Quadratic reciprocity and the sign of the quadratic Gauss sum

January 17, 2014

Today we're going to use the ideas from the last lecture to prove the quadratic reciprocity law. First let me remind you about that law.

It's one of the strangest facts you encounter in elementary number theory. First, let p be an odd prime. A question we pretend to be interested in is the following: what are the squares (mod p)? Of course, 0 is always a square (mod p). Usually we implicitly discard that one, so we really mean to ask about the nonzero squares (mod p).

It's easy to give an abstract answer to this question. The nonzero squares (mod p) are exactly the elements of the unique index 2 subgroup $H \subset (\mathbb{Z}/p\mathbb{Z})^\times$ we discussed in the last lecture. But really that's not such a satisfying answer. To highlight this, let's change the question, and instead ask: Given a natural number $a \in \mathbb{N}$ not divisible by p, how can we tell whether or not a is a square (mod p)? Or rather, maybe, what's the fastest way of telling? It would be nice if you didn't have to first make a list of all the squares (mod p), then check whether a is among them.

One can get started on this question by the following remark: it suffices to consider the case when $a = \ell$ is also a prime. That's because the product of two squares is a square, the product of two non-squares is a square, and the product of a square and a non-square is a non-square. (This follows from the fact that H is an index two subgroup of $(\mathbb{Z}/p\mathbb{Z})^\times$, so the quotient group is isomorphic to ± 1.) So given an arbitrary a, you can write it as a product of primes. Then if you know how to tell whether each of those primes is a square (mod p) or not, you can just see whether the total number of "yes" answers is even or odd and you have your answer for a.

So our question is, given a prime $\ell \neq p$, when is ℓ a square (mod p)? The case $\ell = 2$ is somewhat exceptional, so let's throw it away, so that we're assuming both ℓ and p are odd. Then there's an amazing symmetry, given by the quadratic reciprocity law:

Theorem 0.1. *Let ℓ and p be distinct odd primes. Then ℓ is a square (mod p) if and only if $(-1)^{\frac{p-1}{2}} p$ is a square (mod ℓ).*

This theorem does give a fast way of checking whether an integer is a square modulo a prime. To illustrate this, let's see if we can decide whether 123 is a square (mod 53). Well, a good first step is to reduce 123 mod 53. This gives 17. So 123 is a square (mod 53) if and only if 17 is a square mod 53. Note that both of these are prime, so by the theorem this is the same as asking if -53 is a square (mod 17). But -53 is congruent to 15 (mod 17). Now 15 = 5*3, so now we branch off and see whether 3 and 5 are squares mod 17. First let's think about 3. By the theorem, it's equivalent to asking whether 17 is a square mod 3. But 17 is congruent to 2 mod 3, so it's not a square. As for 5, by the theorem it's equivalent to asking whether 17 is a square mod 5. But 17 is 2 mod 5, also not a square. We deduce that 15 is a square (mod 17), so in the end 123 is a square (mod 53).

In practice this is actually a really good algorithm. But that's not the best thing about the quadratic reciprocity law. The best thing about it is that after all these years, and despite numerous proofs and beautiful structural generalizations, it's still a mystery as to why it should be true. There's just no reason for why ℓ being a square (mod p) should have anything to do with p being a square (mod ℓ), let alone such a simple law as the above theorem posits.

OK, enough waxing on. Let's prove it. Recall from last time that if P and P' denote the sums $\sum_{a \in H} \zeta_p^a$ and $\sum_{b \notin H} \zeta_p^b$, then P and P' are the roots of the quadratic equation

$$X^2 + X + C,$$

where $C = -\frac{p-1}{4}$ if p is 1 (mod 4), and $C = \frac{p-3}{4}$ if p is 3 (mod 4). Thus, if we let $G = P - P'$, then

$$G^2 = (-1)^{\frac{p-1}{2}} p,$$

i.e.

$$G = \pm \sqrt{(-1)^{\frac{p-1}{2}} p}.$$

Now, as we were making those calculations, you were probably imagining everything taking place inside $\mathbb{C}$. But it was actually all algebraic, just having to do with p^{th} roots of unity. So in fact it all makes sense in any algebraically closed field of characteristic different from p. Actually, in the very last step there, we might want to also assume that the characteristic is different from 2, so that there really are exactly two square roots for any nonzero element of the field.

