

Counting points on elliptic curves: Hasse's theorem and recent developments

Igor Tolkov

June 3, 2009

Abstract

We introduce the the elliptic curve and the problem of counting the number of points on the curve when it is reduced modulo a prime. For any such curve that is nonsingular, Hasse's theorem provides a bound for the number of points. We prove Hasse's theorem and discuss more recent developments, namely, the exact formula for the number of points on a special class of curves (the ones with complex multiplication), as well as the Sato-Tate conjecture for the distribution of Hasse error terms on curves for which no exact formula exists.

Contents

1	Introduction	2
1.1	Acknowledgements	2
2	A certain cubic curve	3
3	Elliptic Curves and their Properties	5
3.1	Group structure	6
3.2	Singularity	8
3.3	Reduction modulo p	9
4	Endomorphisms and Hasse's Theorem	9
4.1	Isogenies and Isomorphisms	10
4.2	Another property of the degree	10
4.3	Hasse's Theorem	11
5	Beyond Hasse's Theorem	11
5.1	Elliptic Curves with Complex Multiplication	12
5.2	Elliptic Curves Without Complex Multiplication	12

1 Introduction

Number theory is a field full of seemingly simple problems that have as of today not been resolved. It is not known, for example, if any even integer greater than 2 can be expressed as a sum of two primes (Goldbach's conjecture), though it has been checked by computer for up to very large n . It is not known that there are infinitely many primes p such that $p+2$ is also prime (twin primes conjecture).

The open problems are not restricted to primes, however. An example is the $3n+1$ problem (Collatz conjecture) that deals with a recursion function defined on the natural numbers. The Sato-Tate conjecture is another, dealing with a question about the number of integral points on an elliptic curve.

Though it is still technically not fully resolved, the proof of the Sato-Tate conjecture is almost complete, with much recent development by Richard Taylor in 2006. Similarly, other questions involving elliptic curves have been answered such as the Taniyama-Shimura-Weil conjecture on modularity of elliptic curves. The methods used in both problems, such as Iwasawa theory and the development of the theory of Galois representations, are also recent.

At the same time, other questions about elliptic curves have been long known and are easily accessible, since they require only simple algebra tools. Among these is Hasse's theorem, a square-root-close estimate to the number of points on an elliptic curve, and exact formulas for the number of points on a certain class of elliptic curves.

In this paper, we will discuss both kinds of problems. We will attempt to give proofs of the easier kind of problems, including Hasse's theorem. We will also mention the recent developments, including the Sato-Tate conjecture, but will not attempt to give proofs to them, as their proofs are long and require very advanced theory.

In chapter 2, we will summarize this paper with an example, without proving anything. In chapters 3 and 4, we will develop the theory of elliptic curves and prove Hasse's theorem. In chapter 5, we will introduce complex multiplication and provide a narrative of the recent developments related to the problem of counting points on an elliptic curve.

1.1 Acknowledgements

I would like to thank my project advisor, Dr. Jim Morrow, for his feedback and patience during the second half of the quarter while my attention was split between class and this paper. Also, Dr. William Stein, my research advisor, for his time and explanations. Also Dr. McGovern for his explanation of field

extensions. I would also like to thank graduate students Chris Swierczewski and Tom Boothby for their help and suggestions of reading material.

All calculations and were performed using Sage: Open Source Mathematics Software (<http://www.sagemath.org>) on a server funded by the National Science Foundation under Grant No. 0555776.

2 A certain cubic curve

Consider the curve E defined by the following equation:

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

This is a cubic, whose graph is presented in Figure 1.

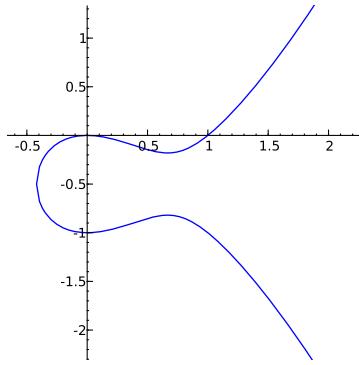


Figure 1: Implicit plot of E

Via a linear shift, $x \rightarrow x - 1/3$ and $y \rightarrow y + 1/2$, we transform the equation into

$$y^2 = x^3 - 1/3x + 19/108$$

But we will not be working much with this form.

