Essays on the arithmetic of quadratic fields

Bill Casselman
University of British Columbia
cass@math.ubc.ca

Indefinite binary forms and real quadratic fields

Suppose $N > 0$ square-free, $F = \mathbb{Q}(\sqrt{N})$. Assume F to be embedded into $\mathbb{R}$, and let $\sqrt{N}$ be the positive square root. In this essay I'll discuss the classification of lattices in F. This is at once more complicated and more interesting than in the case $N < 0$.

The basic question is, *how to compute the proper equivalence classes of forms associated to lattices in F?* The theory again comes down to a matter of classifying **reduced forms**, but the problem does not involve the same kind of geometry that it does when $N < 0$, because the group $\mathrm{SL}_2(\mathbb{Z})$ does not act discretely on the appropriate domain. Instead, the computation turns out to be related to the computation of continued fractions of quadratic irrationals.

Contents

1. Reduced forms

An element γ of F is called **reduced** if

$$|\gamma| > 1, \quad |\overline{\gamma}| < 1, \quad \gamma\overline{\gamma} < 0$$

Of course $-\gamma$ is reduced if and only if γ is, but there is a more interesting duality:

1.1. Lemma. *The element γ is reduced if and only if $-1/\overline{\gamma}$ is.*

Suppose $Q(x, y)$ to be a primitive quadratic form with factorization

$$Q(x, y) = ax^2 + bxy + cy^2 = c(y - \gamma x)(y - \overline{\gamma} x) \tag{1.2}$$

where γ is irrational and lies in F. I make this specification unambiguous by setting

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

The first root γ is called the **characteristic root** of the form.

The form is called **reduced** if and only if its characteristic root is reduced. This can be formulated as conditions on a, b, c, which amount to a direct translation:

$$|\sqrt{D} - b| < 2|c| < |\sqrt{D} + b|, \quad b^2 < D\,.$$

,

1.3. Lemma. *Suppose $ax^2 + bxy + cy^2$ to be a primitive integral form. The following are equivalent:*

(a) the form is reduced;
(b) $0 < b < \sqrt{D}$ *and* $\sqrt{D} - b < 2|c| < \sqrt{D} + b$*;*
(c) $|\sqrt{D} - 2|c|| < b < \sqrt{D}$*.*

Also equivalent are conditions (b) and (c) with c *replaced by* a*. Finally,* a *and* c *necessarily have opposite signs.*

Proof. Suppose the form to be reduced.

Step 1. If $x, y > 0$ then $|x + y| = x + y$, while

$$|x - y| = \begin{cases} x - y & \text{if } y \le x \\ y - x & \text{if } y > x. \end{cases}$$

So certainly $|x + y| > |x - y|$. If we apply this to $|x - z|$ when $z < 0$, we see that $|x + z| < |x - z|$. From the inequality

$$|\sqrt{D} + b| > |\sqrt{D} - b|$$

we therefore deduce that • $0 < b$.

Step 2. Since $\gamma\overline{\gamma} < 0$

$$(-b - \sqrt{D})(-b + \sqrt{D}) = b^2 - D < 0\,.$$

This tells us that $b^2 < D$. Since $b > 0$, • $b < \sqrt{D}$.

Step 3. The inequalities

$$|\sqrt{D} - b| < 2|c| < |\sqrt{D} + b|$$

now become

$$\sqrt{D} - b < 2|c| < \sqrt{D} + b\,.$$

This concludes the proof that (a) implies (b). The rest of the implications are straightforward.

As for the final claim, $D = b^2 - 4ac$ and $b^2 < D$, so $4ac < 0$. It follows also from Lemma 1.1 and the observation that the characteristic root of the form

$$-cx^2 + bxy - ay^2$$

is $-1/\overline{\gamma}$. 

REDUCTION.

1.4. Proposition. *Every indefinite quadratic form with irrational* D *is properly equivalent to a reduced form.*

Proof. I offer the algorithm found in §183 of [Gauss:1801] or §73 of [Dirichlet:1863]. For brevity, I'll write $ax^2 + bxy + cy^2$ as (a, b, c), or sometimes even $(a, b, \ldots)$ or $(\ldots, b, c)$. Given that the discriminant remains the same, these designations are unambiguous.

