}{2a} \right\rceil = \left\lceil \frac{0 + \sqrt{4d}}{2 \cdot 1} \right\rceil = \left\lceil \sqrt{d} \right\rceil,$$

and we see that this matches the value of n_0 in (5.1.4). Given $a = 1$, $b = 0$, and $c = -d$, we see from (3.1.4) that $a' = n_0^2 - d = k_0$, with the second equality holding by (5.1.5) and (5.1.9). By (3.1.5), we have $b' = 2n_0 = 2r_0$, with (5.1.5) used again in the second equality. By (3.1.6), we have $c' = 1 = k_{-1}$, where (5.1.2) is used in the second equality. Thus, the first step of the reduction algorithm applied to f_0 gives us

$$f_0 = [1, 0, -d\,] \xrightarrow{n_0} [k_0, 2r_0, k_{-1}],$$

and the second form is equal to f_1, as defined in (5.2.2). More generally, let $j \geq 1$ be a given integer and consider the form $f_j = [k_{j-1}, 2r_{j-1}, k_{j-2}] = [a, b, c]$. We have

$$n_j = \left\lceil \frac{b + \sqrt{D}}{2a} \right\rceil = \left\lceil \frac{2r_{j-1} + \sqrt{4d}}{2k_{j-1}} \right\rceil = \left\lceil \frac{r_{j-1} + \sqrt{d}}{k_{j-1}} \right\rceil,$$

in concordance with (N). Referring back to (3.1.4), (3.1.5), and (3.1.6), if we can confirm that each of the second equalities below hold:

$$a' = an_j^2 - bn_j + c = k_j, \tag{5.2.4}$$

$$b' = 2an_j - b = 2r_j, \quad \text{and} \tag{5.2.5}$$

$$c' = a = k_{j-1}, \tag{5.2.6}$$

then the reduction algorithm is shown to take f_j to f_{j+1}, completing the proof of this theorem by induction. Clearly, (5.2.6) holds. By (M), we have $k_{j-1}n_j = r_{j-1} + r_j$, or $r_j = k_{j-1}n_j - r_{j-1}$, so that $2r_j = 2k_{j-1}n_j - 2r_{j-1}$, which confirms (5.2.5). To prove that (5.2.4) holds, first note that by (C) we have $r_j^2 - s_j = d = r_{j-1}^2 - s_{j-1}$, and so

$$s_j = r_j^2 - r_{j-1}^2 + s_{j-1}. \tag{5.2.7}$$

By (I), we have $s_j = k_{j-1}k_j$ and $s_{j-1} = k_{j-2}k_{j-1}$, so that (5.2.7) may be rewritten as

$$k_j = \frac{(r_j^2 - r_{j-1}^2)}{k_{j-1}} + k_{j-2},$$

or

$$\begin{aligned} k_j &= \frac{r_{j-1}^2 + 2r_{j-1}r_j + r_j^2}{k_{j-1}} + \frac{-2r_{j-1}^2 - 2r_{j-1}r_j}{k_{j-1}} + k_{j-2} \\ &= k_{j-1}\left[\frac{r_{j-1} + r_j}{k_{j-1}}\right]^2 - (2r_{j-1})\left[\frac{r_{j-1} + r_j}{k_{j-1}}\right] + k_{j-2}. \end{aligned}$$

By (M), we conclude that $k_j = k_{j-1}n_j^2 - 2r_{j-1}n_j + k_{j-2}$, which finally confirms (5.2.4) as well. □

Note that f_0, as defined above, is equal to $x^2 - dy^2$, which is exactly the form on the left hand side of (5.2.1). If we set $x = p_{-1} = 1$ and $y = q_{-1} = 0$, we obtain one of the two trivial integer pair solutions to (5.2.1). We now wish to show that there exists an integer $j \geq 0$ such that $x = p_j$ and $y = q_j$ provides yet another integer pair solution to (5.2.1). By (D) and (E), we know that p_j and q_j are both *positive* integers for each index $j \geq 0$, and thus if (p_j, q_j) is an integer pair solution to (5.2.1) for some index $j \geq 0$, then it is necessarily a *nontrivial* solution pair to (5.2.1). Recall from (G) that $p_j^2 - dq_j^2 = k_j$ for all integers $j \geq 0$. The left hand side of this relation is the form $x^2 - dy^2$ with $x = p_j$ and $y = q_j$, and thus if we can prove that at least one member k_i of the infinite sequence of positive integers $k_0, k_1, k_2, \ldots$ is equal to 1, then we have in hand a nontrivial solution pair (p_i, q_i) to the Fermat-Pell 1-Equation. It is exactly the crucial link established in Theorem 5.2.1 that allows us to prove that $k_i = 1$ for some $i \in \mathbb{Z}^{\geq 0}$. To see the connection, we quickly review what happens when we apply the Zagier reduction algorithm to the *principal* form $f_0 = [1, 0, -d]$ of discriminant $D = 4d \in \mathbb{Z}^+$. As we confirmed earlier (see Table 3.1.2 and the discussion immediately

