FINDING MEANING IN ERROR TERMS

BARRY MAZUR

In memory of Serge Lang

INTRODUCTION

Four decades ago, Mikio Sato and John Tate predicted the shape of probability distributions to which certain “error terms” in number theory conform. Their prediction—known as the Sato-Tate Conjecture—has been verified for an important class of cases thanks to the recent work of Laurent Clozel, Michael Harris, and Richard Taylor [3], and of Michael Harris, Nicholas Shepherd-Barron, and Richard Taylor [18], combined with Richard Taylor’s most recent [54], which establishes this advance in our understanding.

Part of the beauty of this breakthrough is how it pulls together progress made over the past quarter century and work from significantly different viewpoints—from the theory of automorphic representations, from algebraic geometry, and from Galois deformation theory—a demonstration, yet again, of the intense unity of mathematical thought.

My aim is to discuss, in concrete terms, two “sample problems”—one still open, and one settled by the recent work—that give rise to error terms, about which the Sato-Tate Conjecture makes precise predictions.

I thank Andrew Granville, Michael Harris, Nick Katz, Mark Kisin, Phillipe Michel, Michael Rubinstein, Peter Sarnak and Richard Taylor for much enlightening discussion and for their exceedingly helpful comments. I’m grateful to William Stein for conversations and advice about the substance and the format of this article and for the computations and graphs that appear in it, and also to Christopher Swierczewski for computations, among which are those that produced the q-q plots. All plots were created using the free software SAGE (sagemath.org). For a file containing all figures in this article, with codes, see http://wstein.org/mazur/sato.tate.figures. I also want to thank Susan Holmes, who explained to me the nature and utility of q-q plots.
Contents

Part I. The general question of error terms. Our first “sample problem”. 187
1. Error terms and the Sato-Tate Conjecture 187
1.1. Why are there still unsolved problems in number theory? 187
1.2. Much of the depth of the problem is hidden in the structure of the error term 188
1.3. Strict square-root accuracy 189
1.4. Some sample arithmetic problems 189
1.5. The “next question” 192
1.6. The distribution of scaled error terms 192
1.7. Rates of convergence (first version) 194
1.8. Error term roulette 196
1.9. Eisenstein series as good approximation and error term as cusp form 196

Part II. An elliptic curve. Our new “sample problem”. 198
2. The number of points of an elliptic curve when reduced mod \( p \), for varying \( p \) 198
2.1. The elliptic curve that we will be working with 198
2.2. Rates of convergence (second version) 200
2.3. Overcounts versus undercounts 201
2.4. Error terms modulo \( m \) 204
2.5. Correlations 205

Part III. About the proof of Sato-Tate for the elliptic curve \( E \). 205
3. Reducing the problem to a question about analytic continuation of \( L \)-functions 205
3.1. The Sato-Tate distribution 205
3.2. Bases for the ring of polynomials 207
3.3. \( L \)-functions 208
3.4. Sato-Tate and the Generalized Riemann Hypothesis 209
3.5. Meromorphic extension of \( L \)-functions 210
4. Replacing the problem of analytic continuation of \( L \)-functions by questions about automorphic forms 210
4.1. The reciprocity “divide” 210
4.2. Automorphic representations, automorphic forms 211
4.3. Galois representations associated to the symmetric \( m \)-th powers of our data 215
4.4. Digression on compatible families and Galois characters 216
4.5. Langlands reciprocity 218
4.6. Potential automorphy 218
4.7. Galois deformation theorems and the pivotal role played by residual representations 219
4.8. Hopping from one prime to another 221
4.9. A rich source of potentially automorphic Galois representations 222
4.10. Concluding the theorem 223
4.11. Interpreting Sato-Tate as a statement about equidistribution 223
4.12. Expository accounts of this recent work 225
About the author 225
References 226
PART I. THE GENERAL QUESTION OF ERROR TERMS. OUR FIRST
“SAMPLE PROBLEM”.

1. ERROR TERMS AND THE SATO-TATE CONJECTURE

1.1. Why are there still unsolved problems in number theory? Eratosthenes, to take an example—and other ancient Greek mathematicians—might have imagined that all they needed were a few powerful insights and then everything about numbers would be as plain, say, as facts about triangles in the setting of Euclid’s *Elements of Geometry*. If Eratosthenes had felt this, and if he now—transported by some time machine—dropped in to visit us, I’m sure he would be quite surprised to see what has developed.

To be sure, geometry has evolved splendidly but has expanded to higher realms and more profound structures. Nevertheless, there is hardly a question that Euclid could pose with his vocabulary about triangles that we can’t answer today. And, in stark contrast, many of the basic naive queries that Euclid or his contemporaries might have had about primes, perfect numbers, and the like would still be open.

Sometimes, but not that often, in number theory we get a complete answer to a question we have posed, an answer that finishes the problem off. Often something else happens: we manage to find a fine, simple, *good approximation* to the data or phenomena that interests us—perhaps after some major effort—and then we discover that yet deeper questions lie hidden in the error term, i.e., in the measure of how badly our approximation misses its mark.

A telling example of this, and of how in the error term lies richness, is the manner in which we study $\pi(X) :=$ the number of prime numbers less than $X$. The function $\pi(X)$ is shown in Figure 1.1 in various ranges as step functions giving the “staircase” of numbers of primes.

As is well known, Carl Friedrich Gauss, two centuries ago, computed tables of $\pi(X)$ by hand, for $X$ up to the millions, and offered us a probabilistic “first” guess for a nice smooth approximating curve for this data, a certain beautiful curve that, experimentally, seems to be an exceptionally good fit for the staircase of primes.

The data, as we clearly see, certainly cries out to us to guess a *good approximation*. If you make believe that the chances that a number $N$ is a prime is inversely proportional to the number of digits of $N$, you might well hit upon Gauss’s guess, which produces indeed a very good fit. In a letter written in 1849 Gauss claimed that as early as 1792 or 1793 he had already observed that the density of prime

![Figure 1.1. The step function $\pi(N)$ counts the number of primes up to $N$.](image-url)
numbers over intervals of numbers of a given rough magnitude \(X\) seemed to average \(1/\log X\). (Here \(\log\) is the natural logarithm; i.e., to the base \(e\).)

The Riemann Hypothesis is equivalent to saying that the integral \(\int_2^X dx/\log x\) (i.e., the area under the graph of the function \(1/\log x\) from 2 to \(X\)) is essentially square root close to \(\pi(X)\). That is, if we take the difference between \(\pi(X)\) and \(\int_2^X dx/\log x\) as the error term in our attempt to estimate \(\pi(X)\), i.e., if we set

\[
\text{Error}(X) = \pi(X) - \int_2^X dx/\log x,
\]

then the Riemann Hypothesis is equivalent\(^1\) to saying that for every \(\epsilon > 0\), we have that

\[
|\text{Error}(X)| < X^{\frac{1}{2} + \epsilon}
\]

for \(X\) sufficiently large.

1.2. Much of the depth of the problem is hidden in the structure of the error term. In a general context, once we make what we hope to be a good approximation to some numerical data, we can focus our attention on the error term that has thereby been created, namely:

\[
\text{Error term} = \text{Exact Value} - \text{Our “good approximation”}
\]

In our attempt to understand \(\pi(X)\), i.e., the placement of primes in the sequence of natural numbers, we chose in the previous subsection—with Gauss—our good approximation to be the smooth function \(\int_2^X dx/\log x\), so all the essential prime placement information is still contained in the piece-wise continuous function: \(\text{Error}(X) = \pi(X) - \int_2^X dx/\log x\).

It is Riemann’s analysis of this error term that first showed us the immense world of structure packaged in it [40]. For Riemann did what is, in effect, a Fourier analysis\(^2\) of \(\pi(e^t)\) expressing \(\text{Error}(x)\) as an exact infinite sum of corrective terms, each of these corrective terms easily described in terms of the value of a zero of the Riemann zeta function; all of these corrective terms are square root small if and only if “his” hypothesis holds.\(^3\)

---

\(^1\)The Riemann Hypothesis is also equivalent to a more exacting inequality, namely, the existence of a constant \(B\) such that \(|\text{Error}(X)| < BX^{\frac{1}{2}}\log X\). For Serge Lang’s discussion of this, with a comment from his audience, see the lecture Prime Numbers in [33]. The equivalence was first proved by von Koch in 1901. See [55] and also page 113 of [6].

\(^2\)To be more precise, Riemann’s ideas provide a Fourier analysis of (the corresponding error term for) the distribution—in the sense of Schwartz—given by the derivative of the step function \(\psi(e^t)\), where \(\psi(X) := \sum_{n \leq X} \Lambda(n)\) where \(\Lambda(n)\) is equal to \(\log p\) if \(n\) is a power of the prime \(p\) and is zero otherwise. The function \(\psi(e^t)\) is a close relative to \(\pi(e^t)\) and—in the structure of its discontinuities—still packages the same basic information regarding the placement of primes among all natural numbers that \(\pi(e^t)\) does.

\(^3\)William Stein and I are writing a short book entitled What Is Riemann’s Hypothesis?, in which there will be few formulas but lots of graphs and a link to a website where people can experiment with parameters displaying data using Stein’s new computational program SAGE.
1.3. **Strict square-root accuracy.** We will be considering a somewhat different class of number theoretic problems than the example that we have been discussing, and for those problems an even stronger notion of *square-root approximation* is relevant. We will be interested in situations where the *error term* is less than a *fixed constant* times the square root of the quantity being approximated; let us say that an approximation to numerical data has **strict square-root accuracy** if its error term has this property.

We have witnessed great successes in the last century in obtaining good approximations to important problems in number theory, with error terms demonstrated to be strictly square-root accurate. Specifically, through the work of Helmut Hasse [19] in the 1930s, André Weil [56] in the 1940s and Pierre Deligne [7] in the 1970s, a large class of major approximations were proved to have this kind of accuracy. See [21] for an account of this, and for a general discussion see Joe Silverman’s book [52].

1.4. **Some sample arithmetic problems.** It has been known since the time of Fermat, and proved by Euler, that a prime $p$ can be written as a sum of two square numbers if and only if $p \not\equiv 3$ modulo 4; and if it can be written as a sum of two squares, it can be done so in only one way (not counting the order of the two squares). For example:

$$401 = 1^2 + 20^2$$

is the only way (up to changing the order of the two summands) to express the prime number 401 as a sum of two square numbers. The question of determining in how many ways a prime can be written as a sum of two squares leads, for many reasons, to a much more central and important inquiry than one might first anticipate. This problem, which seems to mix *prime numbers* with *geometry* (squares of distances to the origin of integral lattice points in the plane) has the virtue that its answer
is equivalent to knowledge of the splitting properties of primes and the validity of the unique factorization theorem in the ring of Gaussian integers.

In how many ways can the prime \( p \) be expressed as a sum of the squares of three integers? The answer for \( p \geq 5 \)—due to Gauss—is intimately connected to the structure of nonprincipal ideals in the rings of integers in quadratic imaginary number fields, and therefore—it would seem—follows, in perhaps a more complicated way, the pattern set by the question of expressing \( p \) as a sum of two squares. Specifically, the number of ways that \( p \geq 5 \) can be expressed as a sum of the squares of three integers can be given in terms of the function \( h(-d) \) the class number of the imaginary quadratic field of discriminant \(-d\) and is:

- \( 12h(-4p) \) if \( p \equiv 1, 5 \) modulo 8;
- \( 24h(-p) \) if \( p \equiv 3 \) modulo 8;
- \( 0 \) if \( p \equiv 7 \) modulo 8.

The rules of the game here are that the ordering of the summands and the signs of the integers chosen count in the tally, so for \( p = 2 \) we have \( 2 = 0^2 + (\pm 1)^2 + (\pm 1)^2 = (\pm 1)^2 + 0^2 + (\pm 1)^2 = (\pm 1)^2 + (\pm 1)^2 + 0^2 \), and therefore we have that 2 can be written “as a sum of three squares” in \( 3 \cdot 2^2 = 12 \) ways.

These two problems are simply the first two of a series of companion questions that have a long history:

In how many ways can the prime \( p \) be expressed as a sum of the squares of \( r \) integers?

Despite what happens for \( r = 2, 3 \) the pattern of answers to this question for larger \( r \) does not get consistently more complicated as \( r \) rises. To get some sample problems that drive home a point I want to make in this exposition—and for no other reason—I’ll restrict consideration to certain select values of \( r \).

For \( r = 4 \) we have a simply statable, exact, solution: the prime \( p \) can be expressed as a sum of four squares in \( 8p + 8 \) ways.

For \( r = 8 \), any odd prime number \( p \) can be expressed as a sum of eight squares in \( 16p^3 + 16 \) ways.

In both of these cases (resolved by Jacobi in the early part of the 19th century) the answer to our problem (at least for \( p > 2 \)) is a polynomial in \( p \) of degree \( r/2 - 1 \) (i.e., of degree 1 and 3, respectively). Things, however, don’t remain as simple for larger values of \( r \)—probably for most\(^4\) larger values of \( r \). To illustrate how things can change, let us focus on \( r = 24 \).

