

ARITHMETIC PROGRESSIONS OF FOUR SQUARES

KEITH CONRAD

1. INTRODUCTION

Suppose a , b , c , and d are rational numbers such that a^2 , b^2 , c^2 , and d^2 form an arithmetic progression: the differences $b^2 - a^2$, $c^2 - b^2$, and $d^2 - c^2$ are equal. One possibility is that the arithmetic progression is constant: a^2, a^2, a^2, a^2 . Are there arithmetic progressions of four rational squares which are not constant? This question was first raised by Fermat in 1640. There are no such progressions with small rational squares, but that doesn't preclude the possibility of a rational solution altogether. After all, the smallest positive integer solution to $x^2 - 61y^2 = 1$ is $(x, y) = (1766319049, 226153980)$.

We will show how to turn 4-tuples of numbers (not all 0) whose squares form an arithmetic progression into points on the elliptic curve

$$(1.1) \quad y^2 = (x - 1)(x^2 - 4).$$

A brief check reveals 8 rational points of the homogenized projective curve: $[0, 1, 0]$ and

$$(1.2) \quad (1, 0), (2, 0), (-2, 0), (0, 2), (0, -2), (4, 6), (4, -6).$$

We will see that whether or not there is a nonconstant 4-term arithmetic progression of rational squares is equivalent to the existence of an additional rational point on this curve.

To turn squares in arithmetic progression into points on (1.1), the main idea is to let geometry dictate the right changes of variables. This is carried out in Section 2. In Section 3 we will apply theorems about elliptic curves over \mathbf{Q} to find all rational points on (1.1) and thus all arithmetic progressions of rational squares.

2. GEOMETRIC CONSIDERATIONS

To say a^2, b^2, c^2 , and d^2 are in an arithmetic progression means $b^2 - a^2 = c^2 - b^2$ and $c^2 - b^2 = d^2 - c^2$. We write these conditions as

$$(2.1) \quad a^2 + c^2 = 2b^2, \quad b^2 + d^2 = 2c^2.$$

These equations are homogeneous, so we will consider a , b , c , and d as a point $[a, b, c, d]$ in $\mathbf{P}^3(\mathbf{R})$.

Each equation in (2.1), considered separately, cuts out a surface in $\mathbf{P}^3(\mathbf{R})$. The eight points $[\pm 1, \pm 1, \pm 1, 1]$, with independent choices of sign, lie on both surfaces. (There are not 16 points if we allow an independent sign in

the last coordinate, since we are working projectively, *e.g.*, $[1, -1, 1, -1] = [-1, 1, -1, 1]$.) Finding common solutions to both equations in (2.1) means looking at the intersection of the two surfaces, which will be a curve. Call it C . Rational points on C are what we are interested in, since they will tell us the 4-tuples of rational numbers whose squares are in arithmetic progression.

To find an equation for C itself, so we can study its rational points, we will project C into the plane $\{[a, b, c, 0]\}$ and work in this plane. We of course need to be sure that our projection is one-to-one on C so no information is lost. For instance, projecting C to the plane $\{[a, b, c, 0]\}$ in the simple-minded way by $[a, b, c, d] \mapsto [a, b, c, 0]$, which is projection from the point $[0, 0, 0, 1]$, will be 2-to-1 on C since $[a, b, c, \pm d]$ both go to the same point. That's no good! We will be more careful by projecting into the plane from a point already on C ; lines through the point will turn out to meet C in at most one other point. Different points on C will have different projections into the plane.

Set

$$P := [1, 1, 1, 1], \quad \Pi := \{[a, b, c, 0]\} \subset \mathbf{P}^3(\mathbf{R}).$$

Let $f: \mathbf{P}^3(\mathbf{R}) - P \rightarrow \Pi$ by $f(Q) = \overline{PQ} \cap \Pi$. So $f(Q)$ is the point on the line \overline{PQ} which lies in the plane Π . To find an explicit formula for $f(Q)$, write $Q = [a, b, c, d]$. The line \overline{PQ} is

$$\overline{PQ} = \{[\lambda + \mu a, \lambda + \mu b, \lambda + \mu c, \lambda + \mu d] : [\lambda, \mu] \in \mathbf{P}^1(\mathbf{R})\}.$$