Consider this curve over different characteristic fields. Characteristic 2 is not interesting, since $y^2 = 0$, $x^2 = 0$, and we are left with $y = 0$. This isn't even a cubic. Neither is the result over characteristic 3, since then $x^3 = 0$.

The first interesting prime field is characteristic 5. Any prime- p field is isomorphic to the field of integers modulo p , so a curve over a prime field can be represented as a collection of integral points on the grid $[0..p-1] \times [0..p-1]$.

It is easy to check that the ordered pairs

$$\{(0,0), (1,0), (0,-1), (1,-1)\}$$

satisfy the curve equation in any field, and are distinct, unless p is 2 or 3. There are no other points in \mathbb{F}_5 satisfying the equation.

For reasons to be discussed later, we add to the curve a "point at infinity". Call it ∞ and add it to the list of points over any field. Then, E has 5 points in \mathbb{F}_5 .

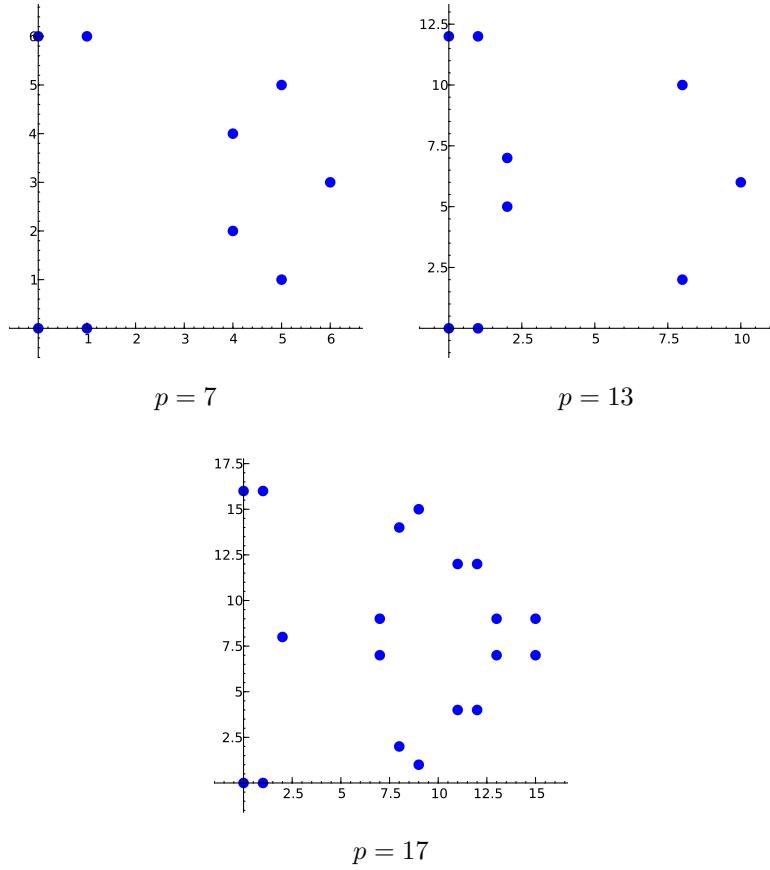


Figure 2: E over finite fields of characteristic 7, 13, and 17, respectively

E has 10 points in \mathbb{F}_7 , that is, the point at infinity, the 4 trivial points, and 5 other points. E has 10 points in \mathbb{F}_{13} and 20 points in \mathbb{F}_{17} . Plots of the curves over these fields are shown in Figure 2. We exclude $p = 11$ because the curve has "bad reduction" modulo 11 - this will be explained later.

How many points does E have over \mathbb{F}_p for arbitrary p ? In general, this is an open problem, but estimates can be made. For example, since the curve is

a cubic, no line can intersect it more than 3 times. Vertical lines can intersect the curve no more than twice, since for a fixed x , E is a quadratic in y . Thus the number of points on E over \mathbb{F}_p is certainly no greater than $2p + 1$, including the "point at infinity". A better bound was proven by Hasse in 1933:

Theorem 2.1. *The number of points on a non-singular cubic curve over the finite field \mathbb{F}_p is $p + 1 + \epsilon$ with $|\epsilon| < 2\sqrt{p}$.*

Hasse's theorem tells us that over \mathbb{F}_p , our curve has about p integer points plus the point at infinity. The error term, $e_p = N - p - 1$, is bounded in magnitude by $2\sqrt{p}$. In the case of E , Hasse's theorem seems to be satisfied, as can be seen in Figure 3.