Define, then, \( N(p) \) to be the number of ways in which \( p \) can be written as a sum of 24 squares of whole numbers. Recall that squares of positive numbers, negative numbers and zero are all allowed, and the ordering of the squares of the numbers that occur in this summation also counts. Thus, the first prime number, 2, can already be written as a sum of 24 squares of whole numbers in 1,104 ways. So: \( N(2) = 1,104 \). What about \( N(p) \) for the other prime numbers \( p = 3, 5, 7, 11, \ldots \)?

\(^4\)For a discussion of this problem and its history for small values of \( r \), see page 316 of Hardy and Wright’s classic introductory text [16].
Here is some data.

<table>
<thead>
<tr>
<th>2</th>
<th>1104</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>16192</td>
</tr>
<tr>
<td>5</td>
<td>136236</td>
</tr>
<tr>
<td>7</td>
<td>44981376</td>
</tr>
<tr>
<td>11</td>
<td>6631997376</td>
</tr>
<tr>
<td>13</td>
<td>41469483552</td>
</tr>
<tr>
<td>17</td>
<td>793229226336</td>
</tr>
<tr>
<td>19</td>
<td>2697825744960</td>
</tr>
<tr>
<td>23</td>
<td>22063059606912</td>
</tr>
<tr>
<td>29</td>
<td>282507110257440</td>
</tr>
<tr>
<td>31</td>
<td>588326886375936</td>
</tr>
<tr>
<td>37</td>
<td>4119646755044256</td>
</tr>
<tr>
<td>41</td>
<td>12742799887509216</td>
</tr>
<tr>
<td>43</td>
<td>21517654506205632</td>
</tr>
<tr>
<td>47</td>
<td>5724259902057216</td>
</tr>
<tr>
<td>53</td>
<td>214623041906680992</td>
</tr>
<tr>
<td>59</td>
<td>69825476567746880</td>
</tr>
<tr>
<td>61</td>
<td>100755848394235776</td>
</tr>
<tr>
<td>67</td>
<td>2827903926520931136</td>
</tr>
<tr>
<td>71</td>
<td>5351602023957373056</td>
</tr>
<tr>
<td>73</td>
<td>7264293802635839712</td>
</tr>
<tr>
<td>79</td>
<td>17319684851070915840</td>
</tr>
<tr>
<td>83</td>
<td>29819539398107307072</td>
</tr>
<tr>
<td>89</td>
<td>64258709626203556320</td>
</tr>
<tr>
<td>97</td>
<td>165626956557080594016</td>
</tr>
</tbody>
</table>

If we eyeball the data, it is already convincingly clear that \(N(p)\) is growing less than exponentially, for otherwise the shadow of figures on the page would probably look triangular. Following the pattern we’ve seen for the smaller values of \(r\) we have considered, we might expect \(N(p)\) to be a polynomial in \(p\) of degree \(r/2 - 1 = 11\). If we had enough data, I imagine we might “curve-fit” a polynomial approximation. But happily, without having to lean on numerical experimentation, certain theoretical issues—which I will hint at in subsection 1.9 below—allow us to guess the following good approximation for the values \(N(p)\), namely the polynomial in \(p\) of degree 11:

\[
N_{\text{approx}}(p) := \frac{16}{691}(p^{11} + 1).
\]

The difference, then, between the data and our good approximation is:

\[
\text{Error}(p) := N(p) - N_{\text{approx}}(p) = N(p) - \frac{16}{691}(p^{11} + 1).
\]

This error term has been proven to be square-root small, and this is hardly an elementary result: it is a consequence of deep work of Deligne [7]. In fact, using the work of Deligne I am alluding to, you can show that

\[
|\text{Error}(p)| \leq \frac{66,304}{691}\sqrt{p^{11}}.
\]

What with that hefty constant, \(\frac{66,304}{691}\), the “smallness” of our error term here may not impress us for quite a while as we systematically tabulate the values of \(N(p)\),
but, of course, this result tells us that as we get into the high prime numbers our data will hug startlingly close to the simple smooth curve

\[ f(x) = \frac{16}{691}(x^{11} + 1). \]

1.5. The “next question”. Whenever some element of some theory is settled or is considered settled, many of us mathematicians propose a subsequent plan of inquiry with that phrase “So, the next question to ask is . . .”

Here too. Given the precise inequality

\[ |\text{Error}(p)| \leq \frac{66,304}{691}\sqrt{p^{11}} \]

described in the previous subsection and given the fact that this represents one consequence of what has been a great project that has spanned half a century of progress in number theory, some natural (and related) “next” questions arise. We might, for example, ask:

- Is the bound on this error term (e.g., the constant \( \frac{66,304}{691} \)) the best possible?
- Is \( f(x) = \frac{16}{691}(x^{11} + 1) \) the best polynomial approximation to our data?
- Might we, more specifically, find another polynomial \( g(x) \) which beats \( f(x) \) in the sense that the absolute values of the corresponding error terms \( |N(p) - g(p)| \) are \( \leq C\sqrt{p^{11}} \) with a constant \( C \) that is strictly less than \( \frac{66,304}{691} \)?
- For any given constant \( C < \frac{66,304}{691} \) is there a positive proportion of prime numbers \( p \) for which \( |N(p) - f(p)| \leq C\sqrt{p^{11}} \)?
- We might ask what that proportion is as a function of \( C \).
- We might ask for the proportion of primes \( p \) for which the error term is positive, i.e., where our good approximation is an undercount.

To be sure, we would want to phrase such questions not only about our specific “sample problem” but about the full range of problems for which we have—thanks to Deligne et al.—such good square-root close approximations.

It is the Sato-Tate Conjecture that addresses this “next”, more delicate, tier of questions.\(^5\)

1.6. The distribution of scaled error terms. Given that in our sample problem we know the bound

\[ |\text{Error}(p)| \leq \frac{66,304}{691}\sqrt{p^{11}}, \]

let us focus our microscope on the fluctuations here. Namely, consider the scaled error term

\[ \text{Scaled Error}(p) := \frac{\text{Error}(p)}{\frac{66,304}{691}\sqrt{p^{11}}} = \frac{N(p) - \frac{16}{691}(p^{11} + 1)}{\frac{66,304}{691}\sqrt{p^{11}}} \]

so that we have:

\[-1 \leq \text{Scaled Error}(p) \leq +1.\]

\(^5\)As is only to be expected, there are whole books of questions about this sample problem that one could ask and mathematicians have asked—some of these questions being structurally important and some at least traditionally of great interest. E.g., how often is our approximate value \( N_{\text{approx}}(p) \) above exactly equal to the actual value \( N(p) \)? A conjecture of Lehmer would say that this never happens.
About this type of *scaled error value distribution*, let me recall the words of Susan Holmes, a mathematician and statistician at Stanford, who—when I sent her some numerical computations related to a similar number theoretic problem for which I had some statistical questions—exclaimed: “What beautiful data!”

But what can we say further about this data? How do these scaled error values distribute themselves on the interval \([-1, +1]\)? That is, what is the function \( I \mapsto P(I) \) that associates to any subinterval \( I \) contained in \([-1, +1]\) the probability \( P(I) \) that for a randomly chosen prime number \( p \) its scaled error term \( \text{Error}(p) \) lies in \( I \)?

In 1963, Mikio Sato (by studying numerical data) and John Tate (following a theoretical investigation) predicted—for a large class of number theoretic questions including many problems of current interest, of which our example is one—that the values of the scaled error terms for data in these problems conform to a specific probability distribution.\(^6\) Usually the Sato-Tate conjecture predicts that this distribution is no more complicated than the elementary function \( x \mapsto \frac{2}{\pi} \sqrt{1 - x^2} \), i.e., the thing whose graph is a semi-circle of radius 1 centered at the origin but squished vertically to have its integral equal to one. This makes it far from the Gaussian normal distribution! Indeed, Sato and Tate predict this type of behavior in our example problem, so that their conjecture would have it that

\[
P(I) = \frac{2}{\pi} \int_I \sqrt{1 - x^2} \, dx.
\]

This is still an open question for our sample problem! Nevertheless, we have an impressive amount of data in support of it (see below).

---

\(^6\)For a formulation of a Sato-Tate Conjecture for general motives, see pp. 344-347 of [45].
1.7. Rates of convergence (first version). The open problem of whether or not the distribution data as in Figure 1.3 converges to the Sato-Tate distribution is, in a sense, the gateway to a number of finer questions (these being therefore all the more open) such as the following. If our distribution of data “evens out” to yield the Sato-Tate law in the limit, how fast does it do this? There are various ways of formulating (and visualizing) rates-of-convergence, and we will be revisiting such issues in Part II below.

For now, consider quantile-quantile plots (as statisticians call them), which offer a slightly different way of displaying data such as $p \mapsto \text{Scaled Error}(p)$ as pictured in Figure 1.3.

Fix an interval $(a, b) \subset (-1, +1)$ and for any number $T \in (a, b)$ let

$$X(T) := \frac{\int_T^b \sqrt{1 - x^2} \, dx}{\int_a^b \sqrt{1 - x^2} \, dx}.$$ 

Fix a cutoff $C$ and let $Y_C(T)$ be the ratio

$$Y_C(T) := \frac{\#\{p < C \mid a < \text{Scaled Error}(p) < T\}}{\#\{p < C \mid a < \text{Scaled Error}(p) < b\}}.$$ 

Now plot $(X(T), Y_C(T))$ in the plane as a “curve” lying over the interval $(a, b)$ of the $x$-axis; this is the q-q-plot of our data. See Figures 1.4 and 1.5.

We want to understand rates of convergence for q-q-plots of our data over an interval $(a, b)$ and, even more importantly, to understand what structural issues need to be understood to allow us to pinpoint these rates of convergence. Specifically, how far off is the curve $T \mapsto (X(T), Y_C(T))$ from a straight line, and how fast (as $C$ goes to $\infty$) does it approach a straight line? E.g., a somewhat exacting measure for how far off the curve $T \mapsto (X(T), Y_C(T))$ is from a straight line—called the discrepancy in the literature—is the $L^\infty$-norm of the difference between $X(T)$ and $Y_C(T)$; explicitly, set:

$$D(C) := \max_{|T| \leq 1} |X(T) - Y_C(T)|.$$ 

![Figure 1.4. The q-q-plot for our scaled error terms in the interval $(0, +1)$ for the cutoff $C = 100$.](image)
The Sato-Tate Conjecture is equivalent to saying that $D(C)$ tends to 0 as $C$ goes to $\infty$. As for rates of convergence, it is natural to make the conjecture below, following the lead of Shigeki Akiyama and Yoshio Tanigawa.\textsuperscript{7}

**Conjecture 1.1.** For any positive $\epsilon$

$$D(C) = O(C^{-\frac{1}{2} + \epsilon}).$$

Readers should consult the article of Akiyama and Tanigawa [2] for analogous numerical data about related problems. Also, William Stein and Christopher Swierczewski are running computations of the $L^2$-distance between $X(T)$ and $Y_C(T)$ over intervals $T \in [a, b]$ for various choices of $a < b$ to get a further view of such convergence issues. Specifically, consider the integral

$$\Delta^b_a(C) := \sqrt{\int_a^b (X(T) - Y_C(T))^2 dT}.$$

**Definition 1.2.** The $L^2$ Sato-Tate exponent $\epsilon(a, b)$ for our scaled error terms is the lim sup of all positive numbers $e$ such that

$$\Delta^b_a(C) < C^{-e}$$

for $C$ sufficiently large.

Preliminary numerical experiments suggest that $\epsilon(a, b)$ is going to be $1/2$. For example, for $a = 0$, $b = 1$, and $C = 5000$, Stein and Swierczewski tell me that $-\log \Delta^1_0(C) / \log C = 0.482$.

\textsuperscript{7}In their article [2] Akiyama and Tanigawa make the analogue of this “rate of Sato-Tate convergence” conjecture for elliptic curves over $\mathbb{Q}$ without CM, and they accumulate numerical evidence for it. They also show that their conjecture for an elliptic curve $E$ implies the General Riemann Hypothesis for the $L$-function attached to that elliptic curve. For more about this, see subsection 3.4.
1.8. **Error term roulette.** The symmetry predicted by Sato and Tate in the data of our problem implies that in the limit our estimate would *undercount* the data about as much as it would *overcount* it. As in roulette, where instead of betting on a precise number you can simply place a bet on whether the ball lands on red or black, let us in this subsection not worry about the size of the error term but just compare undercounts versus overcounts; specifically we will plot

\[
\#\{p < C \mid \text{Error}(p) > 0\} - \#\{p < C \mid \text{Error}(p) < 0\}
\]

as a function of the cutoff \(C\) (for any \(C < 10^6\) this difference never climbs above 150):

![Figure 1.6. The difference between undercounts and overcounts.](image)

In this stretch of data there seems to be a definite bias towards the *positive*. I am thankful to Peter Sarnak, who explained to me why (using GRH and other standard and/or plausible conjectures) we might expect such a bias\(^8\) and why it is analogous to the phenomena studied in *Chebyshev’s Bias* [41], his joint article with Michael Rubinstein. Also why we might expect this bias to even out “in the end”.