above this Table), the reduction algorithm takes the principal form f_0 to a *reduced* form f_1 in the principal class in exactly one step. With reference to Theorem 5.2.1, this means that by Theorem 3.1.4, $f_1, f_2, f_3, \ldots$ are all reduced forms, making up the principal cycle of discriminant D. From (5.2.2), we have

$$f_1 = [k_0, 2r_0, 1], \tag{5.2.8}$$

and we now know that f_1 is a reduced form of discriminant $D = 4d \in \mathbb{Z}^+$. If $k_0 = 1$, then $p_0^2 - dq_0^2 = r_0^2 - d = 1$ by (5.1.10), and we have right off the bat a nontrivial solution pair $(p_0, q_0) = (r_0, 1)$ to (5.2.1). Note that in this case, d is one less than a perfect square:

$$r_0^2 - 1 = d, \quad \text{if } \ k_0 = 1. \tag{5.2.9}$$

Since $f_2 = [k_1, 2r_1, k_0]$, $f_3 = [k_2, 2r_2, k_1], \ldots$, proving that $k_i = 1$ for some $i \in \mathbb{Z}^{\geq 0}$ is tantamount to proving that one of the forms in the principal cycle of discriminant D has an a–coefficient equal to 1. Therefore, once Theorem 5.2.3 below is established, a straightforward algorithm to compute a nontrivial solution pair to (5.2.1) in a finite number of steps is readily available since the principal cycle of discriminant D contains a *finite* number of forms, even if the number of such forms can be surprisingly large for some discriminants! In other words, the following result is an immediate corollary to Theorem 5.2.3.

Corollary 5.2.2. *If $d \in \mathbb{Z}^+$ is a fixed positive integer that is not a perfect square, then the Fermat-Pell 1-Equation* (5.2.1) *possesses an integer pair solution $(p, q) \in \mathbb{Z}^2$ with $q \in \mathbb{Z}^+$.*

Theorem 5.2.3. *If $d \in \mathbb{Z}^+$ is a fixed positive integer that is not a perfect square, then the principal cycle of discriminant $D = 4d$ contains a form whose a–coefficient is equal*

to one.

Proof. We noted above that $f_1 = [k_0, 2r_0, 1]$ is a reduced form lying in the principle cycle of discriminant D. We claim that the form $f^* = [1, 2r_0, k_0]$ is also a reduced form lying in the principle cycle of discriminant D. Note that our proof is complete once this claim is confirmed. Since $f_1 = [a, b, c]$ is a reduced form of discriminant D, it is immediate by Definition 3.1.1 that $f^* = [c, b, a]$ is a reduced form of discriminant D as well. We just need to show that f^* lies in the principal cycle of discriminant D. By Theorem 3.1.11, it suffices to prove that f^* lies in the same equivalence class as f_1. To show that $f^* \sim f_1$, we apply the reduction algorithm to f^*. We first obtain

$$n^* = \left\lceil \frac{2r_0 + \sqrt{4d}}{2 \cdot 1} \right\rceil = \left\lceil r_0 + \sqrt{d} \right\rceil = 2r_0, \tag{5.2.10}$$

with the last equality holding since $r_0 = \left\lceil \sqrt{d} \right\rceil$. If we represent this first step of the reduction algorithm by

$$f^* \xrightarrow{n^*} [a', b', c'], \tag{5.2.11}$$

then

$$a' = (n^*)^2 - 2r_0 n^* + k_0 \tag{5.2.12}$$

$$b' = 2n^* - 2r_0 \tag{5.2.13}$$

$$c' = 1 \tag{5.2.14}$$

by (3.1.4), (3.1.5), and (3.1.6), respectively. From (5.2.10), we see that the right hand side of (5.2.12) is equal to $(2r_0)^2 - 2r_0(2r_0) + k_0 = k_0$, which matches the a–coefficient of f_1. Similarly, the right hand side of (5.2.13) is equal to $2(2r_0) - 2r_0 = 2r_0$, which matches the b–coefficient of f_1. Finally, the right hand side of (5.2.14) is equal to the

c–coefficient of f_1, and therefore (5.2.11) may be replaced by

$$f^* \xrightarrow{n^*} f_1, \tag{5.2.15}$$

confirming that $f^* \sim f_1$, and completing the proof of Theorem 5.2.3. □

In Example 5.2.4 below, we illustrate the algorithm for finding a nontrivial solution pair to (5.2.1) arising naturally from the statement of Theorem 5.2.3.