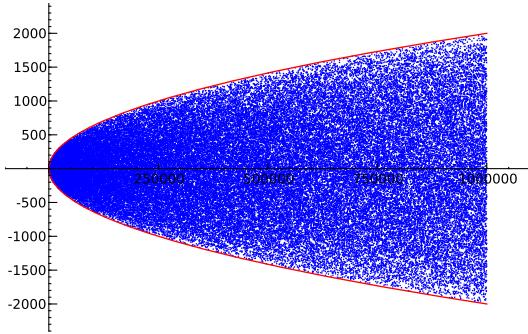


Figure 3: Blue: Hasse's error terms e_p vs. p for $p < 10^6$; red: Hasse's bound on e_p - the graph of $-2\sqrt{p} < e_p < 2\sqrt{p}$

The Sato-Tate conjecture is a statistical statement about the distribution of the error terms in Hasse's estimate. If the normed error term

$$a_p = \frac{e_p}{2\sqrt{p}}$$

is computed for different primes (excluding 2, 3, and 11), the error terms follow a semicircular distribution centered at 0 and with range $[-1, 1]$. For primes $p < 10^6$, the distribution is indeed roughly semicircular, shown in Figure 4.

The Sato-Tate conjecture was proved only 3 years ago for most, but not all elliptic curves. In the next few sections, we will examine in more detail the question about counting points on elliptic curves, including Hasse's theorem and the Sato-Tate conjecture.

3 Elliptic Curves and their Properties

An elliptic curve can be defined as follows:

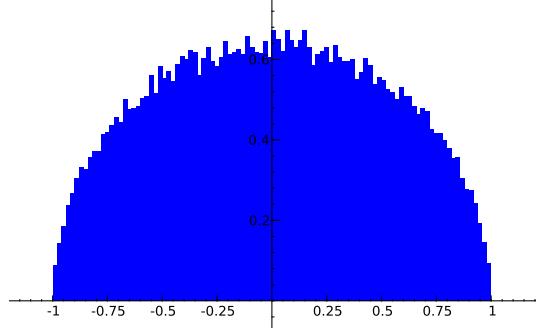


Figure 4: A histogram of a_p , $p < 10^6$.

Definition 3.1. An elliptic curve E over a field K of characteristic different from 2 or 3 is a curve that can be defined by the equation

$$y^2 = x^3 + ax + b$$

with a, b in K .

We shall concern ourselves only with elliptic curves over the rational numbers and their reduction to prime fields.

3.1 Group structure

Let E be an elliptic curve over \mathbb{Q} . Build a group on its set of rational points as follows:

1. Define the identity element to be the point at infinity.
2. For $A, B \in E$, define addition as follows: draw a line through A and B , and label its third point of intersection with the curve C . In the case of a tangent, count multiplicities. We allow $A = B$. We also allow any of the points to be the point at infinity; say that any vertical line intersects E at infinity. Then say that $(A + B) + C = \mathcal{O}$.

This allows us to calculate $A + B$, since $(A + B) + C + \mathcal{O} = \mathcal{O}$ corresponds to a line through $A + B$, C and infinity. By definition, this is a vertical line passing through C . Thus, $A + B$ is a reflection of C through the x-axis.

For this to make sense, we first need a theorem.

Theorem 3.1. *Let A, B be rational points on E . Allow $A = B$. Then the line passing through A and B intersects E in 3 rational points, counting multiplicity.*

Proof. Let E be in the form $y^2 = x^3 + ax + b$. If A and B lie on a vertical line, there is a third intersection at infinity (by definition) and we are done.