1.9. **Eisenstein series as good approximation and error term as cusp form.** Ever since Euler we have acquired the instinct of packaging arithmetic functions

\[ a \mapsto M(a) \]

for \(a = 0, 1, 2, \ldots\) (or at least those arithmetic functions that are of interest to us) as the coefficients of a power series in an abstract variable, say, \(q\), i.e., to form

\[
\mathcal{M}(q) := \sum_{a=0}^{\infty} M(a)q^a,
\]

and then to hope that formal properties of this power series will saliently express interesting relations satisfied by the initial \(a \mapsto M(a)\). The primordial example of this is the packaging of the constant function \(a \mapsto 1\) as a geometric series viewed as

\(^8\)And explaining, at the same time, the bias exhibited in the graphs given in subsection 2.1.
a rational function of \( q \) with a pole at \( q = 1 \). Ever since Riemann we have acquired the further instinct of applying the full power of complex function theory to these \( M(q) \)'s.

Consider, as a germane example, our running problem, which we now state for all positive integers \( a \) and not just primes \( p \): namely, start with the arithmetic function

\[
a \mapsto N(a) := \text{the number of ways in which } a \text{ can be expressed as a sum of } 24 \text{ squares of whole numbers},
\]

and form the corresponding generating function

\[
N(q) := \sum_{a=0}^{\infty} N(a) q^a.
\]

The surprise here is that \( N(q) \) satisfies a “hidden symmetry” that can be easily expressed once one replaces the (abstract) variable \( q \) by \( e^{2\pi iz} \) and notes that \( N(e^{2\pi iz}) \) converges to yield an analytic function \( N(e^{2\pi iz}) = f(z) \) on the upper half-plane \( z = x + iy \) \( (y > 0) \). This “hidden symmetry” is simply

\[
f(-1/4z) = (2z)^{12} f(z).
\]

For evident reasons, we think of the series \( N(q) \) as the Fourier series of \( f(z) \) and the original arithmetic function \( a \mapsto N(a) \) as the Fourier coefficients of \( f(z) \).

As will be discussed at some length in later parts of this article, this hidden symmetry establishes \( f(z) \) (and its Fourier series \( N(q) \)) as a modular form of a specific sort (e.g., level 4 and weight 12). One of the miracles of the theory of modular forms of this type (i.e., of a given level and weight) is that \( N(q) \) admits a canonical expression as a sum of two modular forms of the same level and weight,

\[
N(q) = N_{\text{Eis}}(q) + N_{\text{Cusp}}(q),
\]

where the first of these modular forms,

\[
N_{\text{Eis}}(q) = \sum_{a=0}^{\infty} N_{\text{Eis}}(a) q^a,
\]

is—in the parlance of the theory—an Eisenstein series and the second,

\[
N_{\text{Cusp}}(q) = \sum_{a=0}^{\infty} N_{\text{Cusp}}(a) q^a,
\]

a cusp form.

Avoiding technical definitions, in our particular case we can pinpoint this decomposition among all other decompositions of our \( N(q) \) as a sum of two modular forms of the same level 4 and weight 12 in the following curious way:

- **The Eisenstein part:** The arithmetic function \( p \mapsto N_{\text{Eis}}(p) \) for odd primes \( p \) is a polynomial function of \( p \).
- **The Cuspidal part:** For primes \( p \) the absolute value of \( N_{\text{Cusp}}(p) \) is less than a constant times \( p^{11/2} \) (this following from the deep theorem of Deligne, previously cited).

In other words, the theory of modular forms provides us with a **conceptually elegant** choice of “good approximation”, namely

\[
N_{\text{Approx}}(p) := N_{\text{Eis}}(p),
\]

and it provides us with the ability to conceptually understand the “error term”, i.e.

\[
\text{Error}(p) := N_{\text{Cusp}}(p).
\]
For readers familiar with the theory of modular forms—see Part III for a very
brief expository discussion—here are some particulars about this decomposition.

Let

- \( \Theta(q) := \sum_{n=0}^{\infty} q^{n^2} \),
- \( \Delta(q) := q \prod_{n=1}^{\infty} (1 - q^n)^{24} = \sum_{n=1}^{\infty} \tau(n)q^n \),

so that \( \Theta(q) \) is the classical modular form of weight 1/2 and \( \Delta(q) \) is the unique
cuspidal modular form of level 1, weight 12, normalized so that it is 1 \cdot q + O(q^2);
its Fourier coefficients, \( n \mapsto \tau(n) \), are given by Ramanujan’s “tau-function”. Add
to this list the modular form \( E(q) \), the Eisenstein series of level 1 and weight 12
normalized so that for any prime number \( p \) its \( p \)-th Fourier coefficient is \( p^{11} + 1 \).

We have the equation of formal power series,

\[ N(q) = \Theta(q)^{24}, \]

as can be checked by simply multiplying things out. I thank William Stein for the
computation expressing the modular form \( \Theta(q)^{24} \) as a sum of Eisenstein series and
cusp forms of weight 12 and level 4, the answer being:

\[ N_{\text{Eis}}(q) = \frac{16}{691} E(q) - \frac{32}{691} E(q^2) + \frac{65536}{691} E(q^4) \]
and

\[ N_{\text{Cusp}}(q) = \frac{33152}{691} \Delta(q) + \frac{1525760}{691} \Delta(q^2) + \frac{135790592}{691} \Delta(q^4). \]

For an odd prime number \( p \) the \( p \)-th Fourier coefficient of \( N_{\text{Eis}}(q) \) is then \( \frac{16}{691} (p^{11} + 1) \), i.e., is our “good approximation” to \( N(p) \). The \( p \)-th Fourier coefficient of \( N_{\text{Cusp}}(q) \), i.e., our “error term”, is \( \text{Error}(p) = \frac{33152}{691} \tau(p) \).

A curious phenomenon is that although there exists a cuspidal newform of the
same weight (12) and level (\( \Gamma_0(4) \)) as \( \Theta^{24} \), this newform does not enter into the
eigenform decomposition of \( \Theta^{24} \) (i.e., \( \Theta^{24} \) is “old” in its minimal level).

**PART II. AN ELLIPTIC CURVE. OUR NEW “SAMPLE PROBLEM”**.

2. **THE NUMBER OF POINTS OF AN ELLIPTIC CURVE WHEN REDUCED MOD \( p \),
FOR VARYING \( p \)**

2.1. **The elliptic curve that we will be working with.** The example we will
use is one of the favorites of many number theorists, namely the curve in the plane,
call it \( E \), cut out by the equation

\[ y^2 + y = x^3 - x^2. \]

This is an elliptic curve that is something of a showcase for number theory in that
it has been extensively studied—much is known about it—and yet it continues to
repay study, for, as with all other elliptic curves, its deeper features have yet to be
understood. A detailed numerical discussion of the properties of this curve can be
found in section 8, part I, of [34]; for more recent numerical information about this
as well as all the other elliptic curves of low conductor, see [5].

This curve \( E : y^2 + y = x^3 - x^2 \), when extended to the projective plane, has
exactly one point on the line at infinity; and if you stipulate that its unique point at
infinity be the origin, there is a unique algebraic group law on \( E \), allowing us—for
any field \( k \) of characteristic different from 11 (i.e., any field where 11 \( \neq 0 \))—to
endow the set consisting of \( \infty \) and the points of \( E \) with values \((x, y) = (a, b) \in k \)
with the structure of an abelian group. Let \( k \) be of characteristic different from 11.
and let us denote by $E(k)$ this group of $k$-rational points of $E$. The reason why we have to exclude 11 is that the polynomial equation above modulo 11 has a singular point.

Every one of these groups $E(k)$ contains the five rational points

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

and it isn’t difficult to check that these five points compose a cyclic subgroup of $E(k)$ of order five. The data we shall be focussing on in this problem is the number of rational points that $E$ has over the prime field containing $p$ elements (excluding, again, $p = 11$). So, let $p$ be a prime number (different from 11) and let $\mathbf{F}_p$ denote the field of integers modulo $p$, and define

$$N_E(p) := \text{the number of elements in the finite group } E(\mathbf{F}_p).$$

There is much that is surprising in the numerical function $p \mapsto N_E(p)$, and here is what it looks like for small primes $p$:

<table>
<thead>
<tr>
<th>$p$</th>
<th>$N_E(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>23</td>
<td>29</td>
</tr>
<tr>
<td>29</td>
<td>31</td>
</tr>
<tr>
<td>31</td>
<td>37</td>
</tr>
<tr>
<td>37</td>
<td>41</td>
</tr>
<tr>
<td>41</td>
<td>43</td>
</tr>
<tr>
<td>43</td>
<td>47</td>
</tr>
<tr>
<td>47</td>
<td>53</td>
</tr>
<tr>
<td>53</td>
<td>59</td>
</tr>
<tr>
<td>59</td>
<td>61</td>
</tr>
<tr>
<td>61</td>
<td>67</td>
</tr>
<tr>
<td>67</td>
<td>71</td>
</tr>
</tbody>
</table>

Since, from the first of the two definitions, $N_E(p)$ is the order of a finite group that contains a cyclic group of order five, we know from Lagrange’s theorem of elementary group theory that $N_E(p)$ is divisible by 5, but what more can we say about this data?

This, now, will constitute our second sample problem on which we will be focussing for the rest of this article.

For starters, following the format of the previous sections of this article, we should look for a “good approximation” to $N_E(p)$. An old result due to Helmut Hasse [19] tells us that a square-root accurate approximation to $N_E(p)$ is given by the simple expression $p + 1$, which is, by the way, just the number of points on a line in the projective plane over $\mathbf{F}_p$.

It is a deep theorem (proved in the PhD thesis of Noam Elkies; see [11]) that for an infinite number of primes $p$, $N_E(p)$ is equal to precisely this simple expression $p + 1$. But it is generally true that the error term for this approximation is quite small. Explicitly, writing

$$\text{Error}(p) := N_E(p) - (p + 1)$$

Hasse proved the inequality

$$|\text{Error}(p)| = |N_E(p) - (p + 1)| \leq 2\sqrt{p}.$$  

Another way of saying this is that there is a conjugate pair of complex numbers $e^{i\theta_p}$ and $e^{-i\theta_p}$ for which the error term can be written as

$$\text{Error}(p) := N_E(p) - (p + 1) = \sqrt{p}(e^{i\theta_p} + e^{-i\theta_p}) = 2\sqrt{p}\cos(\theta_p).$$

As is often the case in number theory, there are other surprising ways of expressing this same data; for example, expand the infinite product

$$q \prod_{n=1}^{\infty} (1 - q^n)^2 (1 - q^{11n})^2 = \sum a_n q^n$$

and we have that (for prime numbers $p \neq 11$

$$\text{Error}(p) = -a_p.$$
Following, again, the format of our example-problem of the previous sections, we might ask for the distribution of error values, and here we can do this just by asking for the statistics of the rule that assigns to prime numbers $p$ the conjugate-pair of complex numbers on the unit circle in the complex plane

$$p \mapsto e^{\pm i \theta_p}.$$

The distribution to which this data converges as we accumulate larger and larger primes $p$ had been conjectured over forty years ago by Sato and Tate. It is only very recently that it (and many other issues of a similar genre) has finally been settled!

2.2. Rates of convergence (second version). The recent result due to Taylor et al. gives us that the data

$$p \mapsto \cos(\theta_p) = 1/2(e^{i \theta_p} + e^{-i \theta_p})$$

of the previous subsection conforms to the Sato-Tate distribution $\frac{2}{\pi} \sqrt{1 - t^2}$. That is,

**Theorem 2.1.** For any continuous function $F(t)$ on the interval $[-1, +1]$ we have that the limit

$$\lim_{C \to \infty} \frac{1}{\pi(C)} \sum_{p \leq C} F(\cos \theta_p)$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use
exists and is equal to the integral
\[ \frac{2}{\pi} \int_{-1}^{+1} F(t) \sqrt{1-t^2} \, dt. \]

Given this advance in our knowledge, we have the next question—just as in our initial sample problem in subsection 1.7 of Part I—of how fast the Sato-Tate distribution is achieved. This next question is already screaming at us, as we gaze at Figure 2.1 and at the hefty spike—i.e., very tall column—at \( a_p = 0 \), which tells us that there are lots of small primes where the error term vanishes (these coincide in the case of our example with the class of supersingular primes for the elliptic curve \( E \), the class of primes that, as we mentioned, has been shown to be infinite [11]). This seeming superfluity of supersingular primes in our diagram will eventually “even out” and settle into the predicted Sato-Tate distribution as the cutoff \( C \) proceeds to infinity. More specifically, Lang and Trotter conjecture [34] that the number of primes of supersingular primes \( < C \) for our elliptic curve \( E \) is asymptotic to a positive constant times \( C^{\frac{3}{4}}/\log C \) as \( C \) tends to infinity, while Elkies showed that it is \( O(C^{\frac{3}{4}}) \) in [12].