Example 5.2.4. Assume that $d = 5$, and so $f_0 = [1, 0, -5]$. According to Theorem 5.2.1, we obtain the sequence of integers $n_0, n_1, n_2, \ldots$ defined in Section 5.1 by applying the Zagier reduction algorithm to the starting form f_0. We proceed until we obtain a form f_m, with $m \in \mathbb{Z}^+$, whose a–coefficient is equal to 1. A straightforward computation yields the following:

$$f_0 = [1, 0, -5] \xrightarrow{3} [4, 6, 1] \xrightarrow{2} [5, 10, 4] \xrightarrow{2} [4, 10, 5] \xrightarrow{2} [1, 6, 4].$$

We see that $m = 4$, and since $f_4 = [k_3, 2r_3, k_2]$ with $k_3 = 1$, our nontrivial solution pair to $x^2 - 5y^2 = 1$ is (p_3, q_3). With the values $n_0 = 3, n_1 = 2, n_2 = 2, n_3 = 2$ in hand, we may compute p_0, p_1, p_2, p_3 recursively using (O), and similarly q_0, q_1, q_2, q_3 using (P). We find that $p_0 = 3, p_1 = 5, p_2 = 7$, and $p_3 = 9$. We also have $q_0 = 1, q_1 = 2, q_2 = 3$, and $q_3 = 4$, and we verify indeed that

$$9^2 - 5(4)^2 = 81 - 80 = 1,$$

giving a nontrivial solution pair $(9, 4)$ to the Fermat-Pell 1-Equation with $d = 5$ as expected. There is also a nice pattern to the k–values, starting with k_{-1}:

$$k_{-1} = 1, \quad k_0 = 4, \quad k_1 = 5, \quad k_2 = 4, \quad k_3 = 1. \tag{5.2.16}$$

With some further refinements, the following algorithm may be justified using our discussion above as a jumping off point.

Algorithm 5.2.5. *If $d \in \mathbb{Z}^+$ is a fixed positive integer that is not a perfect square, and there are $m \in \mathbb{Z}^+$ distinct forms in the principal cycle of discriminant $D = 4d$, then (p_{m-1}, q_{m-1}) is a nontrivial integer pair solution to the Fermat-Pell 1-Equation $x^2 - dy^2 = 1$.*

We require two lemmas to confirm the effectual use of Algorithm 5.2.5.

Lemma 5.2.6. *If $d \in \mathbb{Z}^+$ is a fixed positive integer that is not a perfect square, then the principal cycle of discriminant $D = 4d$ contains exactly one form whose a–coefficient is equal to 1.*

Proof. By Theorem 5.2.3, we know that there exists at least one reduced form $[a, b, c]$ lying in the principal cycle of discriminant D with $a = 1$. To prove that there are no other such forms, recall from (3.1.41) that *any* reduced form of discriminant D is of the shape

$$\left[a, k + 2a, k + a - \frac{D - k^2}{4a}\right], \tag{5.2.17}$$

where the integers a and k simultaneously satisfy the four conditions in (3.1.42). Recall that the b–coefficient

$$b = k + 2a \tag{5.2.18}$$

has the same parity as D by the Remark at the end of Section 2.1, and by (5.2.18) we see that the integer k has the same parity as b and D. Since D is even by assumption, k is even as well. If $a = 1$, then we have

$$0 < \sqrt{D} - k < 2 \tag{5.2.19}$$

by the 4th condition in (3.1.42). We know that $\sqrt{D}$ is an irrational number and there is exactly one *even* integer ℓ lying to the left of $\sqrt{D}$ whose distance to $\sqrt{D}$ is less than 2. By (5.2.19), we have $k = \ell$, and this shows that if $a = 1$, then k is uniquely determined. Going back to (5.2.17), we obtain exactly one reduced form of this shape with $a = 1$ and $k = \ell$, which completes the proof. □

Lemma 5.2.7. *The uniquely defined form of discriminant $D = 4d$ specified in Lemma 5.2.6 is the form $f^* = [1, 2r_0, k_0]$ introduced in the proof of Theorem 5.2.3. The form f^* is the left neighbor of the form f_1 in* (5.2.8).