Otherwise, Let A have coordinates (a_x, a_y) and B have coordinates (b_x, b_y) . We determine the slope and intercept of the line in two cases:

Case 1: $A = B$. To find the slope of the tangent, differentiate.

$$2y \frac{dy}{dx} = 3x^2 + a$$

$$\lambda = \left. \frac{dy}{dx} \right|_{(a_x, a_y)} = \frac{3a_x^2 + a}{2a_y}$$

Case 2: $A \neq B$. Then

$$\lambda = \frac{b_y - a_y}{b_x - a_x}$$

The intercept of the line is

$$\beta = a_y - \lambda a_x$$

Then

$$y = \lambda x + \beta$$

Plug this into the elliptic curve equation.

$$(\lambda x + \beta)^2 = x^3 + ax + b$$

Which is a cubic in x :

$$x^3 - x^2 + (a - 2\lambda\beta)x + (b - \beta^2) = 0$$

This is a cubic, and as such, has three complex roots. However it has rational coefficients and two rational roots, counting multiplicity. Therefore, the third root must also be rational. \square

Theorem 3.2. *The set of rational points on an elliptic curve E with the addition operation defined as above, is an abelian group.*

Proof. We have already proven closure in Theorem 3.1. By construction, addition is commutative since the line through A, B is the same as the line through B, A . Take \mathcal{O} as the identity element. Then for any $A \in E$, it is clear that $-A$ is the third point of intersection, since then $A + (-A)$ intersects E at \mathcal{O} , whose reflection about the x-axis is also \mathcal{O} . Similarly, $\mathcal{O} + A$ is the reflection about the x-axis of $-A$, which is just A . This proves all properties of an abelian group except commutativity.

Proving associativity is tedious, and not relevant to the rest of this discussion. For those interested, a sketch can be found in [7]. For a more rigorous treatment, see [6]. \square

3.2 Singularity

In the preceding discussion, it was necessary that the implicit derivative be well-defined at every point. For $y \neq 0$, this is always the case since there the map $\pm y \leftrightarrow y^2$ is a bijection in both \mathbb{R}^\times and $-\mathbb{R}^\times$. However, there may be cusps or other singular points when $y = 0$ if the defining cubic has a multiple root.

The following is well-known:

Theorem 3.3. *Let $y = x^3 + px + q$ be a cubic. Define the discriminant as*

$$\Delta = -4p^3 - 27q^2$$

Then y has 3 distinct real roots if $\Delta > 0$, a double (or triple) root if $\Delta = 0$ and exactly one real root if $\Delta < 0$.

In the case of elliptic curves, the discriminant is defined in the same way, but with an additional factor of 16. This factor is irrelevant to this discussion, and we will omit it, but it will probably be used in more advanced literature on elliptic curves.

If $\Delta \neq 0$, all roots are distinct and the curve is nonsingular. Depending on the sign of Δ , it will consist of either one or two connected segments, as shown in Figure 5

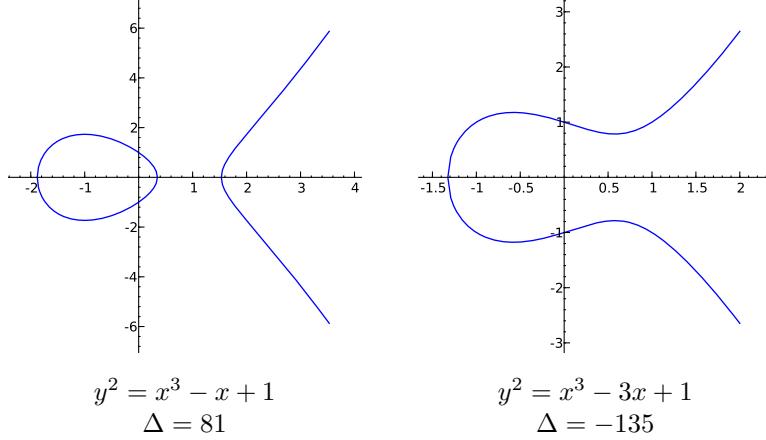


Figure 5: Elliptic curves with positive and negative discriminant, respectively. Notice that the positive (top) halves of each curve “look like” positive halves of an ordinary cubic equation, and it is easy to see how the sign of the discriminant affects the topology of the curve.

If $\Delta = 0$, we say the curve is *singular*. The singularity is at the multiple root and is either a cusp, as in the case of a triple root, a node (two intersecting

lines) in the case of a double root with nonnegative values near the root, or an isolated point in the case of a double root with nonpositive values near the root. Examples of the cusp and node cases are shown in Figure 6.