How fast, then, does this distribution even out to yield the Sato-Tate law in the limit?

Akiyama and Tanigawa formulate a conjecture (Conjecture 1 in [2]) that implies

**Conjecture 2.2 (Akiyama-Tanigawa).** Let \( F(t) \) be a real-valued function of bounded variation. Put
\[ \Delta F(C) := \left| \frac{1}{\pi(C)} \sum_{p \leq C} F(\cos \theta_p) - \frac{2}{\pi} \int_{-1}^{+1} F(t) \sqrt{1-t^2} \, dt \right|. \]

For every positive \( \epsilon \) we have
\[ \Delta F(C) < C^{-\frac{1}{4}+\epsilon} \]
for \( C \) sufficiently large.

The phrase “sufficiently large” means, more specifically, that there is a constant \( C(F, \epsilon) \) depending only on \( F \) and \( \epsilon \) such that we have the stated inequality for all \( C > C(F, \epsilon) \). The conjecture of Akiyama and Tanigawa even predicts a certain strong uniformity feature of this inequality with regard to its dependence on the function \( F \); namely, the function \( (F, \epsilon) \mapsto C(F, \epsilon) \) can be taken to depend only on \( V(F) \), the total variation of \( F \).

**2.3. Overcounts versus undercounts.** Again, as in our initial sample problem in subsection 1.7 of Part I, we can ask for statistics in our current sample problem for how often the estimate \( p + 1 \) exceeds the number of rational points on \( E \) modulo \( p \) and how often it falls short of that number. Explicitly, we will plot
\[ D_E(C) := \#\{ p < C \mid N_E(p) < p + 1 \} - \#\{ p < C \mid N_E(p) > p + 1 \}. \]

\(^9\)Earlier, J-P. Serre had shown this same \( O(C^{\frac{3}{4}}) \) bound conditional on GRH by a very different method; see [43].
Here, in contrast to our initial problem of Part I, we actually know—since we know the Sato-Tate conjecture for $E$—that this number is $o(C)$ (i.e., $D_E(C)/C$ tends to zero as $C$ goes to infinity). But the actual data for $C < 10^6$ is a bit more striking than that. Here is the graph of $D_E(C)$:

![Graph of $D_E(C)$](image)

**Figure 2.2.** The race between $N_E(p) < p + 1$ and $N_E(p) > p + 1$.

The difference $D_E(C)$ is no greater than 300 for any $C < 10^6$. At least so far, $N_E(p)$ tends to be a tiny bit more often $< p + 1$ than it is $> p + 1$. This is not necessarily the case for other elliptic curves; the pattern we will see in the data below (for $C < 10^6$) seems reminiscent of one of the very important heuristics in the modern history of the arithmetic of elliptic curves, namely the idea due to Bryan Birch and Peter Swinnerton-Dyer that if the rank of the group of rational points of an elliptic curve $E$ is large, one might be able to detect this by discovering that the numbers $N_E(p)$ are—in some statistical sense—larger than expected (given, of course, that these numbers are constrained to be smaller that $1 + p + 2\sqrt{p}$ and that the statistics conforms to the Sato-Tate law). The elliptic curve $E$ we are working with has rank zero (it has only five rational points), so it is interesting to choose other elliptic curves $E$ with infinitely many rational points and compute comparable data for the race between $N_E(p) < p + 1$ and $N_E(p) > p + 1$ for these curves $E$ and compare with the graph above. This computation William Stein and Chris Swierczewski do for the elliptic curves usually denoted 37A, 389A, and 5077A, which have ranks 1, 2 and 3, respectively, and for which the recent work we are reporting in this article also proves that the Sato-Tate conjecture holds. For these elliptic curves, $N_E(p)$ tends to be more often $> p + 1$ than it is $< p + 1$, at least as far as the data has been computed, i.e., up to $C = 10^6$. Here is the graph of

$$D_E(C) := \#\{p < C \mid N_E(p) < p + 1\} - \#\{p < C \mid N_E(p) > p + 1\}$$

for each of these in turn:
Peter Sarnak gave the following explanation of the qualitative features of these graphs, namely that—in the stretch of data depicted—the higher the Mordell-Weil rank of the elliptic curve, the more pronounced is the data biased towards the negative.\footnote{Sarnak explained this in a letter to me which, as I understand it, was the fruit of conversations with Granville. I’m grateful for that and for illuminating discussions with Granville, Rubinstein, and Sarnak about this phenomenon. For more on a similar topic, see [15].} Very briefly, assuming a list of standard conjectures about the behavior
of $L$-functions, together with some very plausible but less standard conjectures, Sarnak begins by showing that

$$X \mapsto \log X \sum_{p \leq X} \frac{a_E(p)}{\sqrt{p}}$$

has a limiting distribution with mean equal to $1 - 2r(E)$ where $r(E)$ is the Mordell-Weil rank of the elliptic curve $E$. The variance of this limiting distribution is the sum of the squares of the reciprocals of the absolute values of the nonreal zeros of the $L$-function of $E$. The argument for this follows the line of reasoning in the article Chebyshev’s Bias [41]. If, however, we apply similar reasoning to the quantity specifically measuring the race depicted in our graphs, i.e.,

$$X \mapsto \log X \left( \# \{p \leq X; a_E(p) > 0 \} - \# \{p \leq X; a_E(p) < 0 \} \right),$$

one computes (given reasonable conjectures) the mean, and one discovers that it conforms fairly well with the data; the variance, however, is infinite, so whatever bias we see in our finite stretch of data will eventually wash out.\(^{11}\)

2.4. Error terms modulo $m$. Our main subject is the statistics governing the position of the real numbers $\frac{a_p}{\sqrt{p}}$ in the interval $(-1, +1)$. But the $a_p$’s are integers, and so it is also perfectly reasonable to ask for the statistics of their congruence classes modulo a given positive integer $m$. For any $\alpha$ modulo $m$ how often is $a_p \equiv \alpha \mod m$? This is a genuine “companion” to the question that this article is devoted to; it is an older question and has long been answered and even (given

\(^{11}\)This is specific to elliptic curves $E$ with no complex multiplication, as our examples above all are. The nonfiniteness of the variance is related to the fact that the (expected) number of zeros—in intervals $(1/2, i/2 + iT)$ ($T > 0$)—of the $L$ function of the $n$th symmetric power of the newform $f_E$ attached to $E$ grows at least linearly with $n$.\)
the Generalized Riemann Hypothesis) with precise information about convergence rates. So, let us briefly discuss it.

First, returning to the data of the $N_E(p)$’s given in section 2.1, one suspects (and, as it turns out, with good reason) that the question of congruences modulo 5 might be idiosyncratic. (This is related to the fact that our elliptic curve has a rational point of order 5.) Questions of congruences modulo 11 and 2 also have some (minor) peculiarities: 11 because the elliptic curve has bad reduction at 11, and 2 for other more general reasons. So, to get a clean statement let us restrict our attention to a modulus $m$ that is not divisible by 2, 5, or 11.

**Theorem 2.3.** Fix $m$ an integer not divisible by 2, 5, or 11, and $\alpha$ a congruence class modulo $m$. For any cutoff $C$, let $Y_C(\alpha; m)$ denote the proportion of prime numbers $p < C$ such that $a_p \equiv \alpha \mod m$. Let $X(\alpha; m)$ denote the proportion of nonsingular $2 \times 2$ matrices with coefficients in $\mathbb{Z}/m\mathbb{Z}$ that have trace $\alpha$. Then

$$\lim_{C \to \infty} Y_C(\alpha; m) = X(\alpha; m).$$

This is a particular consequence of the classical theorem of Cebotarev, and we have a strikingly effective version of this theorem due to Lagarias and Odlyzko [37] (see also Théorème 2 of section 2.2 in [43]). If we assume the Generalized Riemann Hypothesis (for the Dedekind zeta function of the splitting field of the group of $m$-torsion points in our elliptic curve $E$), we would have that the analogue of Conjecture 2.2 holds. That is, for any positive $\epsilon$,

$$|Y_C(\alpha; m) - X(\alpha; m)| < C^{-1/2} + \epsilon$$

for $C$ sufficiently large (Théorème 4 of section 2.4 in [43]).

It might be amusing to rephrase the standard proof of the Cebotarev theorem to follow a bit more closely than it does the scenario for the proof of the Sato-Tate Conjecture discussed in Part III below.

### 2.5. Correlations.

Having discussed both the statistics governing the position of the $\frac{a_p}{\sqrt{p}}$ in the interval $(-1, +1)$ and statistics of the congruence classes of the $a_p$’s modulo $m$, it is natural to ask whether the two kinds of data we have been discussing are correlated or not. Specifically, fixing a congruence class modulo an $m$ (not divisible by 2, 5, or 11) and restricting attention only to the primes $p$ for which $a_p$ falls in that congruence class, do we still get the Sato-Tate distribution for the statistics giving the placement of $\frac{a_p}{\sqrt{p}}$ in the interval $(-1, +1)$? We don’t yet know the answer to this.\(^{12}\)

**Part III. About the proof of Sato-Tate for the elliptic curve $E$.**

### 3. Reducing the problem to a question about analytic continuation of $L$-functions

#### 3.1. The Sato-Tate distribution.

As discussed in subsection 2.2 above we now know that the data

$$p \mapsto \cos(\theta_p) = 1/2(e^{i\theta_p} + e^{-i\theta_p})$$

\(^{12}\)But, quite recently, Michael Harris [17] has made a major stride in giving a plan of attack that may eventually establish a noncorrelation theorem of another sort: the error term statistics of two nonisogenous elliptic curves, both having multiplicative reduction at some prime, each follows the Sato-Tate prediction (as has been shown) and are noncorrelated.
associated to our elliptic curve $E : y^2 + y = x^3 - x^2$ conforms to the Sato-Tate distribution $\frac{2}{\pi} \sqrt{1 - t^2}$. That is, Theorem 2.1 formulated in section 2.2 tells us that for any continuous function $F(t)$ on the interval $[-1, +1]$, the limit

$$\lim_{C \to \infty} \frac{1}{\pi(C)} \sum_{p \leq C} F(\cos \theta_p)$$

exists and is equal to the integral $\frac{2}{\pi} \int_{-1}^{+1} F(t) \sqrt{1 - t^2} dt$. How does one prove such a theorem?

To express our expected distribution in terms of the $\theta_p$‘s, one could make the change of variables ($t \mapsto \cos \theta$)

$$\frac{2}{\pi} \int_{-1}^{+1} F(t) \sqrt{1 - t^2} dt = \frac{1}{\pi} \int_{-\pi}^{+\pi} F(\cos \theta) \sin^2 \theta d\theta;$$

i.e., expressing things in terms of $\theta$ we get a “sine-squared” distribution. Here is what the data looks like in these terms:

![Figure 3.1](https://example.com/figure3.1)

**Figure 3.1.** The horizontal axis is the interval $0 \leq \theta \leq \pi$, segmented into subintervals. The height above a subinterval is proportional to the percentage of primes $p < 10^6$ that have the property that $\theta_p$ lies in the given subinterval.

The rest of this article is devoted to saying some things about the proof (see also [53] and Serre’s letter to Shahidi [44] and comments in [42]). To prove the theorem, it would be enough, thanks to the Weierstrass approximation theorem, to show Theorem 2.1 true for all real-valued polynomial functions $F(t)$, and since our task is linear, we could concentrate on proving this for $F(t) = \text{all the powers of the variable } t$, i.e.,

$$1, t, t^2, t^3, \ldots,$$

or, for that matter, it would suffice to prove it for $F(t) = \text{any other } \mathbb{R}\text{-basis of the ring of real-valued polynomials.}^{13}$

---

13As mentioned in the discussion related to Conjecture 2.2, this luxury—of proving things for a dense basis—is not yet quite enough if we aim to prove the finer rate-of-convergence result formulated by that question.

For some explicitness in our application of the Weierstrass approximation theorem for the continuous function $F$, we might make use, for example, of the (S.N.) Bernstein polynomials...
3.2. Bases for the ring of polynomials. Write the variable $t$ as a sum $\alpha + \alpha^{-1}$ so that any polynomial in $t$ (with, e.g., real coefficients) is a polynomial in $\alpha$ and $\alpha^{-1}$ invariant under the interchange $\alpha \leftrightarrow \alpha^{-1}$, and conversely: any polynomial in $\alpha$ and $\alpha^{-1}$ invariant under the above interchange is a polynomial in $t$. Consider, then, these polynomials (let’s call them symmetric power polynomials):

\[
\begin{align*}
 s_0 &= 1 \\
 s_1 &= \alpha + \alpha^{-1} \\
 s_2 &= \alpha^2 + 1 + \alpha^{-2} \\
 s_3 &= \alpha^3 + \alpha^1 + \alpha^{-1} + \alpha^{-3} \\
 s_4 &= \alpha^4 + \alpha^2 + 1 + \alpha^{-2} + \alpha^{-4} \\
 s_5 &= \alpha^5 + \alpha^3 + \alpha^1 + \alpha^{-1} + \alpha^{-3} + \alpha^{-5}
\end{align*}
\]

(3.1)

... which, when expressed as polynomials in $t$ look like:

\[
\begin{align*}
 s_0 &= 1 \\
 s_1 &= t \\
 s_2 &= t^2 - 1 \\
 s_3 &= t^3 - 2t \\
 s_4 &= t^4 - 3t^2 + 1 \\
 s_5 &= t^5 - 4t^3 + 3t
\end{align*}
\]

(3.2)

... where $s_n$ is a monic polynomial in $t$ of degree $n$ (they are the Chebychev polynomials of the second kind). They form a basis, as does any collection of products

\[\{s_n s_m\}_\mathcal{I}\]

where $\mathcal{I}$ is a collection of pairs of nonnegative integers such that the sums $n + m$ run through all nonnegative numbers with no repeats.