This line meets Π when $\lambda + \mu d = 0$, so $\lambda = -\mu d$. Thus

$$f(Q) = [\mu(a - d), \mu(b - d), \mu(c - d), 0] = [a - d, b - d, c - d, 0].$$

We are interested in f not on all of $\mathbf{P}^3(\mathbf{R}) - P$, but specifically on C , which includes P too. What should $f(P)$ mean? There is no line \overline{PP} , but we will use the *tangent line* to C at P . What is this line and where does it meet Π ? The tangent planes of the two surfaces in (2.1) at P are given by

$$a + c = 2b, \quad b + d = 2c.$$

These planes overlap in $\{[a, b, -a + 2b, -2a + 3b] : [a, b] \in \mathbf{P}^1(\mathbf{R})\}$, so this is the tangent line to C at P . This line meets Π where $-2a + 3b = 0$, so the intersection point is $[a, (2/3)a, (1/3)a, 0] = [3, 2, 1, 0]$. Thus, we define $f: C \rightarrow \Pi$ by

$$(2.2) \quad f([a, b, c, d]) = \begin{cases} [a - d, b - d, c - d, 0], & \text{if } [a, b, c, d] \neq [1, 1, 1, 1], \\ [3, 2, 1, 0], & \text{if } [a, b, c, d] = [1, 1, 1, 1]. \end{cases}$$

Table 1 gives the projection to Π of the 8 obvious rational points on C . Note the coordinates of $f(Q)$ are only determined up to an overall scaling. For instance, (2.2) says $f([-1, 1, 1, 1]) = [-2, 0, 0, 0]$, which appears in Table 1 as $[1, 0, 0, 0]$.

Remark 2.1. The formula for f is *not* discontinuous at P . Indeed, let's pick a sequence of points on C tending to P and see their f -values tend to $[3, 2, 1, 0]$. To keep the algebra simple, for any ε let P_ε be a point on C with

Q	$f(Q)$
$[1, 1, 1, 1]$	$[3, 2, 1, 0]$
$[-1, 1, 1, 1]$	$[1, 0, 0, 0]$
$[1, -1, 1, 1]$	$[0, 1, 0, 0]$
$[1, 1, -1, 1]$	$[0, 0, 1, 0]$
$[-1, -1, 1, 1]$	$[1, 1, 0, 0]$
$[1, -1, -1, 1]$	$[0, 1, 1, 0]$
$[-1, 1, -1, 1]$	$[1, 0, 1, 0]$
$[-1, -1, -1, 1]$	$[1, 1, 1, 0]$

TABLE 1

coordinates $d = 1$ and $c = 1 + \varepsilon$. The coordinates a and b are determined (up to sign) by the equations in (2.1). Choosing positive square roots, we take

$$P_\varepsilon = [\sqrt{1 + 6\varepsilon + 3\varepsilon^2}, \sqrt{1 + 4\varepsilon + 2\varepsilon^2}, 1 + \varepsilon, 1].$$

Then $P_0 = P$ and for $\varepsilon \neq 0$, (2.2) says

$$f(P_\varepsilon) = [\sqrt{1 + 6\varepsilon + 3\varepsilon^2} - 1, \sqrt{1 + 4\varepsilon + 2\varepsilon^2} - 1, \varepsilon, 0].$$

To understand the behavior of $f(P_\varepsilon)$ as $\varepsilon \rightarrow 0$ (the limit is *not* $[0, 0, 0, 0]!$), scale the third coordinate to 1:

$$f(P_\varepsilon) = \left[\frac{\sqrt{1 + 6\varepsilon + 3\varepsilon^2} - 1}{\varepsilon}, \frac{\sqrt{1 + 4\varepsilon + 2\varepsilon^2} - 1}{\varepsilon}, 1, 0 \right].$$

Letting $\varepsilon \rightarrow 0$, a derivative calculation shows the limit is $[3, 2, 1, 0]$.

We want to find an equation for $f(C)$ in the plane Π . Considering the formula for f in (2.2) away from P , when $[a, b, c, d] \neq [1, 1, 1, 1]$ set

$$(2.3) \quad u = a - d, \quad v = b - d, \quad w = c - d,$$

so $a = u + d$, $b = v + d$, and $c = w + d$. Using this, (2.1) becomes

$$(u + d)^2 + (w + d)^2 = 2(v + d)^2, \quad (v + d)^2 + d^2 = 2(v + d)^2.$$

Expanding the squares and collecting like terms,

$$(2.4) \quad u^2 - 2v^2 + w^2 = -2d(u - 2v + w), \quad v^2 - 2w^2 = -2d(v - 2w).$$

We can eliminate d by multiplying the first equation by $v - 2w$, the second equation by $u - 2v + w$, and then equating the left sides:

$$(v - 2w)(u^2 - 2v^2 + w^2) = (v^2 - 2w^2)(u - 2v + w).$$

After multiplying out both sides and moving everything to one side, this equation becomes

$$(2.5) \quad uv^2 - u^2v + 2u^2w - 2uw^2 - 3v^2w + 3vw^2 = 0.$$

We should check that the elimination of d is reversible: is d determined by u , v , and w in (2.4)? Yes, provided either $u - 2v + w$ or $v - 2w$ is nonzero.