Proof. We verified in the proof of Theorem 5.2.3 that f^* is a reduced form lying in the principal cycle of discriminant D, and its a–coefficient is equal to 1. We also verified in (5.2.15) that f_1 is the right neighbor of f^*. Since f^* is reduced and $I(f^*) = f_1$, we see from the proof of Proposition 3.1.8 that $J(f_1) = f^*$, proving that f^* is the left neighbor of f_1. □

Justification of Algorithm 5.2.5. If $m = 1$, there is only one reduced form in the principal cycle of discriminant $D = 4d$, and so the reduced form $f_1 = [k_0, 2r_0, 1]$ must be the same as the reduced form $f^* = [1, 2r_0, k_0]$. This implies that $k_0 = 1$ and so $p_0^2 - dq_0^2 = 1$, confirming that (p_{m-1}, q_{m-1}) is a nontrivial integer pair solution to (5.2.1). If $m > 1$, the principal cycle of reduced forms may be visualized by use of the diagram:

$$f_1 \xrightarrow{n_1} \cdots \to \underset{\overset{\longleftarrow}{n_m}}{f_m}. \tag{5.2.20}$$

By Lemma 5.2.7, we see that $f^* = f_m$, and $f_m = [k_{m-1}, 2r_{m-1}, k_{m-2}]$ by (5.2.2). Since $p_{m-1}^2 - dq_{m-1}^2 = k_{m-1} = 1$, our verification is complete in this case as well. If $m > 1$, it is of interest to note that based upon Lemma 5.2.6, no pair of the form (p_j, q_j), for

$j = 0, \ldots, m-2$, will be a solution to (5.2.1) since $p_j^2 - dq_j^2 = k_j > 1$ for these values of j. We do not solve (5.2.1), so to speak, until we arrive at the *last* form f_m in the cycle. □

Using the same argument as was used in Section 2.3 (just above Definition 2.3.6), we can define the unique minimal solution pair (x_1, y_1) to the Fermat-Pell 1-Equation (5.2.1). Both p_{m-1} and q_{m-1} in Algorithm 5.2.5 are positive integers, so it is natural to ask if we sometimes obtain the optimally smallest solution $p_{m-1} = x_1$ and $q_{m-1} = y_1$. Indeed, it is true that Algorithm 5.2.5 not only gives us a nontrivial integer pair solution to (5.2.1), but even better, this algorithm *always* outputs the minimal solution itself. Unfortunately, we do not offer a proof of this statement here since further refined techniques are required for this proof. In the Table 5.2.1 below, we display the nontrivial integer pair solution (p_{m-1}, q_{m-1}) to (5.2.1) arising from Algorithm 5.2.5 for all values of $d \in \mathbb{Z}^+$ less than 16 that are not perfect squares.

Table 5.2.1. Solution pair (p_{m-1}, q_{m-1}) to (5.2.1) using Algorithm 5.2.5

d	(p_{m-1}, q_{m-1})	d	(p_{m-1}, q_{m-1})
2	$(3, 2)$	10	$(19, 6)$
3	$(2, 1)$	11	$(10, 3)$
5	$(9, 4)$	12	$(7, 2)$
6	$(5, 2)$	13	$(649, 180)$
7	$(8, 3)$	14	$(15, 4)$
8	$(3, 1)$	15	$(4, 1)$

As we saw in Section 4.1, corresponding to the diagram

$$f_0 \xrightarrow{n_0} f_1 \xrightarrow{n_1} f_2 \xrightarrow{n_2} \cdots , \tag{5.2.21}$$

with $f_0 = [1, 0, -d\,]$ of discriminant $D = 4d \in \mathbb{Z}^+$, is the parallel diagram

$$\beta_0 \xrightarrow{n_0} \beta_1 \xrightarrow{n_1} \beta_2 \xrightarrow{n_2} \cdots , \tag{5.2.22}$$

with $\beta_0 = Z(f_0) = \sqrt{d}$. According to the theory in Section 4.2, we have the minus continued fraction representation

$$\sqrt{d} = (n_0, n_1, n_2, \dots). \tag{5.2.23}$$

Recall that $n_0 = \left\lceil \sqrt{d} \right\rceil$, which gives the first entry in (5.2.23). The other entries in (5.2.23) are readily extracted from Table 3.1.3. For example, when $d = 5$, we look at the principal cycle of discriminant $D = 20$ to obtain

