1.14 26/07/07 - The Proof of the Nagell-Lutz Theorem

The goal for today is to prove the Nagell-Lutz Theorem (1.13.7). The proof of 1.13.7 is a consequence of the following two propositions:

Proposition 1.14.1. Let $E : y^2 = x^3 + ax^2 + bx + c$ be an elliptic curve over $\mathbb{Q}$ with $a, b, c \in \mathbb{Z}$ and $D = 2^{-4}\Delta$. Suppose $P = (x_0, y_0) \in E(\mathbb{Q}) \smallsetminus \{\mathscr{O}\}$ is a point such that $y_0 \neq 0$, i.e., $2P \neq \mathscr{O}$. Suppose further that P and $2P$ both have integer coordinates. Then $y_0^2 \mid D$.

Proposition 1.14.2. Let $E : y^2 = x^3 + ax^2 + bx + c$ be an elliptic curve over $\mathbb{Q}$ with $a, b, c \in \mathbb{Z}$. If $P = (x_0, y_0) \in \operatorname{Tors} E(\mathbb{Q}) \smallsetminus \{\mathscr{O}\}$, then $x_0, y_0 \in \mathbb{Z}$.

Given these two results, let's prove 1.13.7.

Proof of 1.13.7, assuming 1.14.1 and 1.14.2. Suppose $P = (x_0, y_0) \in \operatorname{Tors} E(\mathbb{Q}) \smallsetminus \{\mathscr{O}\}$. It follows from 1.14.2 that $x_0, y_0 \in \mathbb{Z}$. If $y_0 = 0$, we are done. Else $y_0 \neq 0$, which is equivalent to $2P \neq \mathscr{O}$. Since $2P \in \operatorname{Tors} E(\mathbb{Q}) \smallsetminus \{\mathscr{O}\}$ as well, we conclude from 1.14.2 again that $2P$ has integer coordinates. It then follows from 1.14.1 that $y_0^2 \mid D$. ∎

It remains to prove 1.14.1 and 1.14.2. I think of the first of these (1.14.1) as being algebraic trickery, whereas I think there is much more structure and theory behind the proof of 1.14.2 (although a very short proof of 1.14.2 involving algebraic trickery can be given by using 2.2.10).

1.14.A Proof of 1.14.1

Proof of 1.14.1. Recall the duplication formula 1.12.12 for the coordinates of $x(2P)$; in our case, it takes the form

$$x(2P) = \frac{f'(x)^2}{4f(x)} - a - 2x = \frac{g(x)}{4f(x)}, \tag{1.14.3}$$

where

$$g(x) = f'(x)^2 - 4(a + 2x)f(x) = x^4 - 2bx^2 - 8cx + b^2 - 4ac.$$

From $D = \operatorname{disc}(f) \neq 0$ we know that $f(x)$ and $f'(x)$ don't have any common (complex) roots; from the formula $g(x) = f'(x)^2 - 4(a + 2x)f(x)$, it then follows that $f(x)$ and $g(x)$ don't have any common (complex) roots either. In particular, since $\gcd(f(x), g(x)) = 1$ in the Euclidean domain $\mathbb{Q}[x]$, the Euclidean algorithm can be used to produce polynomials $p(x), q(x) \in \mathbb{Q}[x]$ such that

$$p(x)f(x) + q(x)g(x) = 1.$$

In fact, one can check that we may take $p(x) = D^{-1}P(x)$ and $q(x) = D^{-1}Q(x)$, where $P(x), Q(x) \in \mathbb{Z}[x]$ are the polynomials given by

$$P(x) = 3x^3 - ax^2 - 5bx + 2ab - 27c \text{ and } Q(x) = -3x^2 - 2ax + a^2 - 4b.$$

(Here we are using that $a, b, c \in \mathbb{Z}$.) Clearing denominators, we get an identity of the form

$$P(x) \cdot f(x) + Q(x) \cdot g(x) = D, \tag{1.14.4}$$

where $f(x), g(x), P(x), Q(x) \in \mathbb{Z}[x]$ (check!).