Could $u - 2v + w$ and $v - 2w$ both vanish? In terms of a, b, c , and d , the equation $u - 2v + w = 0$ says $c = -a + 2b$. Feeding this into $a^2 + c^2 = 2b^2$, we get $a^2 - 2ab + b^2 = 0$, so $(a - b)^2 = 0$ and thus $a = b$ and $c = -a + 2b = b$. In a similar way, the condition $v - 2w = 0$ says $d = -b + 2c$, so $d = -a + 2a = a$. Therefore a point $[a, b, c, d]$ where $u - 2v + w$ and $v - 2w$ both vanish must be $[a, a, a, a] = [1, 1, 1, 1] = P$. As long as we are looking at projections of points $Q = [a, b, c, d]$ on C other than P , d is determined by $f(Q) = [u, v, w]$. If $Q \neq P$ then $f(Q) \neq [3, 2, 1]$, and $[3, 2, 1]$ satisfies (2.5), so $f(C)$ is the set of points $[u, v, w, 0]$ in $\mathbf{P}^3(\mathbf{R})$ satisfying (2.5).

All points in $f(C)$ have final coordinate 0, so we may as well just ignore this last coordinate: identify Π with $\mathbf{P}^2(\mathbf{R})$ by $[u, v, w, 0] \leftrightarrow [u, v, w]$. The equation (2.5) defines a smooth cubic curve in $\mathbf{P}^2(\mathbf{R})$, and it does have rational points on it: the second column of Table 1 provides 8 of them when we drop the last coordinate 0 from those points.

In order to write (2.5) in the (affine) form $y^2 = \text{cubic in } x$, we need to find a point on (2.5) whose tangent line to the curve passes through no other point on the curve. Then a linear change of variables $[u, v, w] \mapsto [x, y, z]$ that moves that point to $[0, 1, 0]$ and moves its tangent line to the line $z = 0$ in $\mathbf{P}^2(\mathbf{R})$ will give us a simpler equation for (2.5). But watch out! Even though the point $[0, 1, 0]$ is already on the curve (2.5), its tangent line is $u = 3w$, which meets the curve in another point, $[3, 2, 1]$. So a linear change of variables $[u, v, w] \mapsto [x, y, z]$ that fixes $[0, 1, 0]$ and makes its tangent line $z = 0$ will move $[3, 2, 1]$ to this line, which means the new equation for the curve has two solutions on the line $z = 0$, so the curve won't have the (affine) equation $y^2 = \text{cubic in } x$.

In Table 2 are the tangent lines to (2.5) at all 8 points we already know.

$[u, v, w]$	Tangent Line
$[3, 2, 1]$	$u - 3v + 3w = 0$
$[1, 0, 0]$	$v - 2w = 0$
$[0, 1, 0]$	$u - 3w = 0$
$[0, 0, 1]$	$2u - 3w = 0$
$[1, 1, 0]$	$u - v + w = 0$
$[0, 1, 1]$	$u + 3v - 3w = 0$
$[1, 0, 1]$	$u + v - w = 0$
$[1, 1, 1]$	$u - 2v + 3w = 0$

TABLE 2

An inspection shows the tangents at $[3, 2, 1]$, $[1, 1, 0]$, and $[1, 0, 1]$ pass through $[0, 1, 1]$ and the tangents at $[1, 0, 0]$, $[0, 1, 0]$, $[0, 0, 1]$, and $[1, 1, 1]$ pass through $[3, 2, 1]$. Among the 8 points in Table 2, only $[0, 1, 1]$ has a tangent line which contains no other point on (2.5). (Indeed, if you set $u = 3w - 3v$ in (2.5) the equation simplifies to $12(v - w)^3 = 0$, so $v = w$ and thus $[u, v, w] = [0, v, v] = [0, 1, 1]$.) A linear change of variables $[u, v, w] \mapsto$

$[X, Y, Z]$ that turns $[0, 1, 1]$ into $[0, 1, 0]$ and the line $u + 3v - 3w = 0$ into the line $Z = 0$ is

$$(2.6) \quad X = u, \quad Y = v, \quad Z = u + 3v - 3w.$$

The new coordinates of our 8 points are in Table 3. (For instance, when $[u, v, w] = [0, 0, 1]$ we get $[X, Y, Z] = [0, 0, -3] = [0, 0, 1]$.)