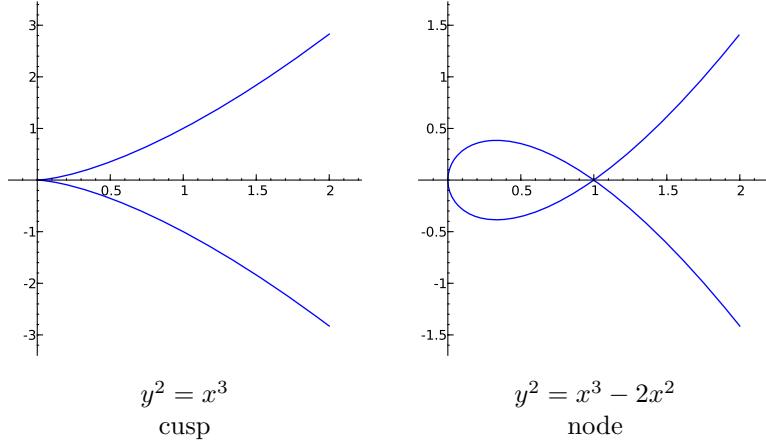


Figure 6: Singular elliptic curves with a cusp and a node, respectively. Again, notice how the positive (top) part of each cubic produces each singularity. The curve $y^2 = x^3 - 2x^2$ is not in the form $y^2 = x^3 + ax + b$, but can be placed in that form by shifting along $x \mapsto x - 2/3$.

In this paper we will not be concerned with singular curves.

3.3 Reduction modulo p

We finally arrive at the theme of this paper: what happens to an elliptic curve when it is reduced modulo a prime.

Let E have the form $y^2 = x^3 + ax + b$ with a and b rational in simplest form. If neither denominator of a, b is a multiple of p , it has an inverse in \mathbb{F}_p , so the coefficients of E in \mathbb{F}_p are integers in \mathbb{F}_p .

A nonsingular curve reduced modulo p can become singular if $p|\Delta$, since then the discriminant in \mathbb{F}_p becomes zero. If this is the case, we say E has *bad reduction* modulo p . If E over \mathbb{F}_p is non-singular, we say E has *good reduction* modulo p .

4 Endomorphisms and Hasse's Theorem

Let E be an elliptic curve. For all primes p over which E has good reduction, there is a square-root-accurate estimate known as Hasse's Theorem.

Theorem 4.1. let E be a non-singular elliptic curve over a prime field \mathbb{F}_p . Let $N_p(E)$ be the number of points on E . Then $N_p(E) \approx p + 1$ and

$$|N_p(E) - (p + 1)| \leq 2\sqrt{p}$$

To prove Hasse's theorem in the general case, it is necessary to develop background in the more advanced algebraic properties of elliptic curves.

4.1 Isogenies and Isomorphisms

For two elliptic curves, we define an isogeny as follows:

Definition 4.1. An isogeny ϕ from E_1 to E_2 which are elliptic curves is a homomorphism such that $\phi(\mathcal{O}) = \mathcal{O}$.

If ϕ is an isogeny that maps E to itself, we say ϕ is an *endomorphism* of E . The set of all endomorphisms of E forms a ring under function addition and composition:

$$\begin{aligned} (\phi + \psi)(P) &= \phi(P) + \psi(P) \\ (\phi \circ \psi)(P) &= \phi(\psi(P)) \end{aligned}$$

If the isogeny is defined by an irreducible polynomial, we define the *degree* of an isogeny as the degree of the polynomial. Since the degree of a polynomial is multiplicative, so is the degree of an isogeny, that is,

$$\deg(p \circ q) = \deg(p) \deg(q)$$

The trivial map $\phi : P \mapsto P$ is an endomorphism with degree 1. We will denote this map by 1, being the multiplicative identity of the endomorphism ring of E . There are two other important endomorphisms of elliptic curves.

Theorem 4.2. Let E be an elliptic curve. The map $\phi_m : E \rightarrow E$ with $\phi(P) = mP$, denoted $[m]$ is an endomorphism of E with degree m^2 .

Theorem 4.3. Let E be an elliptic curve over \mathbb{F}_p where p is prime. The map ϕ with $\phi[(x, y)] = (x^p, y^p)$, called the Frobenius endomorphism is an endomorphism of E with degree p .