Now suppose we are given a $P = (x_0, y_0) \in E(\mathbb{Q}) \smallsetminus \{\mathscr{O}\}$ such that $y_0 \neq 0$ (which is equivalent to $f(x_0) \neq 0$, which is equivalent to $2P \neq \mathscr{O}$), and suppose that P and $2P$ have integer coordinates. In particular, $x_0, y_0 \in \mathbb{Z}$. From $y_0^2 = f(x_0)$, the integrality of $x(2P)$, and the formula 1.14.3, it follows that $y_0^2 = f(x_0) \mid g(x_0)$. It then follows from the identity 1.14.4 evaluated at $x = x_0$ that $y_0^2 \mid D$ as needed. ∎

Example 1.14.5. Note that 1.14.1 has nothing to do with torsion points. Let's revisit the example $E : y^2 = x^3 - 16x + 16$ of 1.13.9, in which $D = 2^{12} \cdot 37$. If we take $P = (0, 4) \in E(\mathbb{Q})$ as in that example, then we see that the fact that $2P$ has integer coordinates implies that $y(P)^2 \mid D$, which is in fact true. The same applies to the y-coordinates of $2P$ and $3P$, but not to the y-coordinate of $4P$, and indeed $8P = (84/25, -52/125)$.

Remark 1.14.6. Here's another perspective on why we might expect an equation of the sort 1.14.4 to be true. We know from the argument above that the coprimality of $f(x)$ and $g(x)$ implies that there is an identity of the form 1.14.4 for some integer polynomials $P(x), Q(x) \in \mathbb{Z}[x]$ and *some* integer $D' \in \mathbb{Z}_{\geq 1}$ instead of D; for instance, we can take D' to be the resultant $\mathrm{Res}_x(f(x), g(x))$. Now if p is a prime with $p \nmid 2D$, then the same equation $y^2 = f(x)$ also defines an elliptic curve over E over p, and the same argument as above–this time over $\mathbb{F}_p$ instead of $\mathbb{Q}$–allows us to conclude that $\overline{f}(x)$ and $\overline{g}(x)$ have no common complex roots over $\overline{\mathbb{F}}_p$, and this is sufficient to conclude that $p \nmid \mathrm{Res}_x(f(x), g(x))$, since the resultant behaves well under reduction mod p. Therefore, the only primes dividing D' can be the primes dividing D (and possibly 2).

Unimportant Remark 1.14.7. The above computation may tempt you to believe that the resultant of $f(x)$ and $g(x)$ with respect to x is actually D. This is incorrect: the resultant (as you can check!) is $\mathrm{Res}_x(f(x), g(x) = D^2$. Therefore, it is somehow stranger than usual that we get the stronger result 1.14.4. (This is remarked also in a footnote to [2, Exercise 2.11].)

1.14.B Proof of 1.14.2

To do this, we need to study more carefully the denominators that appear in rational points on elliptic curves. The first thing we need to study is what denominators even look like. If you have done the exercises (2.1.9), then the following result will not come as a surprise at all.

Proposition 1.14.8. Let $E/\mathbb{Q}$ be an elliptic curve in short Weierstrass form $E : y^2 = x^3 + ax^2 + bx + c$ with $a, b, c \in \mathbb{Z}$. If $P = (x, y) \in E(\mathbb{Q}) \smallsetminus \{\mathscr{O}\}$, then x and y when expressed as fractions in lowest terms look like

$$(x, y) = \left(\frac{m}{e^2}, \frac{n}{e^3}\right),$$

where $m, n, e \in \mathbb{Z}$ with $e \geq 1$ and $\gcd(m, e) = \gcd(n, e) = 1$.

Proof. For any prime p, the equation

$$y^2 = x^3 + ax^2 + bx + c$$

along with $a, b, c \in \mathbb{Z}$ implies that $v_p(x) < 0$ iff $v_p(y) < 0$, in which case $2v_p(y) = 3v_p(x)$. This–along with unique factorization in $\mathbb{Z}$!–is enough to prove the claim (check!). ∎

Exercise 1.14.9. Convince yourself that the above proof works.

Remark 1.14.10. Note that 1.14.8 shows in particular, that if $P = (x, y) \in E(\mathbb{Q}) \smallsetminus \{\mathscr{O}\}$, then $x \in \mathbb{Z}$ iff $y \in \mathbb{Z}$. But we knew this already: this follows directly from the Rational Root Theorem (1.1.2).