$[u, v, w]$	$[X, Y, Z]$
[3, 2, 1]	[3, 2, 6]
[1, 0, 0]	[1, 0, 1]
[0, 1, 0]	[0, 1, 3]
[0, 0, 1]	[0, 0, 1]
[1, 1, 0]	[1, 1, 4]
[0, 1, 1]	[0, 1, 0]
[1, 0, 1]	[1, 0, -2]
[1, 1, 1]	[1, 1, 1]

TABLE 3

Inverting (2.6),

$$u = X, \quad v = Y, \quad w = \frac{1}{3}(X + 3Y - Z),$$

and substitution of this into (2.5) turns the equation into

$$(2.7) \quad 9Y^2Z - 3YZ^2 - 6XYZ - 4X^3 + 2X^2Z + 2XZ^2 = 0.$$

Remark 2.2. The tangent line to the original curve C at $P = [1, 1, 1, 1]$ meets C at no point other than P itself, but after projection to the plane Π , the tangent line to $f(C)$ at $f(P)$ (in Π) meets $f(C)$ at a second point. It is not really a surprise that projecting into a lower-dimensional space can introduce extra intersections, but in fact there is something more going on: the projection of the tangent line to C at P under f is *not* the tangent line to $f(C)$ at $f(P)$. Indeed, the whole tangent line to C at P gets projected onto the single point $f(P)$. (Similarly, each line through P – not just the tangent line to C – gets projected from P to a single point in Π .) For each $Q \neq P$ in Table 1, the relation between projection and tangency is better: the projection f of the tangent line to C at Q is the tangent line to $f(C)$ at $f(Q)$. For example, the tangent line to C at $[-1, 1, 1, 1]$ is $\{[a, b, a + 2b, 2a + 3b] : [a, b] \in \mathbf{P}^1(\mathbf{R})\}$ and the image of this line under f is the line $v = 2w$, which is listed in Table 2 as the tangent line to (2.5) at $f([-1, 1, 1, 1]) = [1, 0, 0]$ (dropping the fourth coordinate 0).

We want to massage (2.7) further so Y only occurs once in the equation. At this point the geometry ends and grinding algebra takes over. To make the algebra easier to follow, let's pass to the affine form of (2.7) with $Z = 1$:

$$9Y^2 - 3Y - 6XY - 4X^3 + 2X^2 + 2X = 0.$$

Put the Y -free terms on the right:

$$9Y^2 - 3Y(1 + 2X) = 4X^3 - 2X^2 - 2X.$$

Complete the square on the left:

$$\left(3Y - \frac{1 + 2X}{2}\right)^2 - \left(\frac{1 + 2X}{2}\right)^2 = 4X^3 - 2X^2 - 2X.$$

Bring the second term on the left over to the other side:

$$\left(3Y - \frac{1 + 2X}{2}\right)^2 = 4X^3 - X^2 - X + \frac{1}{4}$$

Multiply through by 4 to clear the denominator:

$$(6Y - 2X - 1)^2 = 16X^3 - 4X^2 - 4X + 1.$$

Multiply by 4 (again) to absorb the power of 2 everywhere:

$$(12Y - 4X - 2)^2 = (4X)^3 - (4X)^2 - 4(4X) + 4.$$

In homogeneous form, this is

$$(12Y - 4X - 2Z)^2 Z = (4X)^3 - (4X)^2 Z - 4(4X)Z^2 + 4Z^3.$$

This equation tells us to make a final change of variables:

$$(2.8) \quad x = 4X, \quad y = 12Y - 4X - 2Z, \quad z = Z.$$

The equation of the curve is

$$y^2 z = x^3 - x^2 z - 4xz^2 + 4z^3,$$

or

$$(2.9) \quad y^2 = x^3 - x^2 - 4x + 4$$

in affine form. (The cubic polynomial in x factors as $(x - 1)(x^2 - 4)$, thus recovering (1.1).) Table 4 records the final set of coordinates of the 8 points. The 7 points in the last column of the table other than $[0, 1, 0]$ are precisely the points listed in (1.2).

$[a, b, c, d]$	$[u, v, w]$	$[X, Y, Z]$	$[x, y, z]$
$[1, 1, 1, 1]$	$[3, 2, 1]$	$[3, 2, 6]$	$[2, 0, 1]$
$[-1, 1, 1, 1]$	$[1, 0, 0]$	$[1, 0, 1]$	$[4, -6, 1]$
$[1, -1, 1, 1]$	$[0, 1, 0]$	$[0, 1, 3]$	$[0, 2, 1]$
$[1, 1, -1, 1]$	$[0, 0, 1]$	$[0, 0, 1]$	$[0, -2, 1]$
$[-1, -1, 1, 1]$	$[1, 1, 0]$	$[1, 1, 4]$	$[1, 0, 1]$
$[1, -1, -1, 1]$	$[0, 1, 1]$	$[0, 1, 0]$	$[0, 1, 0]$
$[-1, 1, -1, 1]$	$[1, 0, 1]$	$[1, 0, -2]$	$[-2, 0, 1]$
$[-1, -1, -1, 1]$	$[1, 1, 1]$	$[1, 1, 1]$	$[4, 6, 1]$

TABLE 4

We have proved the following theorem.

