SATO-TATE AND BEYOND

DONU ARAPURA

1. Families of Complex Curves

Let $f(x)$ be a nonconstant complex polynomial with distinct roots. Then the set of complex solutions of the equation $y^2 = f(x)$ is a Riemann surface (or a smooth complex curve if you're an algebraic geometer) which can compactified by adding at most two points at infinity. I'll be sloppy and use the same equation for the compactified curve. As a topological space, this is just a sphere with g handles where g is the genus. The classical Riemann-Hurwitz formula implies

$$g = \lceil deg\, f/2 \rceil - 1$$

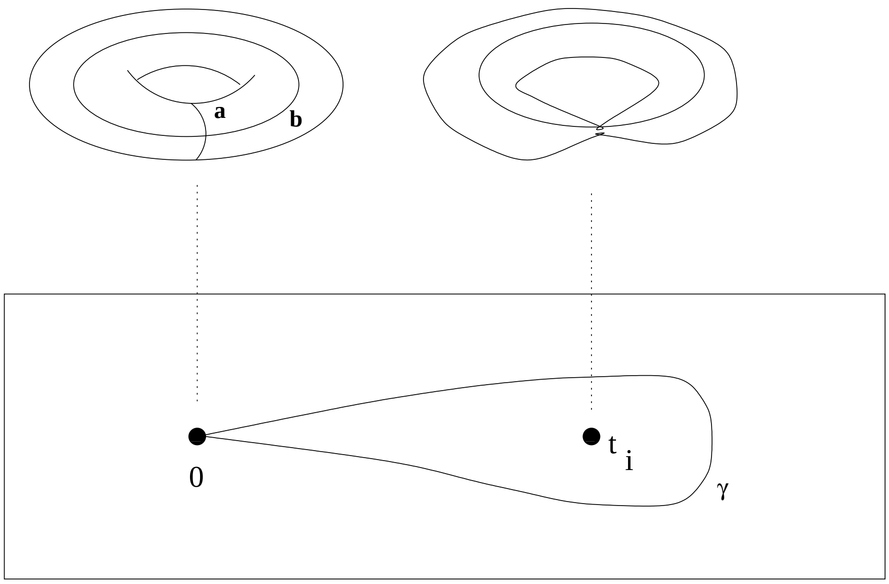
where $\lceil\ \rceil$ means round up. (Aside to experts: the restriction to hyperelliptic curves is not essential for following discussion.)

Now suppose that $f_t(x)$ is a family of polynomials depending polynomially on a complex parameter t. Assume furthermore that $f_t(x)$ has distinct roots for $t = 0$, then it follows that this would be true for all but a finite number of critical values $t_0, t_1, \ldots t_n$. The solutions of $y^2 = f_t(x)$, as t varies, forms a complex surface mapping to $\mathbb{C}$ (or the Riemann sphere after compactifying). If we delete the points lying over $\{t_i\}$, we get a fiber bundle which usually nontrivial.

I want to explain what happens in the first interesting case where $g = 1$ (so $deg f = 3, 4$). Then $y^2 = f_t(x), t \notin \{t_i\}$

These are slides to my colloquium talk at U. of Arizona.

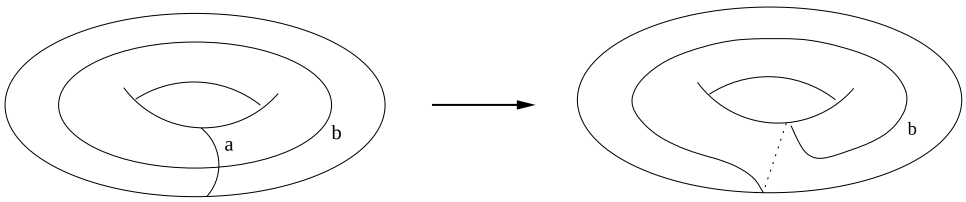
is a family of elliptic curves, which topologically is a bundle of tori. The first homology of the fiber over 0 is free abelian group generated by the curves a, b in the picture.