If $Q = (a, b, c)$ is any primitive form, define new forms $\sigma(Q)$ and $\tau(Q)$ as

$$\sigma(Q) = (c, -b, a), \quad \tau(Q) = (a, b + 2na, 4n^2a + 2nb + c)\,,$$

with n determined uniquely by the condition

$$\sqrt{D} - 2|a| < b + 2na < \sqrt{D}\,.$$

These operations arise (respectively) from the substitutions

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_\circ \\ y_\circ \end{bmatrix}$$
$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_\circ \\ y_\circ \end{bmatrix}$$

Define ρ to be the composite $\tau\sigma$. Thus ρ changes (a,b,c) to $(c,-b+2nc,\dots)$ with

$$\sqrt{D}-2|c|<-b+2nc<\sqrt{D}\,.$$

The reduction algorithm is a consequence of:

1.5. Proposition. *If $Q=(a,b,c)$ is a primitive form then $\rho^n(Q)$ is reduced for $n\gg 0$.*

Proof. Suppose $\rho(Q)=(a_n,b_n,c_n)$. If $|a_n|>|c_n|$ then $|a_{n+1}|=|c_n|<|a_n|$. This decrease in $|a_n|$ can't go on forever, so sooner or later

$$|a_n|<|c_n|,\quad \sqrt{D}-2|a_n|<b_n<\sqrt{D}\,.$$

1.6. Lemma. *If $|a|<|c|$ and $\sqrt{D}-2|a|<b<\sqrt{D}$ then (a,b,c) is reduced.*

Proof. Since $b^2<D=b^2-4ac$, a and c must have opposite signs. Since $(\sqrt{D}-b)(\sqrt{D}+b)=D-b^2>0$, and $\sqrt{D}-b>0$, $\sqrt{D}+b>0$ as well. Since $|a|\le|c|$ we must have

$$(\sqrt{D}-b)(\sqrt{D}+b)=-4ac\ge 4|a|^2\,.$$

Since $\sqrt{D}-b<2|a|$, we must have $\sqrt{D}+b>2|a|$. From

$$\sqrt{D}+b>2|a|>\sqrt{D}-b$$

we deduce that $b>0$.

Remark. The procedure explained above is that of [Gauss:1801], but [Lagarias:1980] has pointed out that it is not as efficient as it might be. He proposes a simple modification that works much faster.

$$\circ\text{————}\circ$$

1.7. Theorem. *The set of reduced forms with a given discriminant $D>0$ is finite.*

Proof. It is easy to make a list of all them. Let $d=\lfloor\sqrt{D}\rfloor$. Since $D=b^2-4ac$, $b\equiv_2 D$. So we scan through all $b\equiv_2 D$ such that $b^2<D$. For each of these we scan through all a dividing $(D-b^2)/4$, and check whether $(d+1)-b\le 2a\le d+b$. For each a that passes, we set $c=-(D-b^2)/(4a)$ and add both (a,b,c) and $(-a,b,-c)$ to the list.

Since the number of divisors of any $n\ge 1$ is $O(n^\varepsilon)$ for any $\varepsilon>0$, the number of reduced forms if $O(D^{1/2+\varepsilon})$ for any $\varepsilon>0$.

2. Cycles

If (a,b,c) is a form, a **right neighbour** is a form $(a_\circ,b_\circ,c_\circ)$ of the same discriminant such that $a_\circ=c$ and $b\equiv -b_\circ$ modulo $2c$. For a left neighbour, c is replaced by a.

2.1. Proposition. *A reduced form has unique reduced right and left neighbours. The unique right neighbour is $\rho(Q)$, and the map taking Q to its unique reduced left neighbour is the inverse of ρ on the set of reduced forms.*

Proof. For the first part, the argument is almost the same for right and left, so I'll look just at right neighbours.

A necessary condition that $(a_\circ,b_\circ,c_\circ)$ be reduced is that

$$\sqrt{D}-2|c|<b_\circ=-b+2nc<\sqrt{D}\,.$$

Such an n is unique, and can be calculated by integral division. Hence there is at most one reduced right neighbour. What is more difficult is to show that the unique candidate is in fact reduced.

Let $b_\circ = -b + 2nc$. The candidate will be the unique $(c, b_\circ, c_\circ)$ with

$$\sqrt{D} - 2|c| < b_\circ < \sqrt{D}\,.$$

Specify $c_\circ$, as mentioned above, by the requirement that $b_\circ^2 - 4a_\circ c_\circ = D$. It remains to show that

$$\begin{aligned} 0 &< b_\circ \\ 2|c| &< \sqrt{D} + b_\circ\,. \end{aligned}$$

Step 1. Since (a, b, c) is reduced, according to the last remark in Lemma 1.3:

$$\begin{aligned} \sqrt{D} + b - 2|c| &> 0 \\ \sqrt{D} - 2|c| &< b_\circ \\ -\sqrt{D} + 2|c| + b_\circ &> 0 \\ 2n\,|c| = b + b_\circ &> 0\,. \end{aligned}$$

Hence • $n > 0$.