Theorem 2.3. *The 4-tuples $[a, b, c, d] \in \mathbf{P}^3(\mathbf{R})$ such that a^2, b^2, c^2, d^2 form an arithmetic progression are parametrized by the points on the elliptic curve*

$$E: y^2 = x^3 - x^2 - 4x + 4.$$

A formula for $[x, y, z]$ in terms of $[a, b, c, d]$ is obtained by composing the changes of variables (2.3), (2.6), and (2.8): for any $[a, b, c, d] \neq [1, 1, 1, 1]$,

$$(2.10) \quad [x, y, z] = [4(a - d), -6(a - b - c + d), a + 3b - 3c - d],$$

while for $[a, b, c, d] = [1, 1, 1, 1]$, $[x, y, z] = [2, 0, 1]$. To invert this overall change of variables, invert (2.3), (2.6), and (2.8), and then solve for d in terms of u, v , and w . Once we find d ,

$$(2.11) \quad a = d + \frac{x}{4}, \quad b = d + \frac{x + y + 2z}{12}, \quad c = d + \frac{2x + y + 2z}{12}.$$

Looking at the first and last columns of Table 4, the eight rational points in the last column of Table 4 correspond to 4-tuples whose squares form a constant arithmetic progression (common difference 0), allowing for sign changes in the four numbers before they are squared. To find a nonconstant arithmetic progression of rational squares is therefore equivalent to finding another rational point on E .

3. RATIONAL POINTS ON E

The eight rational points we identified already in $E(\mathbf{Q})$ are all torsion points. More precisely, these points form a group of size 8 generated by $(0, 2)$ (of order 4) and $(1, 0)$ (of order 2). Table 5 expresses each of the seven non-identity points in terms of the chosen generators.

P	$2P$	$3P$	Q	$P + Q$	$2P + Q$	$3P + Q$
$(0, 2)$	$(2, 0)$	$(0, -2)$	$(1, 0)$	$(4, 6)$	$(-2, 0)$	$(4, -6)$

TABLE 5. Torsion on $E(\mathbf{Q})$

Theorem 3.1. *The eight points found already in $E(\mathbf{Q})$ form its full torsion subgroup.*

Proof. We give two proofs, first using Nagell-Lutz and then using reduction mod p .

To use the Nagell-Lutz theorem, we rewrite the Weierstrass equation (2.9) for E in the form $y'^2 = x'^3 + Ax' + B$. Taking $x = (x' + 3)/9$ and $y = y'/27$, we obtain from (2.9) the new equation

$$\begin{aligned} y'^2 &= x'^3 - 351x' + 1890 \\ &= (x' - 6)(x' - 15)(x' + 21). \end{aligned}$$

Here $4A^3 + 27B^2 = -2^4 \cdot 3^{14}$, so if (x', y') is a rational torsion point then x' and y' are integers with $y' = 0$ or $y' | 2^2 \cdot 3^7$. When $y' = 0$ we have $x' = 6, 15$, or -21 . When y' runs through the 24 positive factors of $2^2 \cdot 3^7$, only $y' = 54$

and $y' = 162$ produce a corresponding integral x' . The results, and the conversion back to x, y coordinates, are in Table 6. We recover the known torsion points and nothing further.

(x', y')	(x, y)
(6, 0)	(1, 0)
(15, 0)	(2, 0)
(-21, 0)	(-2, 0)
(-3, 54)	(0, 2)
(-3, -54)	(0, -2)
(33, 162)	(4, 6)
(33, -162)	(4, -6)

TABLE 6. Checking Nagell-Lutz

For the second proof of the theorem, consider reduction $E(\mathbf{Q}) \rightarrow E(\mathbf{F}_p)$. The discriminant of the cubic is divisible only by the primes 2 and 3, so reduction is injective on the torsion subgroup of $E(\mathbf{Q})$ when $p > 3$. A calculation shows $E(\mathbf{F}_5)$ has size 8, and we already have 8 torsion points, so the points we found in $E(\mathbf{Q})$ are its full torsion subgroup. \square