In particular, all of the above works in $\overline{\mathbb{F}_\ell}$, an algebraic closure of $\mathbb{F}_\ell$. We find that if ζ_p is a primitive p^{th} root of unity in $\overline{\mathbb{F}_\ell}$, then the element

$$G = \sum_{a \in H} \zeta_p^a - \sum_{b \notin G} \zeta_p^b \in \overline{\mathbb{F}_\ell}$$

is, up to sign, the unique square root of $(-1)^{\frac{p-1}{2}} p$. Thus, $(-1)^{\frac{p-1}{2}} p$ is a square (mod ℓ) if and only if $G \in \mathbb{F}_\ell$. Now, we can test this by considering the action of Frobenius on G. Recall that Frobenius is the map $x \mapsto x^\ell$ from $\overline{\mathbb{F}_\ell}$ to itself. It's a field homomorphism because we're in characteristic ℓ. Furthermore, the fixed points of Frobenius are exactly the elements of $\mathbb{F}_\ell$. Indeed, to be fixed by Frobenius is to be a root of the polynomial $X^\ell - X$. Each element of $\mathbb{F}_\ell$ satisfies this polynomial, and there can be no more roots since the degree is ℓ.

So, $(-1)^{\frac{p-1}{2}} p$ is a square (mod ℓ) if and only if $Frob(G) = G$.

But on the other hand, $Frob(\zeta_p^a) = \zeta_p^{\ell a}$. If ℓ is a square (mod p), so ℓ lies in H, then $a \mapsto \ell a$ preserves H and $G - H$, so that $Frob(G) = G$. But if ℓ is not a square (mod p), then $a \mapsto \ell a$ switches H and $G - H$, so $Frob(G) = -G$. Since $2 \neq 0$ in $\overline{\mathbb{F}_\ell}$, we deduce that $Frob(G) = G$ if and only if ℓ is a square (mod p).

Combining our knowledge, we deduce the theorem: $(-1)^{\frac{p-1}{2}} p$ is a square (mod ℓ) if and only if $Frob(G) = G$ if and only if ℓ is a square (mod p).

It's worth saying some words about this proof. The essential fact was that $\mathbb{Q}(\zeta_p)$ contains $\mathbb{Q}(\sqrt{(-1)^{\frac{p-1}{2}} p})$ as a subfield. (You also saw this on your first homework.) Then the fact that Frobenius is easy to understand on $\mathbb{Q}(\zeta_p)$ implies that Frobenius is easy to understand on $\mathbb{Q}(\sqrt{(-1)^{\frac{p-1}{2}} p})$. This means that

it's easy to understand when $(-1)^{\frac{p-1}{2}}p$ is a square (mod ℓ). When you unwind it, you get the quadratic reciprocity law.

Actually, this is part of a beautiful and general story. Recall that, in fact, Frobenius is easy to understand on any cyclotomic field $\mathbb{Q}(\zeta_m)$. (This can be measured by the fact that the effect of the ℓ-Frobenius only depends on the value of ℓ (mod m) — a very regular pattern.) Thus, Frobenius is easy to understand on any subfield of any cyclotomic field. Thus the question arises as to how one can recognize which number fields $F/\mathbb{Q}$ are subfields of some cyclotomic field. The answer is provided by the Kronecker-Weber theorem: a number field F is a subfield of a cyclotomic field if and only if $F/\mathbb{Q}$ is Galois with abelian Galois group. So the upshot is that it's easy to understand Frobenius on abelain Galois extension of $\mathbb{Q}$; so if you have a polynomial over $\mathbb{Q}$ with abelian Galois group, the pattern of its factorizations modulo primes ℓ is very regular. For quadratic extensions, this regularity is equivalent to the quadratic reciprocity law.

A subject called "class field theory" establishes an analog of this, where you replace the "base field" $\mathbb{Q}$ by an arbitrary number field F. (Imagine that you are intrinsically interested in arithmetic in $\mathcal{O}_F$, as opposed to arithmetic in $\mathbb{Z}$). The result is that there exist certain extension $F_{\mathfrak{m}}/F$, the "ray class fields", on which Frobenius is very regular (only depends on certain congruence conditions). And again you have the result that every abelian extension F'/F is a subfield of a ray class field.