Here is an elementary calculus exercise:

**Proposition 3.1.** If $F(t) = s_n(t)s_m(t)$ with $n \neq m$, then

\[
\frac{2}{\pi} \int_{-1}^{+1} F(t) \sqrt{1 - t^2} dt = 0.
\]

**Corollary 3.2.** Theorem 2.1 would follow if for every positive integer $k$ there is a pair of distinct nonnegative integers $(n, m)$ with $n + m = k$ and such that

\[
\lim_{C \to \infty} \frac{1}{\pi(C)} \sum_{p \leq C} s_m(\cos \theta_p)s_n(\cos \theta_p) = 0.
\]

A colloquial way of expressing the existence and vanishing of the above limit is to say: the mean value of the quantities $s_m(\cos \theta_p)s_n(\cos \theta_p)$ is zero.

defined (for $n \geq 0$) as

\[
P_{F,n}(t) := \frac{1}{2^n n!} \sum_{k=-n}^{n} F\left(\frac{n+k}{2n}\right) \binom{2n}{n+k} (1 + t)^{n+k}(1 - t)^{n-k},
\]

for this family of degree $n$ polynomials, $P_{F,n}$ tend uniformly to $F$ on the interval $[-1, +1]$. 

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use
But how can we show that such mean values exist and vanish? The standard strategy—in fact, it seems, the only known strategy—is to invoke $L$ functions. So we turn to:

3.3. $L$-functions. To study $p \mapsto \theta_p$ effectively it is a good idea to “package this data” into complex analytic functions (Dirichlet series) whose behavior will tell us about the limits described in Corollary 3.2.

Let us do this. For any choice of prime number $p$ different from 11 and for any pair of nonnegative numbers $0 \leq m \leq n$, define the local factor at $p$ of the $L$-function $L_{m,n}(s)$ as follows:

$$L_{m,n}^{\{p\}}(s) := \prod_{j=0}^{m} \prod_{k=0}^{n} (1 - e^{i(m+n-2j-2k)\theta_p s})^{-1}.$$

If $m$ (or $n$) is zero, the factors in “$$\prod_{j=0}^{m}$$” (or “$$\prod_{k=0}^{n}$$”) don’t amount to much, so, for example:

$$L_{0,n}^{\{p\}}(s) := \prod_{k=0}^{n} (1 - e^{i(n-2k)\theta_p s})^{-1}.$$

Now form the infinite product over all prime numbers $p$:

$$L_{m,n}(s) := \prod_{p \neq 11} L_{m,n}^{\{p\}}(s)$$

and expand this to get a Dirichlet series

$$L_{m,n}(s) = \sum_{r=0}^{\infty} a_{m,n}(r) r^{-s}.$$

Here we rely on analytic number theory in the form of a classical theorem of Ikehara, which gives us that if we know enough analytic facts about these Dirichlet series $\sum a_{m,n}(r) r^{-s}$, we can control limits of the form

$$\lim_{C \to \infty} \frac{\sum_{p<C} a_{m,n}(p)}{\pi(C)};$$

i.e., since $a_{m,n}(p) = s_m(\cos \theta_p) s_n(\cos \theta_p)$, these are exactly the limits we are interested in.\(^{16}\)

---

\(^{14}\)As mentioned, one can establish the distribution of values of our error terms once we know—for some basis $\{F_i(t)\}$, ($i = 1, 2, \ldots$) of the vector space of polynomials—the mean values of the quantities $F_i(\cos \theta_p)$ for all $i$. The basis we chose to work with in Corollary 3.2 has to do with the $L$-functions that will be available to us. Another way of dicing the problem, as mentioned to me by Andrew Granville, uses the basis of polynomials in $t = \alpha + \alpha^{-1}$ given by $F_1(t) = \alpha^{\nu} + \alpha^{-\nu}$, allowing us to conclude that the Sato-Tate conjecture for our data is equivalent to the statement that, for each $\nu > 0$, the mean values of the quantities $a_{\nu^p}/p^{\frac{\nu}{2}}$ are zero, where the $a_{\nu^p}$ are the $p^\nu$-th Fourier coefficients of the cuspidal modular form of level 11 and weight two introduced in section 2.1 above.

\(^{15}\)These are the local factors of the Hasse-Weil $L$-function associated to the symmetric $m$-th power tensored with the symmetric $n$-th power of the fundamental Galois representation $\rho$ of our elliptic curve. If these symmetric powers of $\rho$ are automorphic, an issue we shall discuss later, then $L_{m,n}(s)$ would be (up to some elementary factors) the $L$-function attached to the pair of corresponding automorphic representations.

\(^{16}\)For a related discussion see [39].
Proposition 3.3. Let $m < n$. If $L_{m,n}(s)$ extends to a meromorphic function on the entire complex plane, holomorphic on the right half-plane $\Re(s) \geq 1$ and nonzero on all points $\Re(s) \geq 1$ other than $s = 1$, then
\[
\lim_{C \to \infty} \frac{1}{\pi(C)} \sum_{p \leq C} s_m(\cos \theta_p)s_n(\cos \theta_p) = 0.
\]

If, by the way, $L_{m,n}(s)$ extended to a meromorphic function on the entire complex plane, holomorphic and nonzero on $\Re(s) \geq 1$ except for having a pole of order $k$ at $s = 1$ (which it does not), the analytic proposition above\(^{17}\) would tell us that the limit is $k$ rather than 0.

3.4. Sato-Tate and the Generalized Riemann Hypothesis. It is striking that—upon assuming $L_{m,n}(s)$ extends to an entire function on the complex plane, satisfying a functional equation as expected—a proof of Conjecture 2.2 for the polynomial $F_{n,m}(t) := s_m(t)s_n(t)$ would imply the Generalized Riemann Hypothesis for the Dirichlet series $L_{m,n}(s)$. The proof of the implication (which is mutatis mutandis the proof of this same statement for $L_{0,1}(s)$ as given in the article of Akiyama and Tanigawa [2]) is briefly as follows (and we assume below that $n \neq m$). Noting that, under our initial hypothesis,
\[
\log L_{m,n}(s) = \sum_p \{ \sum_{j=0}^{m} \sum_{k=0}^{n} e^{i(n+m-2j-2k)\theta_p} \} p^{-s} + A(s) = \sum_p F_{n,m}(\cos \theta_p)p^{-s} + A(s),
\]
where $A(s)$ is holomorphic in the right half-plane $\Re(s) > \frac{1}{2}$, GRH for $L_{m,n}(s)$ will follow if we show holomorphicity of $\sum_p F_{n,m}(\cos \theta_p)p^{-s}$ for $\Re(s) > \frac{1}{2}$. A partial summation argument gives:

Lemma 3.4. If for any positive $\epsilon$, $\sum_{p < C} F_{n,m}(\cos \theta_p)$ is $O(C^{\frac{1}{2}+\epsilon})$, then
\[
\sum_p F_{n,m}(\cos \theta_p)p^{-s}
\]
converges to yield a holomorphic function in the region $\Re(s) > \frac{1}{2}$.

Proof. For $k = 1, 2, \ldots$ set $a_k := F_{n,m}(\cos \theta_p)$ if $k = p$ is a prime number, and otherwise set $a_k := 0$. So our Dirichlet series is now denoted $\sum_k a_k k^{-s}$, and we have (for any positive $\epsilon$) \( \sum_{k \leq N} a_k = O(N^{\frac{1}{2}+\epsilon}) \) for any $N$. Partial summation gives
\[
\sum_{k < N} a_k k^{-s} = \sum_{k < N} a_k \cdot N^{-s} - \sum_{n < N} \{ \sum_{k < n} a_k \} \cdot \{(n+1)^{-s} - n^{-s}\}.
\]
The first term on the right-hand side of this equation is bounded by $N^{\frac{1}{2}+\epsilon-\epsilon}$, which, if $\Re(s) > \frac{1}{2}$, is bounded independent of $N$ for an appropriate choice of $\epsilon$. Moreover, since $|(n+1)^{-s} - n^{-s}| \leq n^{-s-1}$ the second term is bounded by $\sum_n n^{\frac{1}{2}+\epsilon-\epsilon-1}$, which again, if $\Re(s) > \frac{1}{2}$, is bounded independent of $N$ for an appropriate choice of $\epsilon$. \( \square \)

It remains, then, to show the following.

---

\(^{17}\)For a concise expository summary of variant hypotheses that might be considered in the above proposition yielding a similar conclusion, see Nick Katz’s MSRI lecture (available on the MSRI website). Also see [42], IA.2; [39]; and [8] (specifically, Theorem 2.1.4 in Chapter II ("la Méthode de Hadamard-De La Vallée-Poussin") of Deligne’s paper) for further material relevant to this discussion. For a general reference on Tauberian theorems, of which these propositions are examples, see [31].
Proposition 3.5. Let \( m \neq n \). Assume that \( L_{m,n}(s) \) extends to an entire function on the complex plane and satisfies the expected functional equation. Assume, furthermore, that Conjecture 2.2 holds for the polynomial \( F_{m,n}(t) \). Then \( L_{m,n}(s) \) satisfies the Generalized Riemann Hypothesis; i.e., all its zeroes lie on the line \( \Re(s) = \frac{1}{2} \).

Proof. Assuming Conjecture 2.2 we have

\[
\Delta_{F_{m,n}}(C) := \frac{1}{\pi(C)} \sum_{p \leq C} F_{m,n}(\cos \theta_p) - \frac{2}{\pi} \int_{-1}^{+1} F_{m,n}(t) \sqrt{1 - t^2} dt < C^{-\frac{1}{2} + \epsilon}
\]

for \( C \) sufficiently large. Since the integral vanishes \(((m, n) \neq (0, 0))\) and \((\text{for any } \delta > 0) \pi(C) \geq C^{1-\delta} \) for \( C >> 0 \), we get that

\[
\left| \sum_{p \leq C} F_{m,n}(\cos \theta_p) \right| < C^{\frac{1}{2} + \epsilon}
\]

for \( C \) sufficiently large, and our proposition follows from Lemma 3.4. \( \square \)

3.5. Meromorphic extension of \( L \)-functions. But, returning to our discussion of Sato-Tate, how can we get that Dirichlet series such as \( L_{m,n}(s) \) extend meromorphically to the entire complex plane, and how can we determine the nature of their poles? A standard strategy—in fact, it seems, the only one of two known strategies—is to connect these \( L \)-functions with automorphic forms. The other strategy is closely related—and is only nominally different—and relies directly upon Poisson summation. This latter method was used by Riemann, then extended by a number of mathematicians, including Hecke, to deal with abelian \( L \)-functions and from that to construct automorphic forms of complex multiplication, and this method too will play a (key!) role in the proof, only later.

4. Replacing the Problem of Analytic Continuation of \( L \)-functions by Questions about Automorphic Forms

4.1. The reciprocity “divide”. Consider these two species of mathematical objects:

- Quadratic field extensions of the field of rational numbers, i.e., \( \mathbb{Q}(\sqrt{d})/\mathbb{Q} \) for square-free integers \( d \), and
- Functions \( \chi : \mathbb{Z} \rightarrow \{0, \pm 1\} \) that are multiplicative, i.e. \( \chi(m \cdot n) = \chi(m) \cdot \chi(n) \), nontrivial, and “of congruence type” in the sense that there is some positive integer \( N \) such that \( \chi(a) \) depends only on \( a \mod N \) (for all \( a \)).

To truly understand the first of these structures, the quadratic number fields, surely we should know the splitting properties of prime numbers in these fields; i.e., we should know, for any prime number \( p = 2, 3, 5, 7, 11, \ldots \) , whether

- the ideal generated by \( p \) is a prime ideal in the ring of integers of \( \mathbb{Q}(\sqrt{d}) \),
- the ideal generated by \( p \) splits into a product of two distinct prime ideals \( (p) = PP \) in the ring of integers of \( \mathbb{Q}(\sqrt{d}) \), or
- the ideal generated by \( p \) is the square of a prime ideal \( (p) = P^2 \) in the ring of integers of \( \mathbb{Q}(\sqrt{d}) \),

these being the only three things that can happen to the ideal generated by \( p \) in the ring of integers of \( \mathbb{Q}(\sqrt{d}) \).