Remark 3.2. The Nagell-Lutz theorem is true for Weierstrass equations $y^2 = x^3 + ax^2 + bx + c$ where the coefficients are all integers (a need not be 0). A torsion point (x, y) has integer coordinates and $y = 0$ or $y^2 | \Delta$, where

$$\Delta = 4b^3 + 27c^2 - a^2b^2 + 4a^3c - 18abc.$$

(When $a = 0$ this reduces to the usual formula for the cubic discriminant.) The original Weierstrass equation (2.9) has this form: $a = -1$, $b = -4$, and $c = 4$, giving $\Delta = 144$, so a torsion point (x, y) has $y = 0$ or $y | 12$. Running through the factors of 12 as values for y , we get integral x when $y = \pm 2$ and $y = \pm 6$. The corresponding points on E are $(0, \pm 2)$, and $(4, \pm 6)$, which are torsion. Allowing $y = 0$ leads to the rest of the torsion points in $E(\mathbf{Q})$.

Corollary 3.3. *If there is a nonconstant 4-term arithmetic progression of rational squares then there are infinitely many which are not scalar multiples of each other.*

Proof. A nonconstant 4-term arithmetic progression of rational squares leads to a point in $E(\mathbf{Q})$ besides the 8 known points, which must have infinite order, and we get infinitely many progressions from those infinitely many rational points. \square

Theorem 3.4. *A nonconstant 4-term arithmetic progression of rational squares does not exist.*

Proof. We show $E(\mathbf{Q})$ is finite. This can be done by a descent argument, and that is how Euler proceeded when he solved this problem in 1780. (He

used descent on the pair of equations (2.1), not on an elliptic curve.) We will instead cite Kolyvagin's theorem: if the L -function of an elliptic curve over \mathbf{Q} is nonzero at $s = 1$, then the elliptic curve has finitely many rational points. For our particular elliptic curve E , PARI says $L(E, 1) \approx .53912$, so if we believe this is even approximately correct then $L(E, 1) \neq 0$, so $E(\mathbf{Q})$ is finite. \square

While nonconstant 4-term arithmetic progressions of rational squares don't exist, there are such progressions in other fields, and we can find them using Theorem 2.3. Indeed, the proof of Theorem 2.3 applies not just to rational numbers, but to four elements of any field F (not of characteristic 2 or 3)¹ whose squares are in arithmetic progression, giving a bijection between such 4-tuples in $\mathbf{P}^3(F)$ and points in $E(F)$.

Example 3.5. Consider the arithmetic progression 0, 1, 2, 3. This is not an arithmetic progression of rational squares, but it is an arithmetic progression of squares if it is viewed as $0^2, 1^2, \sqrt{2}^2, \sqrt{3}^2$. Starting with $[a, b, c, d] = [0, 1, \sqrt{2}, \sqrt{3}]$ on C , we obtain by (2.10) the point

$$[x, y, z] = [-4\sqrt{3}, 6 + 6\sqrt{2} - 6\sqrt{3}, 3 - 3\sqrt{2} - \sqrt{3}]$$

on E . In affine form, this is $[x/z, y/z, 1]$, where

$$\frac{x}{z} = 4 + 3\sqrt{2} - 2\sqrt{3} - \sqrt{6}, \quad \frac{y}{z} = 12 + 9\sqrt{2} - 8\sqrt{3} - 5\sqrt{6}.$$

Example 3.6. Let's go the other way, from a point on E to an arithmetic progression of 4 squares. The point $(x, y) = (3, \sqrt{10})$ lies on E so it must correspond to an arithmetic progression of squares in $\mathbf{Q}(\sqrt{10})$. Setting $x = 3$, $y = \sqrt{10}$, and $z = 1$, we run all the changes of variables in reverse:

$$X = \frac{3}{4}, \quad Y = \frac{5 + \sqrt{10}}{12}, \quad Z = 1$$

and

$$u = \frac{3}{4}, \quad v = \frac{5 + \sqrt{10}}{12}, \quad w = \frac{4 + \sqrt{10}}{12}.$$

Then $u = a - d$, $v = b - d$, and $w = c - d$, where d is determined by either of the equations in (2.4). From (2.4) and some algebra, $d = (-9 + \sqrt{10})/24$. Feeding this into (2.11),

$$a = \frac{9 + \sqrt{10}}{24}, \quad b = \frac{1 + 3\sqrt{10}}{24}, \quad c = \frac{-1 + 3\sqrt{10}}{24}, \quad d = \frac{-9 + \sqrt{10}}{24}.$$

The numbers $a^2 > b^2 > c^2 > d^2$ are a decreasing arithmetic progression with common difference $\sqrt{10}/48$.