There is one perplexing aspect of the general theory, though, which is that although one knows has this analog $F_{\mathfrak{m}}$ of the fields $\mathbb{Q}(\zeta_m)$, one has, outside special cases, no analog of the elements ζ_m themselves. We don't know nice generators for the extension $F_{\mathfrak{m}}/F$ on which Frobenius is easy to understand; we just know $F_{\mathfrak{m}}/F$ has to exist by some very involved counting argument.

But let's leave that aside and turn to more concrete matters. There's an interesting question raised by the above proof. Namely, let's work inside $\mathbb{C}$ again, and take $\zeta_p = exp(2\pi i/p)$ in our definition of G. We know that

$$G = \pm\sqrt{(-1)^{\frac{p-1}{2}}p}.$$

But G was a completely specified complex number. So we can well ask whether the value is $+$ or $-$ here. (We adopt the convention that $\sqrt{-x} = i\sqrt{x}$ when $x > 0$.)

The answer is that it's always $+$, irrespective of p. This was noticed by Gauss, and proved by him a while later. It's a really funky fact. Let's give a proof, because it will lead into what we're going to do next.

The idea is the following. In your first problem set, you found another square root of $(-1)^{\frac{p-1}{2}}p$. It was (or can be taken to be)

$$G' = \prod_{k=1}^{\frac{p-1}{2}}(\zeta_p^{-k/2} - \zeta_p^{k/2}).$$

Remember, this came from cleverly splitting the identity $p = (1-\zeta_p)\dots(1-\zeta_p^{p-1})$ in half. Now, also on your first problem set you saw that actually $G' = +\sqrt{(-1)^{\frac{p-1}{2}}p}$: here the sign is a plus. The reason G' is easier to analyze than G is that G' is given by a product, and the argument of complex numbers is additive in products. Each of the factors is basically the sin of some easy to draw angle, so it's not hard to figure out the argument of the product and see it lies in the appropriate half-plane.

So we're just faced with showing that $G = G'$: our two square roots are the same. Now this is a

purely algebraic fact: we're claiming that

$$\sum_{a\in H}\zeta_p^a - \sum_{b\notin H}\zeta_p^b = \prod_{k=1}^{\frac{p-1}{2}}(\zeta_p^{-k/2} - \zeta_p^{k/2}).$$

It doesn't matter here what primitive p^{th} root of unity ζ_p is: since this is algebraic, we can act by Galois to change any one into another.

I don't know any direct proof of the above algebraic identity. Instead we'll argue somewhat indirectly that the two sides are equal.

Here is the idea. We know that G^2 is p, up to a sign. Thus, the maximal ideal factorization of G in $\mathbb{Z}[\zeta_p]$ must be $(G) = (\pi)^{\frac{p-1}{2}}$, where $\pi = 1 - \zeta_p$. This means that G is a unit times $\pi^{\frac{p-1}{2}}$. Now, we won't explicitly identify that unit, but we'll determine its value (mod π). And we'll see that this value (mod π) is the same for G and for G'. Since G and G' differ by a sign, and 1 and -1 are distinct in $\mathbb{Z}[\zeta_p]/\pi = \mathbb{F}_p$, it will follow that necessarily $G = G'$.

To make this analysis, it's helpful to think in terms of π-adic expansions of elements of $\mathbb{Z}[\zeta_p]$. These are defined as follows. Let $x \in \mathbb{Z}[\zeta_p]$. Since $\mathbb{Z}[\zeta_p]/\pi = \mathbb{F}_p$, there is a unique $c_0 \in \{0, 1, \ldots, p-1\}$ such that

$$x = c_0 + \pi \cdot x',$$

where $x' \in \mathbb{Z}[\zeta_p]$. Applying the same procedure to x' and iterating, we obtain a formal expansion

$$x = c_0 + c_1\pi + c_2\pi^2 + c_3\pi^3 + \ldots$$

uniquely characterized by the fact that each coefficient c_i lies in $\{0, 1, \ldots, p-1\}$, and the difference between x and the n^{th} partial sum is divisible by π^{n+1}.