If we restrict this to a loop, we get a bundle on $[0,1]$ with the ends glued together. The gluing gives an automorphism of homology which is just an invertible 2×2 matrix with integer entries. So we get a representation called the monodromy

$$\rho : \pi_1(\mathbb{C} - \{t_0, t_1 \dots\}, 0) \to GL_2(\mathbb{Z}).$$

In the above picture, the gluing around γ is via a Dehn twist along a.



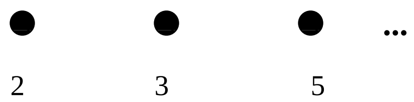
So

$$\rho(\gamma) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

In the general case, the monodromy takes values in $GL_{2g}(\mathbb{Z})$. A bit more analysis, using the intersection pairing, shows that the image lies in $Sp_{2g}(\mathbb{Z})$ (which is $SL_2(\mathbb{Z})$ when $g = 1$).

2. Arithmetic

Now suppose that $f(x)$ is an integer polynomial with distinct complex roots. We can formally replace the parameter t by a prime number $p \in \mathbb{Z}$, and the complex plane by the "line"



Then $f(x)$ gives rise to a family of polynomials $f(x)\ (mod\, p) \in \mathbb{Z}/p\mathbb{Z}[x]$. We can consider set of solutions of the congruences $y^2 \equiv f(x)\ (mod\, p)$, as p varies over the set of primes, as an analogue of the family of curves considered earlier. There is a finite set of "bad primes" which play the role of the critical values $t_0, \ldots$.

There is one important difference in the arithmetic case: the sets of solutions are finite. Let $N_1(p)$ *be the number of solutions of* $y^2 \equiv f(x)(mod\, p)$ (including points at infinity). I now come to the main question:

Question 1. *How do the numbers $N_1(p)$ vary as a function of p?*

The easy case is when $deg f(x) \leq 2$. Then it's not hard to show that $N_1(p) = 1 + p$ for good primes. This is about as nice an answer as one can expect. In general the expected answer will be more qualitative.

Case of elliptic curves.

Assume $deg f = 3$ **for this section.**

Let's consider the example

$$y^2 = x^3 - a$$

then it's possible to give an explicit answer for 50% of all primes. More precisely,

Lemma 1. *$N_1(p) = 1 + p$ if $p \not\equiv 1\,(mod\, 3)$.*

Proof. The condition ensures that the cube map $(\mathbb{Z}/p\mathbb{Z}) \to (\mathbb{Z}/p\mathbb{Z})$ is bijective. Therefore there is a unique solution $((y^2 + a)^{1/3}, y)$ for each $y \in \mathbb{Z}/p\mathbb{Z}$, plus one solution at infinity. □

The above examples are very special. These are all examples of *elliptic curves with complex multiplication*, which roughly means that they have extra symmetries. In general, $N_1(p) \neq 1 + p$, but the difference is bounded.

Theorem 1. *(Hasse) If p is a good prime, then*

$$|1 + p - N_1(p)| \leq 2\sqrt{p}$$

This suggest writing

$$1 + p - N_1(p) = 2\sqrt{p}\cos\theta(p)$$

with $\theta(p) \in [0, \pi]$. Call $\theta(p)$ a Frobenius angle.

For the examples $y^2 = x^3 - a$, we saw that the $\theta(p) = 0$ 50% of the time. In the generic cases the behaviour should be more random:

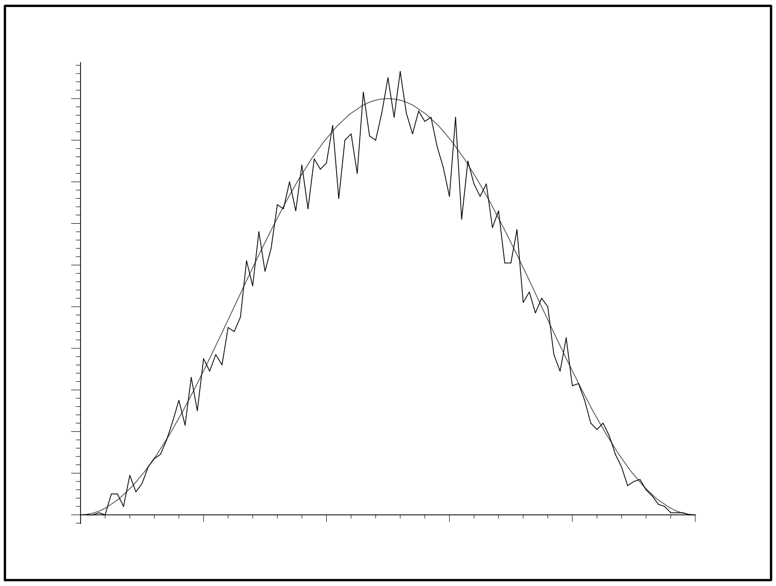
Conjecture 1. *(Sato-Tate) If the curve does not have complex multiplication, the probability that the Frobenius angle $\theta(p)$ lies in $[a,b]$ is given by*

$$\frac{2}{\pi}\int_a^b \sin^2\theta\, d\theta$$

The curve

$$y^2 = x(x-1)(x-3)$$

does not have complex multiplication. Here is a plot[1] of the distribution of angles for good primes $< 100,000$, for this curve. It's superimposed on the Sato-Tate distribution (the smooth graph).



[1] The plotting was done using Maple, however the number crunching was handled by software I wrote myself in C. The sources will be available at http://www.math.purdue.edu/~dvb

Genus two and beyond.

In the general case, it is necessary to introduce several angles. Let $N_n(p)$ denote the number of solutions of $y^2 - f(x) = 0$ in the field with p^n elements. For general g, Weil proved the following refinement of Hasse's theorem.

Theorem 2. *(Weil) If p is a good prime, then*

$$|1 + p^n - N_n(p)| \leq 2p^{n/2}g$$

Using some manipulations involving Zeta functions, it's possible to deduce an equivalent formulation:

Theorem 3. *If p is a good prime, there exists angles $\theta_1(p), \ldots \theta_g(p) \in [0, \pi]$ such that*

$$1 + p^n - N_n(p) = \sum_j 2p^{n/2} \cos(n\theta_j(p))$$

From now on assume that $deg f(x) = 5$**, so that** $g = 2$. In this case, we have a pair of Frobenius angles. If we assume $\theta_1(p) \leq \theta_2(p)$, then the above equations can be solved:

$$\theta_1(p) = \cos^{-1}\left(\frac{P_1 - \sqrt{2P_2 - P_1^2}}{2}\right)$$

$$\theta_2(p) = \cos^{-1}\left(\frac{P_1 + \sqrt{2P_2 - P_1^2}}{2}\right)$$

where

$$P_1 = \frac{1 + p - N_1(p)}{2\sqrt{p}}$$

$$P_2 = \frac{1 + p^2 - N_2(p)}{4p} + 1$$

As in the case of elliptic curves, it's possible to exhibit nonrandom behviour for special cases:

Lemma 2. *For the curve $y^2 = x^5 - a$, the angles $(\theta_1(p), \theta_2(p)) = (\frac{\pi}{2}, \frac{\pi}{2})$ if $p \not\equiv 1 \,(mod\, 5)$.*

To eliminate these kinds of examples, we will impose a monodromy condition which will be explained later. But roughly speaking the idea is to eliminate extra symmetries. Here's my guess for what the analogue of the Sato-Tate conjecture should be:

Conjecture 2. *When the curve $y^2 = f(x)$ has "big monodromy", the angles $(\theta_1(p), \theta_2(p))$ are distributed uniformly, in $\{0 \le \theta_1 \le \theta_2 \le \pi\}$, with respect probability distribution*

$$\sigma(\theta_1, \theta_2) = C \sin^2(\theta_1) \sin^2(\theta_2) \sin^2((\theta_1 + \theta_2)/2) \sin^2((\theta_1 - \theta_2)/2)$$

(where C is the appropriate normalization constant).