The point $(3, \sqrt{10})$ has infinite order on E , since for instance its double is $(89/40, 273\sqrt{10}/800)$, which doesn't have coordinates in $\mathbf{Z}[\sqrt{10}]$, so by an analogue of the Nagell-Lutz theorem for elliptic curves over $\mathbf{Q}(\sqrt{10})$ the

¹Inverting the changes of variables from $[a, b, c, d]$ to $[x, y, z]$ involves division by 2 and 3, so everything works as long as 2 and 3 are not 0 in the field.

double has infinite order and thus the original point has infinite order as well. Since the group $E(\mathbf{Q}(\sqrt{10}))$ is infinite, there are infinitely many arithmetic progressions of 4 squares in $\mathbf{Q}(\sqrt{10})$ which are not scalar multiples of each other.

Example 3.7. The group $E(\mathbf{F}_{43})$ contains the point $(3, 15)$. Performing the changes of variables in reverse modulo 43 starting from $[x, y, z] = [3, 15, 1]$ produces the point $[a, b, c, d] = [1, 27, 9, 11]$ in $\mathbf{P}^3(\mathbf{F}_{43})$, which squares modulo 43 to $[1, 41, 38, 35]$. These coordinates are an arithmetic progression modulo 43.

Remark 3.8. Modulo 43, $15^2 = 10$, so we can think of $15 \bmod 43$ as a square root of 10. So working with $(3, 15)$ in $E(\mathbf{F}_{43})$ is like working with $(3, \sqrt{10})$ in $E(\mathbf{Q}(\sqrt{10}))$. If you replace $\sqrt{10}$ by 15 everywhere in the formulas for a, b, c , and d in Example 3.6 and then reduce the resulting fractions modulo 43, you will reproduce the result of Example 3.7.

When $p \geq 5$ is prime, the 8 points we have found in $E(\mathbf{Q})$ stay distinct in $E(\mathbf{F}_p)$, so there is a nonconstant arithmetic progression of 4 squares in \mathbf{F}_p provided $\#E(\mathbf{F}_p) > 8$. The Hasse bound $|\#E(\mathbf{F}_p) - (p + 1)| \leq 2\sqrt{p}$ tells us $\#E(\mathbf{F}_p) > 8$ as long as $p + 1 - 2\sqrt{p} > 8$, which is the same as $(\sqrt{p} - 1)^2 > 8$. This holds for $p \geq 17$, so in \mathbf{F}_p there is a nonconstant arithmetic progression of 4 squares when $p \geq 17$. An explicit check reveals such a progression in \mathbf{F}_{13} : $10, 12, 1, 3$. There are no nonconstant 4-term arithmetic progressions of squares in \mathbf{F}_p for $p \leq 11$ by an examination of squares in \mathbf{F}_p for small p .

APPENDIX A. GEOMETRIC INTERPRETATION OF SIGN CHANGES

When $[a, b, c, d]$ is a point on C (the common solutions to (2.1)), all the points $[\pm a, \pm b, \pm c, \pm d]$ lie on C . It is interesting to ask what these sign change operations mean in terms of the group law on C as an elliptic curve. Put more simply, if we transfer these operations over to E with its Weierstrass equation $y^2 = x^3 - x^2 - 4x + 4$, what do the operations look like? Since the coordinates are only defined up to scaling, we may focus on the sign changes $[\pm a, \pm b, \pm c, d]$, where d is fixed.

Theorem A.1. *Let $[a, b, c, d] \in \mathbf{P}^3$ satisfy (2.1) and correspond to the point $[x, y, z]$ on E . Then*

- (1) $[-a, b, c, d]$ corresponds to $-[x, y, z] + [4, 6, 1]$,
- (2) $[a, -b, c, d]$ corresponds to $-[x, y, z] + [0, -2, 1]$,
- (3) $[a, b, -c, d]$ corresponds to $-[x, y, z] + [0, 2, 1]$,
- (4) $[-a, -b, c, d]$ corresponds to $[x, y, z] + [-2, 0, 1]$,
- (5) $[a, -b, -c, d]$ corresponds to $[x, y, z] + [2, 0, 1]$,
- (6) $[-a, b, -c, d]$ corresponds to $[x, y, z] + [1, 0, 1]$,
- (7) $[-a, -b, -c, d]$ corresponds to $-[x, y, z] + [4, -6, 1]$.

Proof. We will give two proofs, one a concrete but tedious set of calculations and the other using general theorems about elliptic curves (and no hard computations at all).