Step 2. Again since (a, b, c) is reduced:

$$\begin{aligned} \sqrt{D} - b &> 0 \\ b_\circ - (\sqrt{D} - 2|c|) &> 0 \\ b_\circ - b + 2|c| &> 0 \\ 2b_\circ - 2m|c| - 2|c| &> 0 \\ 2b_\circ > 2(n-1)|c| &\geq 0\,. \end{aligned}$$

Hence • $b_\circ > 0$.

Step 3. An easy deduction:

$$\sqrt{D} + b_\circ - 2|c| = \sqrt{D} - b + 2(n-1)|c| > 0\,.$$

This concludes the proof of the first part of the Proposition. The rest is straightforward.

Let $\sigma(Q)$ be the left neighbour of a reduced Q. Since $\sigma = \rho^{-1}$ on the set of reduced forms, ρ is a permutation of that set. The set of all reduced forms is therefore partitioned into cycles with respect to ρ.

3. Equivalence classes

One of the main results in the subject is this:

3.1. Theorem. *Two reduced forms are properly equivalent if and only if they lie in the same cycle.*

The proof will come down to well known results about continued fractions.

Step 1. I first call how quadratic forms are related to their characteristic elements in F.

Suppose $\Lambda = [\,\lambda \quad \mu\,]$ to be a positively oriented basis in F. To this corresponds the quadratic form

$$(x\lambda + y\mu)(x\overline{\lambda} + y\overline{\mu}) = [\,x \quad y\,]\,{}^t\overline{\Lambda}\cdot\Lambda\begin{bmatrix} x \\ y \end{bmatrix}.$$

On the other hand, Λ gives rise to the element $\omega = \lambda/\mu$ in F. If we make a coordinate substitution

$$\begin{bmatrix} x \\ y \end{bmatrix} = X \begin{bmatrix} x_\circ \\ y_\circ \end{bmatrix}$$

how does ω change?

3.2. Lemma. *In these circumstances, ω changes to*

$$\omega_\circ = {}^tX(\omega)\,.$$

I recall that if

$$X = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

then $X(\omega) = (a\omega + b)/(c\omega + d)$.

Proof. The substitution give us the expression

$$[\,x_\circ \quad y_\circ\,]\,{}^tX\,{}^t\overline{\Lambda}\cdot\Lambda X \begin{bmatrix} x_\circ \\ y_\circ \end{bmatrix}\,.$$

Thus the new basis $\Lambda_\circ$ is ΛX. But it is immediate that the characteristic associated to $\Lambda_\circ$ is ${}^tX(\omega)$.

Step 2. The map ρ is associated to the successive coordinate changes

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$
$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 & w \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$$

and therefore to the matrix

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & w \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & w \end{bmatrix}\,.$$

Therefore

$$\omega_\circ = \frac{1}{-\omega + w} = \frac{1}{w - \omega}\,.$$

so that

$$\omega = n - \frac{1}{\omega_\circ}\,.$$

Step 3. You can see already some relation to the basic step in computing continued fractions, but in order to make this relationship useful and precise, I have to consider signs. I am following here §77 of [Dirichlet:1863].

Since ω and $\omega_\circ$ are characteristic, $|\omega| > 1$ and $|\omega_\circ| > 1$. From the first, we see that $|n - \omega_\circ| < 1$. Since the first coefficients in the quadratic forms Q and $\rho(Q)$ alternate in sign, and the sign of a is the same as that of ω, we see that • the signs of ω and $\omega_\circ$ are opposite.

Step 4. For the same reason, we see that that $|\omega - w| < 1$ and the sign of $\omega - w$ is the opposite of that of $\omega_\circ$, and the same as that of ω. This specifies w uniquely:

$$\bullet\ |w| = \lfloor |\omega| \rfloor \ \text{ and } \ \omega/w > 0\,.$$

Step 5. I am following now §79 of [Dirichlet:1863]. Since signs of the forms alternate, the length of a cycle must be even. Pick one Q_0 in the cycle for which ω is positive, and for $n \geq 0$ set

$$\begin{aligned} \omega_n &= \text{characteristic element of } \rho^n(Q) \\ \kappa_n &= (-1)^n \omega_n \\ k_n &= (-1)^n w_n\,. \end{aligned}$$

We now have

$$\kappa_n = k_n + \frac{1}{\kappa_{n+1}}$$

for all $n \geq 0$. But these are all positive and larger than 1, so we have the continued fraction expressions

$$\kappa_n = \langle\langle\, k_n, k_{n+1}, k_{n+2}, \dots \,\rangle\rangle\,.$$

Because the cycle has finite length, the continued fraction expansion is periodic. (This is in accord with the fact that ω_0 is reduced.)