Turning to more practical issues, computing $N_2(p)$ involves working in a field with p^2 elements, and this takes much longer than computing $N_1(p)$ for large p. I want to describe a consequence of the above conjecture which is easier to test. Let

$$\phi(p) = \cos \theta_1(p) + \cos \theta_2(p) = \frac{1 + p - N_1(p)}{2\sqrt{p}}$$

Let $\psi(x)$ be the function satisfying

$$\int_{-2}^{x} \psi(x)\, dx = \int \int_{R(x)} \sigma/2 \, d\theta_1 d\theta_2$$

where

$$R(x) = \{(\theta_1, \theta_2) \in [0, \pi]^2 | \cos \theta_1 + \cos \theta_2 \le x\}$$

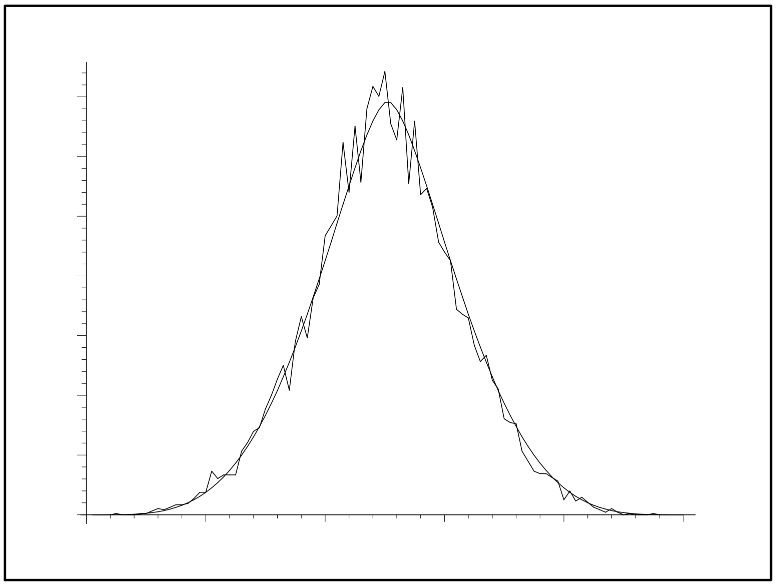
It's easy to determine an integral representation for ψ by differentiating (using Maple!), and this can be used to compute ψ numerically.

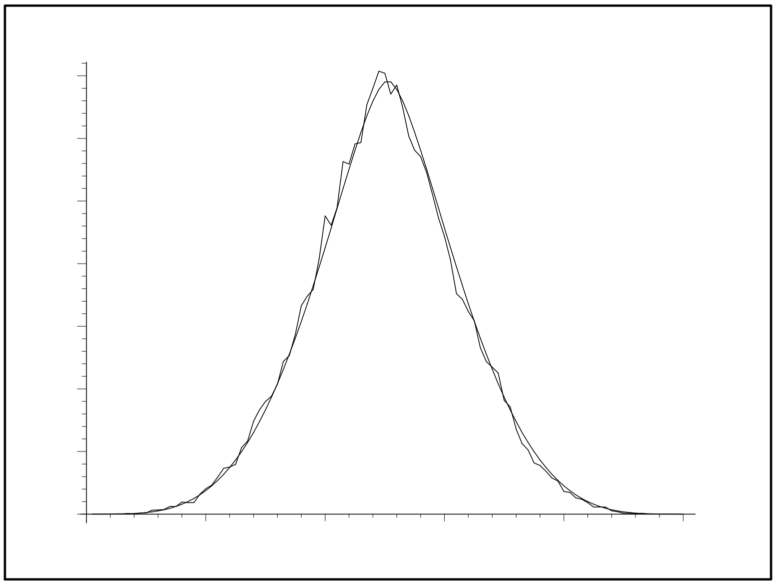
Conjecture 3. *For a genus 2 curve with big monodromy, the numbers $\{\phi(p)\}$ are distributed uniformly with respect to ψ*

Here are plots of the distribution of ϕ for

$$y^2 = x(x-1)(x-2)(x-3)(x-7).$$

for all good primes $< 100,000$, and $< 500,000$. They're superimposed on graphs of ψ.