From the sign change operations in (5), (6), and (7) we can get the other sign change operations by composition: (1) is (7) followed by (5), (2) is (7) followed by (6), (4) is (5) followed by (6), and (3) is (7) followed by (4). A case-by-case check shows the corresponding operations on the elliptic curve compose in the same way as the sign changes. (*E.g.*, the sign change in (1) is (7) followed by (5), while the elliptic curve operation in (1) is (7) followed by (5).) So it suffices to verify (5), (6) and (7).

Here are the formulas on E relevant to (5), (6), and (7):

- (a) $[x, y, z] + [2, 0, 1] = [2x(x - 2z), -4yz, (x - 2z)^2]$,
- (b) $[x, y, z] + [1, 0, 1] = [(x - 4z)(x - z), 3yz, (x - z)^2]$,
- (c) $-[x, y, z] + [4, -6, 1] = [(4x^2z + 4xz^2 - 8z^3 - 12yz^2)(x - 4z), -6x^3 - 60x^2z + 120xz^2 + 36xyz, (x - 4z)^3]$.

We see here that formulas for translation by a 2-torsion point are simpler than formulas for negation and translation by a point of order 4, so it's preferable to work with more formulas of the former kind than the latter kind. This is why it's better to check (5), (6), and (7) explicitly rather than (1), (2), and (3). (We need to deal with at least one operation that is not translation since translations never compose to give a non-translation.)

Equation (2.10) gives the coordinates of $[x, y, z]$ in terms of $[a, b, c, d]$ away from $[1, 1, 1, 1]$. Apply (2.10) to $[a, -b, -c, d]$ instead of to $[a, b, c, d]$ and check the resulting point on E is $[x, y, z] + [2, 0, 1]$. This is a tedious calculation left to the reader, and (2.1) is required for it. This establishes (5). Parts (6) and (7) are established similarly.

For the second proof of the theorem, we replace tedious calculations with ideas. Each of the seven sign changes is an automorphism of order 2 on \mathbf{P}^3 which preserves C . The induced automorphisms of C have order 1 or 2; in fact the order is 2 since $[1, 1, 1, 1]$ gets moved. The curve $C \cong E$ is an elliptic curve. What are the automorphisms of order 2 of an elliptic curve? The elliptic curve E doesn't have endomorphism ring $\mathbf{Z}[i]$ or $\mathbf{Z}[\zeta_3]$ since $j(E) \neq 0$ or 1728 (using the Weierstrass equation for E , $j(E) = 35152/9$), so the only automorphisms of E which fix the identity element are the identity automorphism and negation. Therefore all the automorphisms of E (not necessarily fixing the identity element) are translations and negation followed by translations. A translation $P \mapsto P + Q$ has order 2 only when Q is a point of order 2, while $P \mapsto -P + Q$ has order 2 for any Q .

Consider the sign change $[a, b, c, d] \mapsto [-a, b, c, d]$ on C . To figure out what it looks like on E , let's see where it sends the identity element. The point $[0, 1, 0]$ on E comes from $[1, -1, -1, 1]$ on C (see the first and last columns of Table 4), which is sent by the sign change to $[-1, -1, -1, 1]$. This corresponds to the point $[4, 6, 1]$ on E , so the sign change on C looks like either $[x, y, z] \mapsto [x, y, z] + [4, 6, 1]$ or $[x, y, z] \mapsto -[x, y, z] + [4, 6, 1]$ on E . Since the automorphism has order 2 and $[4, 6, 1]$ has order 4, the first option is impossible. Therefore $[a, b, c, d] \mapsto [-a, b, c, d]$ on C corresponds to $[x, y, z] \mapsto -[x, y, z] + [4, 6, 1]$ on E .

The sign changes $[a, b, c, d] \mapsto [a, -b, c, d]$ and $[a, b, c, d] \mapsto [a, b, -c, d]$ send $[1, -1, -1, 1]$ to $[1, 1, -1, 1]$ and $[1, -1, 1, 1]$, which correspond to $[0, -2, 1]$ and $[0, 2, 1]$. These points on E , like $[4, 6, 1]$, have order 4 and therefore the same analysis as above shows the corresponding automorphisms of E are $[x, y, z] \mapsto -[x, y, z] + [0, -2, 1]$ and $[x, y, z] \mapsto -[x, y, z] + [0, 2, 1]$. We have established (1), (2), and (3). The rest follow from these by composition. \square

It's amusing to note that while we wanted to avoid as much as possible the operations $P \mapsto -P + Q$ when dealing with explicit formulas in the first proof, it is precisely these which we used in the conceptual second proof because $P \mapsto P + Q$ and $P \mapsto -P + Q$ have different orders when Q is not a 2-torsion point.