The above proof shows that it is perhaps wise to look at one prime at a time. Further experimentation with the curves in the exercises would seem to show that if p is a fixed prime and $P = (x, y) \in E(\mathbb{Q}) \smallsetminus \{\mathscr{O}\}$ is such that $v_p(x) < 0$, then $v_p(x(kP)) < 0$ for all $k \geq 1$. This is indeed correct. Let's now formalize this intuition, by defining the subset of points of $E(\mathbb{Q})$ with p "in the denominators". While we're doing this, we might as well keep track of how many times p goes in the denominators. For this reason, for an elliptic curve $E/\mathbb{Q}$ in Weierstrass form, prime p, and integer $r \in \mathbb{Z}_{\geq 1}$, we consider the subset

$$E_{p,r}(\mathbb{Q}) := \{(x, y) \in E(\mathbb{Q}) \smallsetminus \{\mathscr{O}\} : v_p(x) \leq -2r\} \cup \{\mathscr{O}\}.$$

Note that by virtue of 1.14.8, the condition $v_p(x) \leq -2r$ may equivalently be stated as $v_p(y) \leq -3r$. If we write $(x, y) = (me^{-2}, ne^{-3})$ as above (1.14.8), then this condition is further equivalent to $p^r \mid e$.

Remark 1.14.11. Note also that if either $x = 0$ or $y = 0$, then $P \notin E_{p,r}(\mathbb{Q})$, by the convention that $v_p(0) = +\infty$. This is the convention we will stick to throughout.

In this way, we get for each prime p a chain of subsets

$$E(\mathbb{Q}) \supset E_{p,1}(\mathbb{Q}) \supset E_{p,2}(\mathbb{Q}) \supset \cdots \supset \{\mathscr{O}\}.$$

This is known as the *p-adic filtration* on $E(\mathbb{Q})$. Note finally also that for each p we have

$$\bigcap_{r \geq 1} E_{p,r}(\mathbb{Q}) = \{\mathscr{O}\}.$$

The first order of business then, is to show

Proposition 1.14.12. Let $E/\mathbb{Q}$ be an elliptic curve in short Weierstrass form $E : y^2 = x^3 + ax^2 + bx + c$ with $a, b, c \in \mathbb{Z}$. Then for each prime p and integer $r \in \mathbb{Z}_{\geq 1}$, the subset $E_{p,r}(\mathbb{Q}) \subset E(\mathbb{Q})$ is a subgroup.

I like to think of 1.14.12 as saying that the denominators are "persistant": if a prime p shows up in the denominator of (the x- and y-coordinates of) P, then it will do so also for $2P, 3P, \ldots$.

Unimportant Remark 1.14.13. If you know something about the p-adic numbers $\mathbb{Q}_p$, then you should think of the $E_{p,r}(\mathbb{Q})$ as the subgroup where the coordinates get very large p-adically. In other words, the $E_{p,r}(\mathbb{Q})$ form (the intersection with $E(\mathbb{Q})$ of) an open neighborhood basis of $\mathscr{O}$ consisting of subgroups of $E(\mathbb{Q}_p)$. We will not use or need this language, but it is helpful to keep in mind for when you later see $\mathfrak{p}$-adic filtrations more generally for $\mathfrak{p}$ a prime of a general number (or even global) field K.

To prove 1.14.12, it is helpful to have some general criterion to figure out when a subset of an abelian group is a subgroup.

Lemma 1.14.14. Let A be an abelian group. Let $A' \subset A$ be a subset such that

(i) we have $0 \in A'$,
(ii) if $a \in A'$ then also $-a \in A'$, and
(iii) if $a_1, a_2, a_3 \in A$ are such that $a_1 + a_2 + a_3 = 0$, and two of the three a_i are in A', then so is the third.

Then $A' \subset A$ is a subgroup.

Exercise 1.14.15. Prove 1.14.14. Is the generalization to a statement of the form "$n-1$ out of n implies all" true for all n?

The following proof has been adapted from the one found in [2, §2.4].