Step 6. Different cycles have different continued fraction expansions. Because the various ω in one cycle run through all possibilities, and ω determines the form.

Step 7. I recall Theorem 175 of [Hardy-Wright:1960] (see also Corollary 7.2 of [Casselman:2020]) : two real numbers ω_1 and ω_2 have the same tails, matching in parity, in their continued fractions if and only

$$\omega_1 = \frac{a\omega_2 + b}{c\omega_2 + d}\,.$$

with

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

in $\mathrm{SL}_2(\mathbb{Z})$. Therefore the different cycles are parametrized entirely by their periods (as cycles).

This concludes the proof of Theorem 3.1. ∎

4. The geometry of reduced forms

By choosing $\sqrt{N} > 0$, we have already embedded $F = \mathbb{Q}(\sqrt{N})$ in $\mathbb{R}$. Associated to this is an embedding of the F into $\mathbb{R}^2$:

$$\lambda \mapsto (\lambda, \overline{\lambda})\,.$$

A lattice L in F embeds as a lattice in $\mathbb{R}^2$.

REDUCED BASES. I shall call a basis $u = (x_u, y_u)$, $v = (x_v, y_v)$ of $\mathbb{R}^2$ **irrational** if x_u/x_v and y_u/y_v are irrational.

4.1. Lemma. *If L possesses one irrational basis, then all its bases are irrational.*

Proof. Because $\mathrm{GL}_2(\mathbb{Z})$ is generated by matrices

$$\begin{bmatrix} 1 & \pm 1 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 0 & \pm 1 \\ 1 & 0 \end{bmatrix}\,.$$

∎

I shall call a lattice in $\mathbb{R}^2$ irrational if it possesses an irrational basis.

4.2. Lemma. *If L is an irrational lattice in $\mathbb{R}^2$, the points of L approach arbitrarily closely to any rational ray $y = rx$ $(0 < x < \infty)$ in $\mathbb{R}^2$, and from either side.*

Proof. By Lemma 4.1, it suffices to take the ray to be the positive x-axis. In this case, it reduces to the well known result of Kronecker about approximation of irrationals by rationals. ∎

A **reduced element** of L is a point $u = (x_u, y_u)$ with $x_u > 0$ for which the open rectangle

$$B_u = \left\{(x, y) \,\middle|\, |x| < x_u, |y| < |y_u|\right\}$$

intersects L only in $(0, 0)$.

4.3. Lemma. *The lattice L contains reduced elements.*

In fact, as we shall see, it contains lots of them.

Proof. Any point of L in the positive quadrant of minimal (Euclidean) distance from $(0,0)$ will be reduced. ∎

Suppose $u = (x_u, y_u)$ to be a reduced point of L. Let $v = (x_v, y_v)$ be the point in L to its right with the properties (a) $|y_v| < |y_u|$ and (b) its x-coordinate is least among points satisfying (a). This is possible because of Lemma 4.2, since the coordinates of points in L are irrational. The point v is called the **right neighbour** of u.

4.4. Lemma. *In these circumstances, u and v form a basis of L. They lie on opposite sides of the x-axis, and v is also a reduced point of L.*

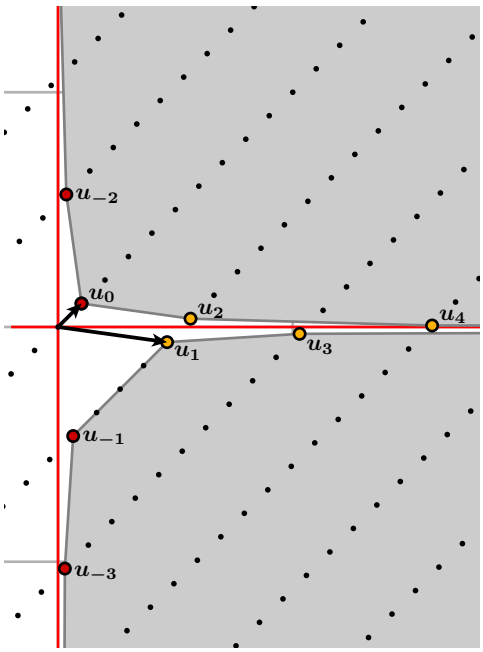
Proof. If v were to lie on the same side of the x-axis as u, then $v - u = (x, y)$ would also lie in L, If $x < x_u$ then $(v - u)$ would lie in the box B_u. So $x > x_u$. If u and $v - u$ lie on the same side of the x-axis, then $y < y_u$ would be the neigbour of u. So v lies on the opposite side of the x-axis from u.