4.2 Another property of the degree

We need the following lemma:

Theorem 4.4. Let ϕ and ψ be endomorphisms of E . Then

$$|\deg(\phi - \psi) - \deg(\phi) - \deg(\psi)| \leq 2\sqrt{\deg(\phi) \deg(\psi)}$$

Note that $\phi - \psi$ may be reducible over E , so the degree of the sum is not necessarily the maximal degree. We will not prove this lemma. (A proof can be found in [2].)

4.3 Hasse's Theorem

We are now ready to prove Hasse's theorem. (This follows the proof given in [8])

Proof. Consider the Frobenius endomorphism on E in \mathbb{F}_p where p is prime. This is the map

$$\phi : (x, y) \mapsto (x^p, y^p)$$

Fermat's little theorem tells us that

$$x^p \equiv x \pmod{p}$$

So the map fixes E pointwise, that is,

$$\phi(P) = P$$

Then $\phi(P) - P = 0$, so $(\phi - 1)(P) = 0$.

$$P \in \ker(\phi - 1)$$

Thus E is isomorphic to the kernel of the map $(\phi - 1)$. This may seem obvious, but it is sufficient to prove Hasse's theorem. The isomorphism yields

$$N_p(E) = \#\ker(\phi - 1) = \deg(\phi - 1)$$

By Theorem 4.4,

$$|\deg(\phi - 1) - \deg(\phi) - \deg(1)| \leq 2\sqrt{\deg(\phi)\deg(1)}$$

But $\deg(\phi - 1) = N_p(E)$, $\deg(\phi) = p$, and $\deg 1 = 1$, so

$$|N_p(E) - p - 1| \leq 2\sqrt{p}$$

And we are done. □

5 Beyond Hasse's Theorem

Hasse's theorem applies to all nonsingular elliptic curves. From Figure 3, we see that it is not possible to do better in general. However, we can make statements concerning the statistical distribution of Hasse error terms over the ensemble of primes p . Specifically, let

$$a_p = \frac{n_p(E) - p - 1}{2\sqrt{p}}$$

Hasse's theorem tells us that $-1 \leq a_p \leq 1$, and we can ask about the distribution of the a_p for a given elliptic curve over all primes p .

To do so, we consider separately two classes of elliptic curves: ones with “complex multiplication” and ones without. Between the two classes, different conclusions can be made about the distribution of the a_p 's.

5.1 Elliptic Curves with Complex Multiplication

Let P be a point in E over \mathbb{F}_p and m be an integer. Then mP is also a point in E and multiplication by m is an endomorphism of E . It follows that the ring of integers is contained in the endomorphism ring of E .

Suppose $\text{End}[E] \neq \mathbb{Z}$, that is, there is an endomorphism of E not corresponding to multiplication by an integer. Then we say that E has complex multiplication, commonly abbreviated CM.

For example (see [7]), consider the curve given by

$$E : y^2 = x^3 + x$$

and the map

$$\mu : (x, y) \mapsto (-x, iy)$$

The image of E in μ is the curve

$$(iy)^2 = (-x)^3 - x$$

Which is the same curve. Thus μ is an endomorphism of E , but μ does not correspond to an integer multiplication. Thus we say E has complex multiplication.

If E is CM, then the following facts are known.

- For half of the primes p , $a_p = 0$, that is, $N_p(E) = p + 1$ exactly.
- For each of the finitely many isomorphism classes of CM elliptic curves, there exist formulas producing exactly the values of a_p for each prime p . These formulas can be proven independently using the properties of the complex endomorphism. (These are proven in [5]. Look for Theorem 1.1.)

The distribution of a_p -s will thus have a peak at 0. If the peak is removed, we can see the additional structure in the distribution, as for two CM curves in Figure 7.

5.2 Elliptic Curves Without Complex Multiplication

In the case when the curve does not have complex multiplication, no explicit formulas exist. However, a precise conjecture has been formulated independently by Sato and Tate concerning the distribution of the a_p 's over non-cm curves. (See [4] for a description of the Sato Tate conjecture and a discussion related to its proof. The proof itself can be found in [9].)