Let us say that a quadratic number field \( \mathbb{Q}(\sqrt{d}) \) and a character \( \chi \) with the properties listed above are linked if
• \( \chi(p) = -1 \) if and only if \( p \) is a prime ideal in the ring of integers of \( \mathbb{Q}(\sqrt{d}) \),
• \( \chi(p) = +1 \) if and only if the ideal generated by \( p \) splits into a product of two distinct prime ideals \((p) = P\bar{P}\) in the ring of integers of \( \mathbb{Q}(\sqrt{d}) \), and
• \( \chi(p) = 0 \) if and only if the ideal generated by \( p \) is the square of a prime ideal \((p) = P^2\) in the ring of integers of \( \mathbb{Q}(\sqrt{d}) \).

So, \( \chi \) is linked to \( K \) if \( \chi \) provides us with complete information about the splitting properties of primes in the field extension \( K/\mathbb{Q} \). Of course, given a quadratic number field \( K \), we can simply construct a multiplicative function, \( \chi_K \), of \( \mathbb{Z} \) with the properties listed in the three bullets above, and the only serious issue is: does the character \( \chi_K \) we have constructed by those rules also have the congruence property? The answer to this is, in fact, yes and goes back to Gauss, it being a consequence of the quadratic reciprocity theorem (whence the title of this subsection).

The \( \chi_K \)'s we have just described are the simplest examples of automorphic representations. One of the goals of the Langlands program is to establish a vast generalization of this type of linkage, where two quite distinct species of mathematical objects are under consideration:
• a number-theoretic structure (such as the quadratic fields of the example just discussed or the sample problem in Part 1 of this article),
• automorphic representations,

and where the type of linkage one envisions is as follows: each member of either of the two species of mathematical objects alluded to above provides, in a natural way, certain numerical data (typically this data takes the form of a function on primes, such as in the example given above). A specific number-theoretic structure and a specific automorphic representation are considered linked if they provide the same numerical data. We will say a bit more about what “number-theoretic structures” are being considered in this linkage in subsection 4.5 below.

Since we will be packaging this type of numerical data into \( L \)-functions, we might hint at what is afoot by mentioning that the number-theoretic structure and the automorphic function are considered linked if they produce (via their respective data) the same \( L \)-function. In specific contexts considered by the Langlands program, if one can establish such a link, one sometimes obtains, as reward, the analytic continuation of the \( L \)-function attached to the corresponding number-theoretic structure alluded to in the bullet above.

4.2. Automorphic representations, automorphic forms. Here I will try to write things that are useful to people not in this specific field so that they might get a sense—admitting a trail of black boxes—of the thread of ideas that led to the recent work on Sato-Tate. For the purposes of this discussion, only the most salient aspects of the type of automorphic form involved in this story will be discussed below. We will be using the phrase Hecke operators with no explanation, but hope that for the moment it is sufficiently evocative and that readers for whom this notion is unfamiliar will go to the literature to seek out the story that I am omitting. A good start would be Diamond and Shurman’s text [9].

I want to say why there are two phrases, automorphic representations and automorphic forms, in the title of this subsection.

Suppose you are faced with \( G \), some group (a Lie group perhaps) acting smoothly on \( M \), some manifold (a homogenous space for the group perhaps). Then whenever

\[ 18 \text{Their official technical name being: quadratic Dirichlet characters over } \mathbb{Q}. \]
you have a function on $M$ or a differential form, $\omega$, on $M$ (or, more generally, a section of any vector bundle over $M$ that admits a compatible action of $\mathcal{G}$), you can use $\omega$ to construct a representation $V$ of the group $\mathcal{G}$ by simply considering the vector space generated by all translates of $\omega$ by elements of the group; you might also pass to the completion of this vector space with respect to some natural metric if there is such and if you want to do that. You, of course, have the option of studying the $\mathcal{G}$-representation space $V$ "abstractly", but you also have a "model" for this representation of $\mathcal{G}$ (e.g., as a space of functions, or differential forms, etc.) which may prove to be useful; even better: you have a certain preferred vector in your representation space, namely the $\omega$ that you started with. The groups $\mathcal{G}$ that are relevant for the discussion of the previous subsection will have as connected component the Lie group $\text{GL}^+_{n+1}(\mathbb{R})$ for $n = 1, 2, 3, \ldots$ (where the $+$ means positive determinant), and the manifold $M$ on which $G$ acts will often have, as connected components, the $\text{GL}^+_{n+1}(\mathbb{R})$ homogenous space $\text{GL}^+_{n+1}(\mathbb{R})/\text{SO}_{n+1} \cdot \mathbb{R}^+$, this being the space of right cosets with respect to the group generated by rotations (i.e., elements of $\text{SO}_{n+1}$) and positive homotheties (i.e., positive scalar matrices). Our automorphic forms will also be required to behave well with respect to the action of a discrete group on $M$, often a discrete subgroup of $\mathcal{G}$ viewed as acting on the right coset space $\text{GL}^+_{n+1}(\mathbb{R})/\text{SO}_{n+1} \cdot \mathbb{R}^+$ via left-multiplication.

We will focus most of our attention on our specific sample problem

$$p \mapsto e^{\pm \theta p},$$

as discussed throughout Part II, and on its "symmetric powers",

$$p \mapsto \{e^{-n\theta p}, e^{-(n-2)\theta p}, e^{-(n-4)\theta p}, \ldots, e^{(n-4)\theta p}, e^{(n-2)\theta p}, e^{n\theta p}\};$$

and we shall be treating each (small value of) $n$ separately and discussing, very briefly, the relationship between the data and automorphy.

- When $n = 0$, the data above just boils down to

$$p \mapsto 1,$$

and this data indeed corresponds to an automorphic form on $\text{GL}_1$, but it plays quite a special role in our proceedings, since its $L$-function is none other than the Riemann zeta-function.

- When $n = 1$, we view the complex upper half-plane $\mathbb{H} = \{z = x + iy \mid y > 0\}$ as a homogeneous space under the action of the group $\text{GL}(2, \mathbb{R})$ via the usual formulas:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az + b}{cz + d}.$$  

The symmetric 1-st power of our data (i.e., our data) is cuspidal automorphic, since there is a holomorphic differential form $\omega = \omega(z)$ on $\mathbb{H}$—linked to our data in a way that we shall mention below—having the following properties: for some positive number $N$ the differential form $\omega$ is invariant under the action of the group, usually denoted $\Gamma_1(N)$, of all matrices of determinant one of the form

$$\begin{pmatrix} a & b \\ Nc & d \end{pmatrix}$$

19This "invariance property" is analogous to the congruence property that the quadratic characters $\chi_K$ possess, as discussed in subsection 4.1.
with \(a, b, c, d\) rational integers and \(a \equiv d \equiv 1\) modulo \(N\). The way in which \(\omega\) is “linked to our data” is that \(\omega\) is an eigenvector under the action of the \(p\)-th Hecke operator with eigenvalue \((e^{-\theta_p} + e^{\theta_p})\sqrt{p} = s_1(e^{-\theta_p} + e^{\theta_p})\sqrt{p}\) for all but finitely many primes \(p\).

We can take \(N = 11\), and there is such an \(\omega\) invariant even under the slightly larger group \(\Gamma_0(11)\) (defined as above but where one does not require \(a \equiv d \equiv 1\) modulo 11). In fact, as a function on the upper half-plane \(z = x + iy\) \((y > 0)\)

\[
\omega = 2\pi i \prod_{\nu \geq 1} (1 - e^{2\pi i \nu z})^2 (1 - e^{22\pi i \nu z})^2 dz = 2\pi i \sum_{n=1}^{\infty} a_n e^{2\pi i n z} dz,
\]

this being a Fourier series that we have already fleetingly referred to in subsection 2.1.\(^20\) The requirement of cuspidality is that the differential form \(\omega\) has sufficiently good behavior as one goes to the points at infinity in the quotient Riemann surface \(\mathbb{H}/\Gamma_1(11)\) so that it extends to a regular differential form on the natural compactification of that Riemann surface.

- When \(n = 2\), the symmetric 2-nd power of our data is **cuspidal automorphic**, since there is a (real analytic) differential 2-form \(\omega_2\) on the homogeneous space \(GL_3(\mathbb{R})/SO_3 \cdot \mathbb{R}^+\) enjoying, as in the previous case, an appropriate (“in”)variance property with respect to an appropriate discrete group; moreover the differential 2-form \(\omega_2\) is an eigenform, with specific eigenvalues, of a certain ring of differential operators, this being the feature analogous to “holomorphicity” discussed in the case \(n = 1\) (see [14]). Furthermore, our \(\omega_2\) exhibits good behavior as one goes to infinity in the symmetric space, and finally the all-important **link to our data** is that for all but finitely many primes \(p\), the differential form \(\omega_2\) is also an eigenvector under the action of certain correspondences (Hecke operators related to \(p\)) and with prescribed eigenvalues related to \((e^{-2\theta_p} + 1 + e^{-2\theta_p}) \cdot p = s_2(e^{-\theta_p} + e^{\theta_p}) \cdot p\).

- Similarly for \(n = 3\) (see [26]).\(^21\)
- Similarly for \(n = 4\) (see [25]).

What happens for \(n \geq 5\)? One has, at the present moment, a somewhat weaker automorphy result (potential automorphy for even \(n\); see subsection 4.6) which is sufficient to establish the Sato-Tate result that this article is discussing (see Corollary 4.2).

The connection between cuspidal automorphy and the desired behavior of the \(L\)-functions we care about is:

**Proposition 4.1.** If for two unequal nonnegative integers \(n\) and \(m\) the symmetric \(n\)-th power of our data and the symmetric \(m\)-th power of our data are both cuspidal

\(^20\) We rigged our sample problem to be given by the elliptic curve that is the quotient of the upper half-plane under the action of \(\Gamma_1(11)\). Thanks to the work on modularity due to Wiles, Taylor-Wiles et al., we could have chosen any elliptic curve over \(\mathbb{Q}\) as well and still enjoy the fact that the symmetric 1-st power of the corresponding data (in short, the “data” itself) is automorphic.

\(^21\) Some of the relevant history of this, and part of the history of Proposition 4.1, is recorded in the introduction of [26], where it is explained that the automorphy of the symmetric cube of a \(GL_2\) representation, a project of Shahidi’s since 1978, following upon Langlands’ work on Eisenstein series ([35], [36]), led Shahidi to develop a machinery [47], [48], [50], all of which is used to prove a (Langlands) functoriality result for \(GL_2 \times GL_3\) and from this to deduce automorphy of the symmetric cube of automorphic forms on \(GL_2\).
automorphic, then $L_{n,m}$ extends to a holomorphic function on the entire complex plane, nonzero on the line $\Re(s) = 1$ (for $z \neq 1$).

In particular, taking $m = 0$ and $n > 0$, one has that if the symmetric $n$-th power of our data is cuspidal automorphic, then $L_n$ extends to a holomorphic function on the entire complex plane. See [46], [49] for proof of holomorphicity and meromorphicity of various symmetric powers.

This proposition in itself is a great piece of mathematics, which when $n$ and $m$ are nonzero involve either

- a method of Langlands and Shahidi (see [46], [27]) where one uses Langlands’ theory of Eisenstein series\(^{22}\) or
- a method of Rankin and Selberg (developed in the context of pairs of automorphic forms for $GL_n$ and $GL_m$ by Jacquet, Piatetski-Shapiro, and Shalika [20] and completed by the publication of [4]).

To see how automorphy might help one to control $L$-functions, consider the special case of $(n, m) = (0, 1)$ of our sample problem and recall the integral expression for the $L$-function $L_{0,1}(s)$ valid for $\Re(s)$ large enough, namely:

$$
\frac{\Gamma(s)}{(2\pi)^s} L_{0,1}(s) = \int_{y=0}^{y=\infty} y^{s-1} \omega(iy) = \int_{y=\sqrt{-1}}^{y=\infty} y^{s-1} \omega(iy) + \int_{y=0}^{y=\sqrt{-1}} y^{s-1} \omega(iy),
$$

where $\omega$ is the differential 1-form discussed previously.

Here the first integral on the right side, i.e., $\int_{y=\sqrt{-1}}^{y=\infty} y^{s-1} \omega(iy)$, has an integrand

$$
y^{s-1} \omega(iy) = 2\pi i \sum_{n=1}^{\infty} a_n e^{-2\pi n y} y^s dy/y,
$$

which goes to zero essentially exponentially as $y$ tends to $\infty$. Therefore this integral converges to an entire function of $s$. The second integral is the troublemaker, for naive estimates will not work to show convergence. Nevertheless, since (miracle!) the differential form $\omega$ is an eigenform for the transformation $z \mapsto -\frac{1}{11} z$, i.e., for the action of the matrix

$$
\begin{pmatrix}
0 & -1 \\
11 & 0
\end{pmatrix}
$$

on the upper half-plane $\mathbf{H}$, it follows that the second integral is easily expressible in terms of the first integral, so the sum of the two integrals on the right-hand side—that is, the $L$-function $L_{0,1}(s)$ decorated by $\frac{\Gamma(s)}{(2\pi)^s}$—converges to an entire function. The essence of this type of proof goes all the way back to Riemann’s famous 1859 article [40].