Proof of 1.14.12. Suppose we are given E, p, and r. We use the criterion of 1.14.14(a). I will leave checking conditions (i) and (ii) to the reader. Suppose we are now given points $P_1, P_2, P_3 \in E(\mathbb{Q})$ such that $P_1 + P_2 + P_3 = 0$, and such that $P_1, P_2 \in E_{p,r}(\mathbb{Q})$; we have to show that $P_3 \in E_{p,r}(\mathbb{Q})$ as well. Again, I will leave the case where one (or more) of the P_i is $\mathscr{O}$, or when $P_1 = P_3$ or $P_2 = P_3$ to the reader, and focus on the case where all three are not $\mathscr{O}$ and $P_3 \notin \{P_1, P_2\}$. Recall by 1.12.17 that $P_1 + P_2 + P_3 = 0$ is equivalent to saying that P_1, P_2, P_3 are collinear (counted with suitable multiplicities!); therefore, we have to show that if $P_1, P_2, P_3 \in E(\mathbb{Q}) \smallsetminus \{\mathscr{O}\}$ are collinear, with $P_3 \notin \{P_1, P_2\}$ and $P_1, P_2 \in E_{p,r}(\mathbb{Q})$, then also $P_3 \in E_{p,r}(\mathbb{Q})$.

At this point, the proof is a calculation best done in (t, s) coordinates as in 1.3.11, so that $(t, s) = (x/y, 1/y)$, and the curve E in these coordinates is

$$s = t^3 + at^2 s + bts^2 + cs^3.$$

Note that this affine patch $Y \neq 0$ does not contain precisely the nontrivial 2-torsion points on $E(\mathbb{Q})$. Note also that the condition of lying in $E_{p,r}(\mathbb{Q})$ when in this patch is totally equivalent to $v_p(t) \geq r \Leftrightarrow$

$v_p(s) \geq 3r$. This automatically includes the point $\mathscr{O} = (0,0)$ when we follow the convention $v_p(0) = +\infty$, as in 1.14.11. Suppose first that P_1, P_2, P_3 are all in this patch; we'll worry about any remaining edge cases at the very end. Therefore, suppose that $P_i = (t_i, s_i)$ with $t_i, s_i \in \mathbb{Q}$ for $i = 1, 2, 3$, and that these points are all collinear; in other words, there is a line ℓ such that $(\ell \cap E)(\mathbb{Q}) = \{P_1, P_2, P_3\}$, counted with multiplicites.

If ℓ is vertical, say given by $t = t_0$ for some $t_0 \in \mathbb{Q}$, then $t_0 = t_1 = t_2 = t_3$. In this case, $P_1 \in \ell(\mathbb{Q}) \cap E_{p,r}(\mathbb{Q})$ implies that $v_p(t_1) \geq r$, and this implies that $v_p(t_3) \geq r$ as needed. Having dealt with this easy case, we may now suppose that ℓ is not vertical; say it has equation

$$\ell : s = \lambda t + \nu$$

for some $\lambda, \nu \in \mathbb{Q}$. We need to figure out an expression for λ. If $t_1 \neq t_2$, then

$$\lambda = \frac{s_2 - s_1}{t_2 - t_1}.$$

Using the fact that

$$s_i = t_i^3 + at_i^2 s_i + bt_i s_i^2 + cs_i^3$$

for $i = 1, 2$, we conclude that

$$s_2 - s_1 = (t_2^3 - t_1^3) + a(t_2^2 s_2 - t_1^2 s_1) + b(t_2 s_2^2 - t_1 s_1^2) + c(s_2^3 - s_1^3).$$

Using that

$$t_2^2 s_2 - t_1^2 s_1 = (t_2^2 - t_1^2)s_2 + t_1^2(s_2 - s_1)$$

and

$$t_2 s_2^2 - t_1 s_1^2 = (t_2 - t_1)s_2^2 + t_1(s_2^2 - s_1^2),$$

we see that

$$(s_2 - s_1)\left(1 - at_1^2 - bt_1(s_1 + s_2) - c(s_1^2 + s_1 s_2 + s_1^2)\right) = (t_2 - t_1)\left((t_1^2 + t_1 t_2 + t_2^2) + as_2(t_1 + t_2) + s_2^2\right),$$

giving us an expression for λ of the form

$$\lambda = \frac{s_2 - s_1}{t_2 - t_1} = \frac{(t_1^2 + t_1 t_2 + t_2^2) + as_2(t_1 + t_2) + bs_2^2}{1 - at_1^2 - bt_1(s_1 + s_2) - c(s_1^2 + s_1 s_2 + s_2^2)}. \tag{1.14.16}$$