As Martin explained to me, this formal expansion need not terminate, unlike an ordinary base-p expansion for a natural number. Indeed, the base-π-expansion of ζ_p^{-1} is

$$\zeta_p^{-1} = 1 + \pi + \pi^2 + \ldots$$

since $\zeta_p^{-1} = 1 + \pi \cdot \zeta_p^{-1}$.

Note that the coefficents $c_0, c_1, \ldots, c_n$ up to some fixed n only depend on the value of x in the quotient $\mathbb{Z}[\zeta_p]/\pi^{n+1}$. In particular, if we're only interested in the first few coefficients $c_0, \ldots, c_{p-2}$, then these only depend on x (mod p). This is convenient, because then we can actually think of the coefficients themselves as being well-defined elements of $\mathbb{F}_p$. Another way of saying all this is that the $p-1$-dimensional $\mathbb{F}_p$-vector space $\mathbb{Z}[\zeta_p]/p$ has a basis given by the first $p-1$ powers of π. That was implicit in the first proof of the irreducibility of $\Phi_p(X)$.

So, let us find the first $p-1$ coefficients in the π-adic expansion of G; or in other words, let's write G (mod p) in terms of the basis $1, \pi, \pi^2, \ldots, \pi^{p-2}$ for $\mathbb{Z}[\zeta_p]/p$. Actually, we know since $(G) = (\pi)^{\frac{p-1}{2}}$ that the first $\frac{p-1}{2}$ of these coefficients have to vanish, and in fact we're only interested in the first nonzero coefficient, the coefficient of $\pi^{\frac{p-1}{2}}$.

Before digging in, it will be convenient to rewrite G as

$$G = \sum_{a\in\mathbb{F}_p}\chi(a)\zeta_p^a,$$

where $\chi(a)$ denotes the unique nontrivial homomorphism $(\mathbb{Z}/p\mathbb{Z})^\times \to \pm 1$, i.e. the map taking the value 1 on H and -1 on the complement of H. OK, now let's dig in. We want to expand this in powers of π, so let's rewrite ζ_p as $1-\pi$. This gives

$$\begin{aligned} G &= \sum_{a\in\mathbb{F}_p} \chi(a)(1-\pi)^a \\ &= \sum_{a=0}^{p} \chi(a) \sum_{k=0}^{p} \binom{a}{k} (-\pi)^k \\ &= \sum_{k=0}^{p} (-\pi)^k \cdot \sum_{a=0}^{p} \chi(a) \binom{a}{k}. \end{aligned}$$

Now we use the fact that we're working (mod p) to write $\chi(a) = a^{\frac{p-1}{2}}$. (This is because $a \mapsto a^{\frac{p-1}{2}}$ (mod p) also defines a nontrivial homomorphism $(\mathbb{Z}/p\mathbb{Z})^\times \to \pm 1$.) Thus the coefficient of $(-\pi)^k$ above is

$$c_k = \sum_{a=0}^{p} a^{\frac{p-1}{2}} \frac{a(a-1)\dots(a-k+1)}{k(k-1)\dots 1}.$$

Now, what's special about this expression is that it's the sum of the values of a polynomial in $\mathbb{F}_p[X]$. Here's a lemma about that:

Lemma 0.2. *Let $p(X) = b_0 + b_1 X + \ldots + b_d X^d \in \mathbb{F}_p[X]$. Then*

$$\sum_{a=0}^{p} p(a) = -(b_{p-1} + b_{2(p-1)} + b_{3(p-1)} + \ldots)$$

Proof. Both sides are additive in $p(X)$, so it suffices to verify the identity when $p(X) = X^d$. Thus we need to see that

$$\sum_{a=0}^{p} a^d$$

is 0 unless $(p-1)|d$ and $d > 0$, in which case it is -1. First suppose $(p-1)|d$ and $d > 0$. Then $a^d = 1$ for all $a \neq 0$, and $0^d = 0$. Thus the sum gives $p - 1 = -1$. If $d = 0$, the sum clearly gives $p = 0$. So now suppose that $(p-1)$ doesn't divide d. We need to see that the sum is 0. Let x be a (for now) arbitrary nonzero element of $\mathbb{F}_p$. Then since multiplication by x is a bijection on $\mathbb{F}_p$,