An example, then, of what would suffice to achieve the Sato-Tate Conjecture for our data is the following corollary of the past work cited and of Proposition 4.1:

**Corollary 4.2.** If for every odd value of $m$ greater than or equal to 7 the symmetric $m$-th power of our data is cuspidal automorphic, then the Sato-Tate conjecture holds for our data.

\(^{22}\)To be a bit more specific, one views $GL_n \times GL_m$ as a Levi component in a parabolic subgroup of $GL_{n+m}$ and relates $L_{n,m}$, initially defined only in some right half-plane, to the constant term of certain Eisenstein series on $GL_{n+m}$. See the three lectures of the Langlands-Shahidi method in [51]. The nonvanishing on $\Re(s) = 1$ is shown in [46].
As readers will see in subsection 4.6 below, somewhat weaker hypotheses will also suffice, and this is a lucky thing.

Proof. We would then have that \( L_{n,m}(s) \) is entire for
\[
(n, m) = (0, 1), (0, 2), (1, 2), (1, 3), (2, 3), (2, 4), (3, 4),
\]
and for \((0, m)\) and \((1, m)\) ranging through all positive odd integers \(m \geq 7\) (here we depend on the earlier work cited to cover \(m < 7\)). The theorem then follows from the previous propositions and Corollary 3.2.

To show the cuspidal automorphy of all the symmetric \(m\)-th powers of our data that are required by Corollary 4.2, it seems that we must, at least at present, connect this data with Galois representations. So we now turn to:

4.3. Galois representations associated to the symmetric \(m\)-th powers of our data. Our elliptic curve \(E\), which we’ve focussed on to provide us with our “sample problem”, whose equation in the finite plane is given by
\[
y^2 + y = x^3 - x^2,
\]
is a commutative algebraic group (the point at infinity playing the role of origin). Therefore, for any positive integer \(N\) we may consider the kernel of multiplication by \(N\) in \(E\), and this subgroup of \(E\) we will denote \(E[N]\).

Working with the points of \(E\) whose coordinates lie in an algebraic closure of \(\mathbb{Q}\), the subgroup \(E[N]\) consists of those points on the algebraic group \(E\) of order dividing \(N\). This group is, on the one hand, a product of two cyclic groups of order \(N\), and on the other hand, if we adjoin to the rational field \(\mathbb{Q}\) all the dehomogenized coordinates of the (finitely many) points of \(E[N]\), we obtain a finite Galois field extension of \(\mathbb{Q}\)—denote it \(K_N/\mathbb{Q}\)—but we also get, along with the field extension itself, a natural injection of \(\text{Gal}(K_N/\mathbb{Q})\) into the automorphism group of \(E[N]\) (via the natural action of the Galois group on the coordinates of the points in \(E[N]\). Since the automorphism group of a product of two cyclic groups of order \(N\) is isomorphic to \(\text{GL}_2(\mathbb{Z}/N\mathbb{Z})\), we emerge from this discussion with quite a beautiful structure. Namely, given our elliptic curve \(E\) we get for every positive integer \(N\) a Galois field extension \(K_N/\mathbb{Q}\) and a two-dimensional representation of its Galois group over the ring \(\mathbb{Z}/N\mathbb{Z}\). Viewing that Galois group as a quotient of the full profinite Galois group \(G\) of the algebraic closure of \(\mathbb{Q}\) over \(\mathbb{Q}\), we may consider this information to be equivalent to having representation
\[
\rho_{E,N} : G \to \text{GL}_2(\mathbb{Z}/N\mathbb{Z}),
\]
the kernel of which restricts to the identity on \(K_N\). Since these \(\rho_{E,N}\)’s compile well, in the sense that if \(N\) divides \(M\) the representation \(\rho_{E,N}\) is equivalent to the composition of \(\rho_{E,M}\) and the natural projection \(\text{GL}_2(\mathbb{Z}/M\mathbb{Z}) \to \text{GL}_2(\mathbb{Z}/N\mathbb{Z})\), we may pass to limits, so that, for example, for any prime number \(\ell\) taking the projective limit of the \(\rho_{E,N}\)’s for the sequence \(N = \ell^n\) (\(\nu\) tending to \(\infty\)) gives us a representation to \(\text{GL}_2(\mathbb{Z}_\ell)\) where \(\mathbb{Z}_\ell\) denotes the \(\ell\)-adic integers, and passing, then, to \(\mathbb{Q}_\ell\) we get representations
\[
\rho_{E,\ell,\infty} : G \to \text{GL}_2(\mathbb{Q}_\ell).
\]
Let \(V_{E,\ell}\) denote the two-dimensional \(\mathbb{Q}_\ell\)-vector space \(\mathbb{Q}_\ell^2\) equipped with a continuous \(\mathbb{Q}_\ell\)-linear action of \(G\) (via \(\rho_{E,\ell,\infty}\)).

\[23\] Potentially cuspidal automorphic is also enough.
The connection between these representation spaces $V_{E,\ell}$ and our “data,” i.e., the data

$$p \mapsto e^{\pm i\theta_p}$$

we have been discussing in the previous sections of this article, is quite neat: For all but finitely many primes $p$ (in fact, in this case, for $p \neq 11, \ell$) there is a well defined class of elements in $G$ (called Frobenius elements at $p$) that have the property that the action of any of these Frobenius elements at $p$ on the $G$-representation space $V_{E,\ell}$ has the same characteristic polynomial, and the roots of this common characteristic polynomial are the quadratic irrationalites $e^{\pm i\theta_p} \sqrt{p}$. The set of these Frobenius elements at $p$ is dense in $G$, and so, since the $G$-representation $V_{E,\ell}$ is irreducible, knowledge of the traces of representation of the action of the Frobenius elements at $p$, i.e., the integer-valued function

$$p \mapsto e^{i\theta_p} \sqrt{p} + e^{-i\theta_p} \sqrt{p} = N_E(p) - (p + 1)$$

for all but finitely many primes $p$ determines the representation.

It should also not escape our notice that we have here a somewhat extraordinary structure: for every prime number $\ell$ we get a two-dimensional $G$-representation space $V_{E,\ell}$ for which the Frobenius elements at $p$ (for $p \neq 11, \ell$) all have the same eigenvalues: the quadratic irrationalites $e^{\pm i\theta_p} \sqrt{p}$. We will refer to such a family, $W_{\ell}$, of $\mathbb{Q}_\ell$-vector space representations of $G$ ($\ell$ running through all prime numbers) possessing the property that the traces of Frobenius elements at $p$ for all but finitely many $p$ are integers independent of $\ell$, as a compatible family of Galois representations.

Of course, for any nonnegative integer $n$, if we take the $n$-th symmetric power of the vector space $V_{E,\ell}$, denote it $\text{Symm}^n(V_{E,\ell})$, and endow it with its induced $G$-action, then the Frobenius elements at $p$ (for $p \neq 11, \ell$) will act on $\text{Symm}^n(V_{E,\ell})$ with eigenvalues

$$e^{n\theta_p} p^{n/2}, e^{(n-2)\theta_p} p^{n/2}, \ldots, e^{-(n-2)\theta_p} p^{n/2}, e^{-n\theta_p} p^{n/2},$$

i.e. with eigenvalues (up to normalization) equal to what we’ve been referring to as the $n$-th symmetric power of our data. In particular, for every positive integer $n$ the $\text{Symm}^n(V_{E,\ell})$ (with $\ell$ running through all prime numbers) is also a compatible family of Galois representations.

### 4.4. Digression on compatible families and Galois characters.

The general notion we have just been considering, of compatible families of Galois representations, is as surprising and elegantly intricate a mathematical concept as—luckily for us—it is ubiquitous. We were working in the previous subsection with representations of $G = G_\mathbb{Q}$, the Galois group of the algebraic closure of $\mathbb{Q}$ over the rational field $\mathbb{Q}$, but we might equally well study, for any number field $K$, the analogous structure, pinned down by “data” that one might call a Galois character over $K$ with values in a number field. The étale cohomology groups of algebraic varieties over number fields give plentiful examples of this kind of mathematical object, so let us briefly discuss it.

Let $K, F$ be number fields, and for $\tilde{K}$ an algebraic closure of $K$, put $G_K := \text{Gal}(\tilde{K}/K)$. Let $S$ be a finite collection of places of $K$ containing all archimedean places and $T$, similarly, a finite collection of places of $F$ containing all archimedean places.
By a **Galois character of degree** $d$ **on** $K$ **with values in** $F$ (relative to the sets of places $S$ and $T$) let us mean a function $\chi$ on the places of $K$ not in $S$ with values in $F$ that has the property that for every place $v$ of $F$ not in $T$ there exists a $d$-dimensional vector space $W_v$ over $F_v$ (where $F_v$ = the completion of $F$ at $v$) endowed with a continuous $F_v$-linear (semisimple) action of $G_K$ that is unramified for all places $w$ of $K$ that are neither in $S$ nor of the same residual characteristic as that of $v$. For each such place $w$ we require that

- the characteristic polynomial $\det(1 - \text{Frob}_w|_{W_v} x)$ of a Frobenius element $\text{Frob}_w$ at $w$ (which is, a priori, only a polynomial in $F_v[x]$) actually have coefficients in the subfield $F \subset F_v$; and, moreover, that

- the polynomial

$$\det(1 - \text{Frob}_w|_{W_v} x) = 1 - \text{Trace}_{F_v}(\text{Frob}_w|_{W_v}) \cdot x + \cdots + (-1)^d \text{Det}_{F_v}(\text{Frob}_w|_{W_v}) \cdot x^d \in F[x]$$

be independent of $v \notin T$; and finally,

- $\chi(w) = \text{Trace}_{F_v}(\text{Frob}_w) \in F \subset F_v$ for all $v \notin T$ and for $w \notin S$ and $w$ not of the same residual characteristic as that of $v$.

Since these Frobenius elements $\text{Frob}_w$ are dense in the image of $G_K$ in $\text{Aut}(W_v)$, knowledge of their traces pins down the character of the representation of $G_K$ on $W_v$, which determines up to isomorphism the representation itself, since we have assumed it to be semisimple. In summary, then, the Galois character $\chi$ over $K$ with values in $F$ determines, and is determined by, the compatible family of $G_K$-representations $\{W_v\}_{v \notin T}$ (taken up to isomorphism).\(^{24}\) One can try to deal with these Galois characters with values in number fields in a manner as close as possible to the way we deal with characters in any other aspect of representation theory.\(^{25}\) For example, the collection of Galois characters over $K$ with values in number fields in $Q$ forms a $\lambda$-ring in the usual sense of representation theory.

Say that a Galois character with values in a number field corresponding to the compatible family $\{W_v\}_{v \notin T}$ is **irreducible** if **every** $W_v$ is irreducible as $G_K$-representation ($v \notin T$).

Galois characters of small degree with values in number fields have an immensely rich history. The study of Galois characters of degree 1 are treated by Class Field Theory (and cf. [42]). The $\chi_K$ already encountered in the exposition above (subsection 4.1) are, in fact, Galois characters of degree 1 over $Q$ with values in $Q$, where the associated compatible family $\{W_{\ell}\}$ of $G_Q$ representations is somewhat atypical in that the entire family comes from a single 1-dimensional $Q$-vector space $W$ with nontrivial $G_Q$-action trivialized when restricted to $G_K$, and where for any prime $\ell$, $W_{\ell} := W \times Q Q_{\ell}$. The Galois character attached to the simplest data $p \mapsto 1$, discussed in the previous subsection, is an even more basic example ("basic", but not elementary, since its associated $L$-function is the Riemann zeta-function; one should have great respect for it).

\(^{24}\)This definition of **Galois character with values in a number field** is just a mild generalization of the concept of **strictly compatible family of rational $\ell$-adic representations** as defined in 1968 in Chapter I of Serre’s treatise [42]. See also (in loc. cit.) Serre’s list of Open Questions regarding these families of representations.

\(^{25}\)Note that if $\chi_1$ and $\chi_2$ are both Galois characters over $K$ with values in $F$ relative to $S$ (and $T$) and if they agree as functions on the complement of any finite set of places $S'$ containing $S$, they agree on $S$. As a result, for any $\chi$ we can always take $S$ and $T$ to be the minimal set of places for which $\chi$ is a Galois character over $K$ with values in $E$ relative to $S$ and $T$. In other words, one can ignore the extra clause "(relative to $S$ and $T$)" except when we wish to make those sets of places precise for a given character.
Galois characters of degree 2 with values in number fields are related to much of the classical theory of modular forms. This brings us to:

4.5. **Langlands reciprocity.** A laudable goal is to link, as far as possible, these two species of mathematical structures:

- irreducible Galois characters \( \chi \) of degree \( d \) of \( K \) with values in a number field \( F \),
- cuspidal automorphic forms \( \omega \) for \( \text{GL}(d) \) over \( K \) that are eigenforms for the appropriate Hecke operators, with eigenvalues in \( F \),

where a given Galois character \( \chi \) would be said to be linked to a cuspidal automorphic form \( \omega \) if for every place \( w \not\in S \) of \( K \) the value \( \chi(w) \) is equal to the eigenvalue of an appropriate (“Hecke”) operator attached to \( w \) acting on the form \( \omega \).