Wait–this expression is only valid if the denominator on the right hand side of this formula is nonzero. Is it? Well, except for the 1, every single quantity has positive p-adic valuation. This forces v_p of the denominator on the right to be 0; in particular, it forces the denominator to be nonzero. Finally, what happens if $t_1 = t_2$? Well, the real formula for λ should then involve derivatives, and this approach gives us

$$\lambda = \frac{3t_1^2 + 2at_1 s_2 + bs_1^2}{1 - at_1^2 - 2bt_1 s_1 - 3cs_1^2},$$

but this is of course the same as 1.14.16 when $(t_1, s_1) = (t_2, s_2)$. (Indeed, this is how people took derivatives in the olden days before calculus!) Thus, 1.14.16 is correct in all cases. Now since $P_1, P_2 \in E_{p,r}(\mathbb{Q})$, we conclude from 1.14.16 that $v_p(\lambda) \geq 2r$. From this and $\nu = s_1 - \lambda t_1$, we get that $v_p(\nu) \geq 3r$.

Let's now look at how the three t_i relate to each other. Of course, these t_i are the roots of the cubic equation

$$\lambda t + \nu = t^3 + at^2(\lambda t + \nu) + bt(\lambda t + \nu)^2 + c(\lambda t + \nu)^3,$$

which can be rearranged to be in the form

$$(1 + a\lambda + b\lambda^2 + c\lambda^3)\, t^3 + \nu(a + 2b\lambda + 3c\lambda^2)t^2 + (\cdots)\, t + (\cdots) = 0,$$

where the $(\cdots)$ is stand-in for some coefficient we don't care about. Arguing as above, we note that $v_p(1 + a\lambda + b\lambda^2 + c\lambda^3) = 0$; in particular, this leading coefficient is nonzero. This implies by Vieta's formulae that this equation *does* genuinely have three roots t_i which are related by

$$t_1 + t_2 + t_3 = -\frac{\nu(a + 2b\lambda + 3c\lambda^2)}{1 + a\lambda + b\lambda^2 + c\lambda^3}. \tag{1.14.17}$$

This equation (1.14.17) is key. Note that since the denominator on the right side has $v_p = 0$, we conclude that the right hand side has p-adic valuation at least $v_p(\nu) \geq 3r \geq r$. Since t_1 and t_2 have p-adic valuation at least r as well, it follows from 1.14.17 that so does t_3, finishing the proof of this case.

It remains to discuss what happens when some of the P_i don't lie on the affine patch $Y \neq 0$, i.e., are nontrivial 2-torsion points. Note that the condition $P_1, P_2 \in E_{p,r}(\mathbb{Q})$ implies that $y(P_i) \neq 0$ for $i = 1, 2$ (see above), so that P_1, P_2 are not 2-torsion. Therefore, the only possibility we need to worry about is when P_3 is 2-torsion. In general, it is possible for two points which are *not* nontrivial 2-torsion points to sum to one (e.g., if $P_1 = P_2$ has order 4); however, we actually see from the above proof that this *cannot* happen if $P_1, P_2 \in E_{p,r}(\mathbb{Q})$, because of the equation 1.14.17. Indeed, the only way P_3 could be a two-torsion point is if its (t, s) coordinates blew-up (i.e., were "infinite"), and this would require us to have $1 + a\lambda + b\lambda^2 + c\lambda^2 = 0$ (check!). Therefore, the fact that this quantity is nonzero and the equation 1.14.17 is *sufficient* to conclude that P_3 is not a two-torsion point either, and the above proof applies. ■

Exercise 1.14.18. Convince yourself that the argument at the end of this proof works, and that we have actually proven what we claimed to.

At this point, we have proven 1.14.12. We will use this result to finish the proof of 1.14.2 and hence 1.13.7 next time.