$$x^d \sum_{a=0}^{p} a^d = \sum_{a=0}^{p} (ax)^d = \sum_{a=0}^{p} a^d.$$

Thus $(x^d - 1)\sum_{a=0}^{p} a^d = 0$, so to see that our sum is zero it's enough to see that there exists an x for which $x^d \neq 1$. But for that we can just take x to be a primitive $(p-1)^{st}$ root of unity, i.e. a generator for $(\mathbb{Z}/p\mathbb{Z})^\times$. □

Now, we're trying to evaluate $c_{\frac{p-1}{2}}$. We saw above that this is the sum of the values of the polynomial

$$X^{\frac{p-1}{2}} \frac{X(X-1)\dots(X-\frac{p+1}{2})}{\left(\frac{p-1}{2}\right)!}.$$

This polynomial has degree $p-1$, so according to the above lemma the sum of its values is just the negative of the leading coefficient, which is obviously $-\frac{1}{(\frac{p-1}{2})!}$. We deduce, in the end, that the $\frac{p-1}{2}^{th}$ coefficient of the π-adic expansion of G is

$$-\frac{(-1)^{\frac{p-1}{2}}}{\left(\frac{p-1}{2}\right)!}.$$

Actually, this expression can be simplified. Note that, modulo p, the product $(\frac{p+1}{2})\dots(p-1)$ identifies with $(-1)^{\frac{p-1}{2}} \cdot \left(\frac{p-1}{2}\right)!$, because we can pair up a number with its negative (mod p). Thus

$$(p-1)! = (-1)^{\frac{p-1}{2}} \cdot \left[\left(\frac{p-1}{2}\right)!\right]^2.$$

But on the other hand $(p-1)! = -1$ (mod p), for the following reason. Imagine pairing each element of $\{1, 2, \dots, p-1\}$ up with its multiplicative inverse (mod p). In general everyone has a partner. But there are two exceptions, the two solutions to $x = x^{-1}$, or $x^2 = 1$: that's 1 and $p-1$. These are their own inverses. We deduce that the product of all these elements is $1 \cdot (p-1) = -1$, as claimed. (Combining with the above formula, we get another proof that -1 is a square when p is 1 mod 4: in fact we found an explicit square root...!)

Thus, actually, the coefficient of $\pi^{\frac{p-1}{2}}$ in the π-adic expansion of G is $\left(\frac{p-1}{2}\right)!$.

Now let's do the analog of this calculation for G'. Thankfully, this is much easier, because G' is a product. It suffices to figure out just the first step of the π-adic expansion of each of the $\frac{p-1}{2}$ factors of G', then to multiply these answers together. Each factor has the form

$$\zeta_p^{-k/2} - \zeta_p^{k/2} = (1-\pi)^{-k/2} - (1-\pi)^{k/2}.$$

Expanding out, we see that (mod π^2) this evaluates to $1 + \frac{k}{2}\pi - (1 - \frac{k}{2}\pi) = k\pi$. Taking the product from $k=1$ to $\frac{p-1}{2}$, we get the desired conclusion: the coefficient of $\pi^{\frac{p-1}{2}}$ in the π-adic expansion of G' is also $\left(\frac{p-1}{2}\right)!$.

Thus, since we already know that $G = \pm G'$, we deduce that $G = G'$, so the sign of G is the same as that of G', which is $+$.

I've always found it amazing that we have these two different square roots of $(-1)^{\frac{p-1}{2}}p$, and that in the end they're actually the same square root. And the way you see it is (sort of) by p-adic analysis. But I should mention that there are other proofs that the sign of G is $+$. Actually, one proof sees it falling out of a more refined analysis of the Dirichlet L-series $L(s, \chi)$. Recall that the same Gauss sum G naturally came up when we calculated $L(1, \chi)$. If we were to have taken the time to prove the "functional equation" for $L(s, \chi)$, then we could've gotten an easy proof of the sign of G.

Anyway, next time we'll investigate more general analogs of this Gauss sum G.