The same reasoning shows that v is a reduced point.

According to a well known result of Minkowski, for the first claim it suffices to show that the closed triangle with vertices at u, v, and the origin contains no point of L other than its vertices. But similar reasoning tells us that any such point would be the neighbour of u, instead of v. ∎

One can define similarly a **left neighbour** of u. Say for example, that $y_0 > 0$. According to Lemma 4.2, the region $|y| < y_0$, $x < x_0$ possesses at least one point of L. It therefore possesses a point w with the smallest value of $|y|$. It must lie in the region $y < 0$, because otherwise the point $u - w$ would have a smaller value of y. The point w is then reduced.

One obtains in this way a sequence $u_0 = u$, $u_1 = v$, ... of reduced points passing off to infinity along the x-axis. One can define similarly a sequence of reduced points $u_{-1}, u_{-2}, \ldots$ passing off to the left. We know that for each n the pair (u_n, u_{n+1}) form a basis of L. Because these points lie on alternate sides of the x-axis, the orientation of this basis also alternates.



These points have a very simple characterization:

4.5. Proposition. *If u_0 lies in the positive quadrant, the points u_n with n even are the vertices of the convex hull of the intersection of L with the positive quadrant, and the points u_n with n odd are the vertices of the convex hull of the intersection of L with the lower right quadrant.*

Proof.  

There is an algorithm for generating the sequence (u_n), given $u = u_0$ and $v = u_1$. We know that $x_v > x_u$, and that u and v lie on opposite sides of the x-axis. Say u lies on top of it. Then the sequence $(u + mv)_m$ is descending from u, and hence for some m we'll have $u + mv$ above the x-axis, and $u + (m + 1)v$ below it. In this case, $u + vm = u_2$. It continues, and in a remarkable way relatyed to the computation of contnued fractions.

$$u_{n+1} = u_{n-1} + \ell_n u_n \quad (\ell_n = \lfloor -y_{n-1}/y_n \rfloor)$$
$$u_{n-1} = u_{n+1} - m_n u_n \quad (m_n = \lfloor x_{n+1}/x_n \rfloor)$$

In fact, the ℓ_n and m_n can be computed to the familiar rules

$$\begin{aligned} \lambda_0 &= y_0/y_1 \\ \ell_n &= \lfloor \lambda_n \rfloor \\ \lambda_{n+1} &= \frac{1}{\lambda_n - \ell_n} \end{aligned}$$

$$\begin{aligned} \mu_0 &= x_1/x_0 \\ m_n &= \lfloor \mu_n \rfloor \\ \mu_{n+1} &= \frac{1}{\mu_n - m_n}. \end{aligned}$$

4.6. Proposition. *Every reduced basis is one of the* (u_n, u_{n+1}).

Proof.  

REDUCED FORMS.

4.7. Proposition. *Let* (u, v) *be a basis of a lattice in* F, $ax^2 + bxy + cy^2$ *the primitive form associated to it. Then the basis is reduced if and only if the form is reduced.*

Suppose $u_0 = (x_0, y_0)$ with $y_0 > 0$. Then $u_1 = (x_1, y_1)$ with $y_1 < 0$. The basis (u_1, u_0) is positively oriented, and the form Q_0 corresponding to it is reduced. This continues:

4.8. Proposition. *The basis* $(u_{2n}, -u_{2n-1})$ *is positively oriented, and the form* Q_{2n-1} *corresponding to it is* $\sigma(Q_{2n-1})$. *Similarly for* (u_{2n+1}, u_{2n}).

5. References

1. J. V. Armitage (editor), **Journées Arithmétiques 1980**, Cambridge University Press, 1982.

2. BillCasselman, 'Approximating irrational numbers by rational ones', preprint, 2020. Available at

`http://www.math.ubc.ca/~cass/research/pdf/cf.pdf`

3. Harold Davenport, **The higher arithmetic**, Cambridge University Press, 1992.

4. Carl Friedrich Gauss, **Disquisitiones arithmeticae**, 1801. An English translation from the original Latin by Arthur A. Clarke was published by Yale University Press in 1965.

5. Jeffrey Lagarias, 'Worst case complexity bounds for algorithms in the theory of integral quadratic forms', *Journal of algorithms* **1** (1980), 142–186.

6. Peter G. Lejeune-Dirichlet, **Vorlesungen über Zahlentheorie**, Braunschweig, 1863.

7. Hendrik W. Lenstra, Jr., 'On the calculation of regulators and class numbers of quadratic fields', pp. 123–150 in [Armitage:1982].