$$\sqrt{5} = (3, 2, 2, 2, 6, 2, 2, 2, 6, 2, 2, 2, 6, \dots). \tag{5.2.24}$$

Given the infinite periodic repeating pattern of the four integers $2, 2, 2, 6$ in (5.2.24), we may abbreviate the representation in (5.2.24) as

$$\sqrt{5} = \left(3, \overline{2, 2, 2, 6}\right), \tag{5.2.25}$$

since no information is lost by using (5.2.25) instead of (5.2.24). Proceeding in this way, we display in Table 5.2.2 below the minus continued fraction representation of $\sqrt{d}$ for all values of $d \in \mathbb{Z}^+$ less than 16 that are not perfect squares.

Table 5.2.2. Minus continued fraction representation of $\beta = \sqrt{d}$

d	$(n_0, n_1, n_2, \dots)$	d	$(n_0, n_1, n_2, \dots)$
2	$\left(2, \overline{2, 4}\right)$	10	$\left(4, \overline{2, 2, 2, 2, 2, 8}\right)$
3	$\left(2, \overline{4}\right)$	11	$\left(4, \overline{2, 2, 8}\right)$
5	$\left(3, \overline{2, 2, 2, 6}\right)$	12	$\left(4, \overline{2, 8}\right)$
6	$\left(3, \overline{2, 6}\right)$	13	$\left(4, \overline{3, 3, 2, 2, 2, 2, 2, 3, 3, 8}\right)$
7	$\left(3, \overline{3, 6}\right)$	14	$\left(4, \overline{4, 8}\right)$
8	$\left(3, \overline{6}\right)$	15	$\left(4, \overline{8}\right)$

Finally, as we saw in Section 4.2, the theory of minus continued fractions provides us with an infinite sequence of rational numbers converging strictly downwards to the limit value β. If β is an irrational number such as $\sqrt{5}$, this is of interest in terms of obtaining a better and better decimal approximation to a number that can be typically difficult to handle otherwise. To give a good idea of just how fast the convergence to the limit value takes place in terms of the minus continued fraction algorithm, we display in Table 5.2.3 below the rational number convergents $r_0, r_1, r_2, \dots, r_{20}$ to $\beta = \sqrt{5}$. We note that r_{20} gives an approximation to $\sqrt{5}$ that is correct to 12 decimal places to the right of the decimal point.

Table 5.2.3. Convergents for $\sqrt{5}$

j	p_j	q_j	$r_j = p_j/q_j$
0	3	1	3.00000000000000000000
1	5	2	2.50000000000000000000
2	7	3	2.33333333333333333333
3	9	4	2.25000000000000000000
4	47	21	2.23809523809523809524
5	85	38	2.23684210526315789474
6	123	55	2.23636363636363636364
7	161	72	2.23611111111111111111
8	843	377	2.23607427055702917772
9	1525	682	2.23607038123167155425
10	2207	987	2.23606889564336372847
11	2889	1292	2.23606811145510835913
12	15127	6765	2.23606799704360679970
13	27365	12238	2.23606798496486353979
14	39603	17711	2.23606798035119417311
15	51841	23184	2.23606797791580400276
16	271443	121393	2.23606797756048536571
17	491045	219602	2.23606797752297337911
18	710647	317811	2.23606797750864507522
19	930249	416020	2.23606797750108167877
20	4870847	2178309	2.23606797749997819409

BIBLIOGRAPHY

[1] David S. Dummit and Richard M. Foote. *Abstract Algebra*. John Wiley and Sons, 3rd Edition, 2004.

[2] Harold M. Edwards. *Fermat's Last Theorem: A Genetic Introduction to Algebraic Number Theory*. Springer-Verlag, 1977.

[3] Carl Friedrich Gauss. *Disquisitiones Arithmeticae*. Springer-Verlag, 1986.

[4] Svetlana Katok. *Continued fractions, hyperbolic geometry and quadratic forms*. MASS selecta, pp. 121–160, American Mathematical Society, 2003.

[5] Edmund Landau. *Elementary Number Theory*. AMS-Chelsea Publishing, 1999.

[6] Winfried Scharlau and Hans Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and Its Historical Development*. Springer-Verlag, 1985.

[7] André Weil. *Number Theory: An approach through history from Hammurapi to Legendre*. Birkhäuser, 1984.

[8] Don B. Zagier. *Zetafunktionen und quadratische Körper: Eine Einführung in die höhere Zahlentheorie*. Springer-Verlag, 1981.