3. The real story

In this final section, I want to explain where all these speculations come from. Given a smooth algebraic curve over $\mathbb{Q}$ (not necessarily defined by $y^2 = f(x)$), there is an arithmetic analogue of the monodromy representation considered in the first section. We need to find a replacement for $\pi_1(\mathbb{C} - \{t_i\})$, which can be defined as the group of covering transformations of the universal cover of the punctured t-line. The algebraic closure $\bar{\mathbb{Q}}$ of $\mathbb{Q}$ will play the role of the universal cover, and the Galois group $Gal(\bar{\mathbb{Q}}/\mathbb{Q})$ is like the group of covering transformations. Instead of homology, we can use Grothendieck's étale cohomology, which has an action of the Galois group. Then we have a monodromy representation

$$Gal(\bar{\mathbb{Q}}/\mathbb{Q}) \to Gl_{2g}(\mathbb{Z}_l),$$

where l is a prime[2], and $\mathbb{Z}_l$ is the ring of l-adic integers. There are noncanonical embeddings $\mathbb{Z}_l \subset \mathbb{C}$. After choosing one, we get a representation

$$\rho : Gal(\bar{\mathbb{Q}}/\mathbb{Q}) \to Gl_{2g}(\mathbb{C}).$$

We can now say what *big monodromy* means. It means that the image of ρ is, roughly speaking, as big as possible (specifically that its Zariski closure contains $Sp_{2g}(\mathbb{C})$). The condition occurred implicitly in the Sato-Tate conjecture because of:

Theorem 4. *(Serre). An elliptic curve has big monodromy if and only if it doesn't have complex multiplication.*

In general, if a curve X has big monodromy then it cannot have extra symmetries (more precisely the Jacobian doesn't have any nontrivial endomorphisms defined over $\bar{\mathbb{Q}}$). For example, the curve $y^2 = x^5 - a$ cannot have big monodromy

[2]I will assume that l is large, but the choice won't really matter

because it has nontrivial automorphisms coming from the action of the 5th roots of unity.

The Galois group $Gal(\bar{\mathbb{Q}}/\mathbb{Q})$ is a pretty mysterious object. However, there are some special elements that are better understood. For each prime p, there is an element F_p, well defined up to conjugacy, called the Frobenius. On a root of unity ζ of order prime to p, it acts by $F_p(\zeta) = \zeta^p$. The relationship with the previous part of the story is the following: *for good primes, the eigenvalues of $\rho(F_p)$ (which depend only on the conjugacy class) are given by $\sqrt{p}e^{\pm i\theta_j(p)}$.*

Let K be a maximal compact subgroup of $Sp_{2g}(\mathbb{C})$, and let $Conjugacy(K)$ be the set of conjugacy classes of K. This has a measure induced by Haar measure. It's not hard, using the previous estimates, to show that the normalized elements $\frac{1}{\sqrt{p}}\rho(F_p)$ lie in a maximal compact subgroup which would be conjugate to K. Hence the conjugacy classes of these elements can be regarded as elements of $Conjugacy(K)$. The real conjecture is the following[3]:

Conjecture 4. *(generalized Sato-Tate) If the curve has big monodromy, the conjugacy classes of $\frac{1}{\sqrt{p}}\rho(F_p)$ are uniformly distributed with respect to the measure induced by Haar measure on $Conjugacy(K)$.*

In the case $g = 1$, take $K = SU(2)$. Any matrix in K is conjugate to a unique matrix of the form

$$\begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}$$

with $\theta \in [0, \pi]$. So we can identity $Conjugacy(K) = [0, \pi]$. The Haar measure on this space is precisely $\frac{2}{\pi}\sin^2\theta$ times Lebesgue measure. So the speculation implies Sato-Tate.

[3]This is stated explicitly in the book "Random matrices, Frobenius eigenvalues, and monodromy" by Katz and Sarnak

For $g = 2$, and suitable K, $Conjugacy(K)$ can be identified with the triangle in $[0, \pi]^2$ lying above the diagonal. The measure is exactly

$$C \sin^2(\theta_1) \sin^2(\theta_2) \sin^2((\theta_1 + \theta_2)/2) \sin^2((\theta_1 - \theta_2)/2)$$

times Lebesgue measure. This is where conjecture 2 comes from.