Conjecture 5.1. *let E be a elliptic curve without complex multiplication. Let a_p be Hasse error terms for E over \mathbb{F}_p . Then for any function $F(t)$,*

$$\lim_{n \rightarrow \infty} \frac{1}{\pi(n)} \sum_{p \leq n} F(a_p) = \frac{2}{\pi} \int_{-1}^1 F(t) \sqrt{1 - t^2} dt$$

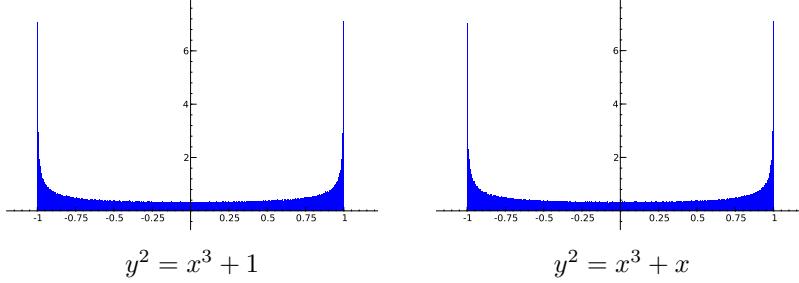


Figure 7: Distribution of the a_p -s corresponding to the first ten million primes for two CM elliptic curves, excluding the points for which $a_p = 0$. The two distributions look similar, but we would be able to compute them exactly if we knew the formulas for the two elliptic curves.

Where $\pi(n)$ is the number of primes less than or equal to n .

For example, if F is a characteristic function, that is, a function whose image is 1 on a set and 0 everywhere else, the statement above tells us that the a_p -s of a non-CM curve conform to the semicircular distribution with probability density function

$$\frac{2}{\pi} \sqrt{1 - t^2}$$

We illustrate the Sato-Tate conjecture with a few curves in Figure 8.

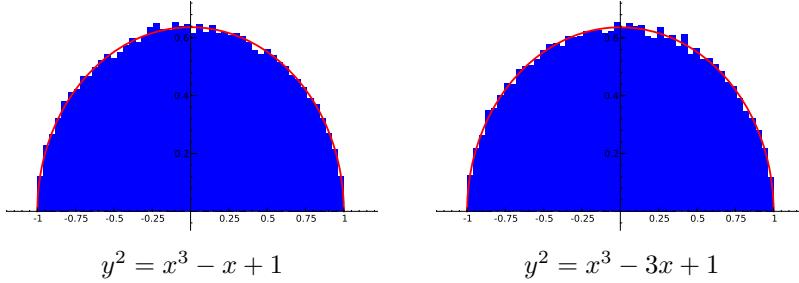


Figure 8: Distribution of the a_p -s corresponding to the first million primes for two elliptic curves (blue). Both distributions visibly conform to the Sato-Tate semicircular distribution (red).

The Sato Tate conjecture has been proven for most non-CM elliptic curves. A stronger result, due to Akiyama and Tanigawa, is conjectured (see [4]) and concerns the rate of convergence to the Sato-Tate distribution.

Conjecture 5.2. *Let $F(t)$ be any characteristic function on $[-1, 1]$. Then for all $\epsilon > 0$ there exists a $N > 0$ such that for all F ,*

$$\left| \frac{1}{\pi(N)} \sum_{p \leq N} F(a_p) - \frac{2}{\pi} \int_{-1}^1 F(t) \sqrt{1-t^2} dt \right| < N^{-1/2+\epsilon}$$

It is known that the Akiyama-Tanigawa conjecture implies the Generalized Riemann Hypothesis, and it is believed that they are equivalent.

References

- [1] Artin, M.: *Algebra*, Prentice Hall (1991)
- [2] Hendricks, K.: *On the Proof of Hasse's Theorem* (term paper)
- [3] Husemoller, D.: *Elliptic Curves*, Graduate Texts in Mathematics, Springer (2004)
- [4] Mazur, B.: *Finding Meaning in Error Terms* (preprint)
- [5] Rubin, K., Silverberg, A.: *Point Counting on Reductions of CM Elliptic Curves* (preprint)
- [6] Silverman, J.: *The Arithmetic of Elliptic Curves*, Springer (2009)
- [7] Silverman, J., Tate, J.: *Rational Points on Elliptic Curves*, Undergraduate Texts in Mathematics, Springer (1992)
- [8] Swierczewski, C., Stein, W.: *Connections Between the Riemann Hypothesis and the Sato-Tate Conjecture* (senior thesis)
- [9] Taylor, R.: *Automorphy for some l -adic lifts of automorphic mod l representations. II* (preprint)