When such a thing happens for an irreducible Galois character \( \chi \) with values in a number field, we will say that the character \( \chi \) itself and also the corresponding compatible family \( \{ W_v \} \), are **cuspidal automorphic**.

In this language, going back to our data as discussed in the previous subsection, we have been asking whether

\[
\text{Symm}^n(V) := \{ \text{Symm}^n(V_\ell), \text{ for all primes } \ell \},
\]

and, equivalently, its corresponding Galois character, can be linked in this way to a cuspidal automorphic eigenform for \( \text{GL}(n+1) \) over \( \mathbb{Q} \), i.e., that they are **cuspidal automorphic**.

But, as already hinted, one can get away with a slightly more malleable notion of automorphy to establish the conjecture of Sato and Tate. Since this is crucial for the recent work, we will now say a few words.

4.6. **Potential automorphy.** Given any compatible family of Galois representations over \( \mathbb{Q} \), i.e., representations of the group \( G_\mathbb{Q} = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \), and given any finite extension \( F/\mathbb{Q} \), we can restrict our compatible family of representations to \( G_F := \text{Gal}(\overline{K}/F) \subset G_\mathbb{Q} \) to get a compatible family of Galois representations over \( F \). We might call this *lifting compatible families of Galois representations from \( \mathbb{Q} \) to \( F \)*, but there is no cause for such a high-sounding name for this evident operation, which is nothing more than “restriction”. There is a “corresponding” operation for automorphic forms that is, in contrast, far from evident. Our automorphic forms have been described as “functions with certain properties” on the symmetric spaces attached to the algebraic groups \( \text{GL}_{n+1} \), these being algebraic groups over the rational field \( \mathbb{Q} \). For every finite extension \( F/\mathbb{Q} \), we may think of \( \text{GL}_{n+1} \) as an algebraic group over \( F \), and this algebraic group has a corresponding symmetric space associated to it and which, in general, is larger in dimension than the symmetric space attached to \( \text{GL}_{n+1} \) over \( \mathbb{Q} \). A good deal of work, over decades, has been devoted to proving aspects of what is called Langlands lifting (of automorphic forms from one number field to a larger one), and every step of progress here has been hard won.

Let us say that a compatible family of Galois representations over \( \mathbb{Q} \) is **strongly potentially cuspidal automorphic** if there is some totally real number field \( K \), Galois over \( \mathbb{Q} \), such that its lifting to a compatible family of Galois representations over \( F \) is cuspidal automorphic over \( K \). (Here the adverb “strongly” reminds one that the field extension \( K/\mathbb{Q} \) over which the family of Galois representations achieves automorphy is required to be Galois and \( K \) is required to be totally real.)
To show the Sato-Tate conjecture for our data, we wish to make use of Proposition 3.3. We would be more than happy, for example, if we knew that for every odd integer \( n \geq 7 \) the \( n \)-th symmetric power of our data were cuspidal automorphic, for the corresponding \( L \)-functions would then be known to be entire and the other requisite hypotheses of Proposition 3.3 would also hold. However, Proposition 3.3 does not require analyticity, but merely meromorphicity (and the appropriate behavior on the right half-plane \( \Re(s) \geq 1 \)). To achieve this, (strong) potential automorphicity—rather than straight automorphicity over \( \mathbb{Q} \)—is enough, using a fundamental result of Langlands (insuring descent of automorphicity for appropriate solvable field extensions) coupled with a classical argument of Brauer\(^{26}\)—an argument first employed in showing meromorphicity of the nonabelian \( L \)-functions of Artin—would gain for us sufficient meromorphic (although not necessarily holomorphic) continuation of the relevant \( L \)-functions—sufficient, that is, to deduce the Sato-Tate conjecture.\(^{27}\)

There is yet another brand of malleability that is critical in the method:

4.7. Galois deformation theorems and the pivotal role played by residual representations. We have discussed how the issue of Sato-Tate is connected to the condition of certain compatible families of Galois representations (of degree \( n + 1 \), for various \( n \) and over some totally real number field \( K \), Galois over \( \mathbb{Q} \)) being cuspidal automorphic.

Now we will change gears and consider single representations rather than compatible families of them. So, let \( K \) be a number field as before, \( \ell \) a prime number, \( \mathbb{Q}_\ell \) an algebraic closure of \( \mathbb{Q}_\ell \), with \( \mathcal{O}_\ell \) its ring of integers, and \( \mathbb{F}_\ell \) the residue field (an algebraic closure of the prime field of characteristic \( \ell \)). Let \( \tilde{W} \) be a \( d \)-dimensional \( \mathbb{Q}_\ell \)-vector space with an irreducible \( \mathbb{Q}_\ell \)-linear \( G_K \)-action represented by a homomorphism

\[
\rho : G_K \rightarrow \text{GL}_d(\mathbb{Q}_\ell) \subset \text{GL}_d(\bar{\mathbb{Q}}_\ell) \cong \text{Aut}_{\mathbb{Q}_\ell} \tilde{W},
\]

the isomorphism on the right given by an appropriate choice of \( \mathbb{Q}_\ell \)-basis of \( \tilde{W} \).

Pass from \( \mathcal{O}_\ell \) to its residue field, \( \mathbb{F}_\ell \), to get the the associated residual representation

\[
\tilde{\rho}_v : G_K \rightarrow \text{GL}_d(\mathbb{F}_\ell),
\]

the semisimplification of which is uniquely determined (up to equivalence) by the equivalence class of the \( G_K \)-representation \( \tilde{W} \).

If, for a number field \( F \), there is a (strongly) potentially automorphic compatible family, \( \{W_v\}_{v \in T} \), of Galois representations over \( K \) with values in \( F \) and a prime \( v \) of \( F \) of residual characteristic \( \ell \) such that for some imbedding \( \iota : F_v \hookrightarrow \mathbb{Q}_\ell \) the base change of the representation \( W_v \) to \( \mathbb{Q}_\ell \) via \( \iota \) is equivalent to \( \tilde{W} \), then we’ll say that the \( G_K \)-representation \( \tilde{W} \) itself is (strongly) potentially automorphic.\(^{28}\)

\(^{26}\)Any character of a finite group \( G \) is a linear combination with integer (possibly negative, however) coefficients of characters induced from characters of elementary subgroups.

\(^{27}\)In fact, as a consequence of potential automorphy (and cuspidality), the above relevant symmetric power \( L \)-functions will be expressed as quotients of finite products of “standard \( L \)-functions” attached to the General Linear Group, and therefore the holomorphicity and nonvanishing of each factor for \( \Re(s) \geq 1 \) gives both holomorphicity and nonvanishing (up to the line \( \Re(s) = 1 \)) of those symmetric power \( L \)-functions.

\(^{28}\)So, for a “single” representation to be potentially automorphic, it must be (the base change to \( \mathbb{Q}_\ell \) of) a member of a compatible family of representations. This might seem like a big restriction on \( \tilde{W} \), for it is indeed a rare occurrence for irreducible \( G_K \)-representations to be the base...
It pays to consider the “inverse problem”; that is, fix a prime number $\ell$ and a specific (irreducible, say) representation

$$\tilde{r} : G_K \to \text{GL}_d(\overline{F}_\ell)$$

and consider the collection $V(\tilde{r})$ of liftings of $\tilde{r}$ to characteristic 0. That is, $V(\tilde{r}) :=$ the set of equivalence classes of Galois representations over $K$ into $d$-dimensional vector spaces $\tilde{W}$ over $\overline{Q}_\ell$, $\rho : G_K \to \text{GL}_d(\overline{Q}_\ell)$, such that their residual representations $\bar{\rho} : G_K \to \text{GL}_d(\overline{F}_\ell)$ are equivalent to $\tilde{r}$.\(^{29}\)

By a residual condition $[\text{RC}]$ we mean a condition imposed on $\tilde{r} : G_K \to \text{GL}_d(\overline{F}_\ell)$. (For example: irreducible, surjective, etc.)

By a global condition $[\text{GC}]$ we mean a condition or a number of conditions of the following sort imposed on a global representation $r : G_K \to \text{GL}_d(\overline{Q}_\ell)$: we might ask that the determinant of the representation be equal to a specified character and/or that the Galois representation be isomorphic to its dual or to the twist of its dual by a specified character or that it be skew-symmetric or that the Galois action preserve some specified tensor.

If $w$ is a place of $K$, by a local condition (at $w$) $[\text{LC}(w)]$ we mean a condition imposed on the restriction of a representation $r : G_K \to \text{GL}_d(\overline{Q}_\ell)$ to the decomposition group $G_{K_w}$ of $w$. (For example: we might ask that the restriction of $r$ to $G_{K_w}$ be unramified, etc.)

A number of theorems of the following shape have been proved, and they therefore might deserve a collective name.

**Definition 4.3.** “$A$” Galois deformation theorem is a theorem with specific hypotheses of the following form:

$$[\text{RC}], [\text{GC}], \text{ and } [\text{LC}(w)] \text{ for all places } w \text{ of } K$$

and with a conclusion of the following form:

For any residual representation $\tilde{r}$ satisfying the specified condition $[\text{RC}]$, if there is some lifting of $\tilde{r}$ to characteristic zero—i.e., element of $V(\tilde{r})$—that

1. satisfies $[\text{GC}]$, and $[\text{LC}(w)]$ for all places $w$ of $K$, and
2. is strongly potentially automorphic in the sense alluded to at the beginning of this subsection, with possible further conditions on the field $K'$ that realizes change of a representation fitting into a compatible family of representations. Nevertheless, there are conjectures [13] that suggest that the main obstruction to this happening comprises (local) conditions that the representation must satisfy when restricted to the decomposition groups of $K$ at primes dividing $\ell$.

\(^{29}\)One of the great advances during the past few years in our understanding of this structure is in the special case where $d = 2$, $K = \mathbb{Q}$ (say $\ell \neq 2$, and $\tilde{r}$ absolutely irreducible), and where complex conjugation does not act as a scalar in the representation $\tilde{r}$. It is in this context that Serre had conjectured that $V(\tilde{r})$ contains a compatible family of Galois representations that is cuspidal automorphic (for $\text{GL}_2$). Of course, a simple consequence of this is that $V(\tilde{r})$ is nonempty if $\tilde{r}$ is of the form described above. This conjecture of Serre has been settled by recent work of Khare and Wintenberger [24]. For more material consult Khare’s webpage (http://www.math.utah.edu/~shekhar/papers.html) and cf. [10]; also the proceedings of the summer school on Serre’s Conjecture held at Luminy in July 2007 will provide—when they appear—background prerequisites for appreciation of these results, as well as an exposition of the proof of Serre’s Conjecture.
the ‘potentiality’ of automorphy,\(^{30}\) then every lifting of \(\bar{r}\) to characteristic zero—i.e., element of \(V(\bar{r})\)—that satisfies (1) also satisfies (2), i.e., is strongly potentially automorphic.

The most powerful such Galois deformation theorems recently proved, with the most flexible and useful conditions [RC], [GC], and [LC(v)], are due to Mark Kisin\(^{31}\) and to Richard Taylor [54]. When such a Galois deformation theorem can be applied, we often get at our disposal large quantities of strongly potentially automorphic Galois representations, all liftings of the same residual representation, and each, of course, fitting into their own family of compatible Galois representations. This allows us to move from residual characteristic to residual characteristic, as we shall now describe.

4.8. Hopping from one prime to another. The Galois deformation theorems discussed in the previous subsection can be applied in one stage, or, at times, they can be applied iteratively in multiple stages, allowing us to hop from residual representations relative to algebraic closures of finite fields \(\bar{F}_\ell\) of different characteristics, obtaining, as corollary, strong potential automorphy for more and more Galois representations. This was already done—a single hop—moving from characteristic 3 to 5 in Wiles’ and Taylor-Wiles’ proof of Fermat’s Last Theorem. Multiple such hops (an inductive argument being in play) were at work in Khare’s orginal work [23] on Serre’s Conjecture for level 1 and also in the full proof of Serre’s Conjecture by Khare and Wintenberger. Moreover, a very elegant such hop plays a role in the recent work on the Sato-Tate Conjecture. Here is the general idea of how a “prime hop” works:

**Stage one:** You might start with \(\bar{r} : G_K \to \text{GL}_d(\bar{F}_\ell)\), for which you know that \(V(\bar{r})\) contains one lifting of \(\bar{r}\) that is strongly potentially automorphic, and then deduce that many other liftings—i.e., those satisfying the specified conditions [GC] and [LC(v)]—lift to a strongly potentially automorphic (characteristic zero) representation

\[
\rho' : G_{K'} \to \text{GL}_d(\bar{Q}_\ell),
\]

for \(K'/K\) some totally real finite extension of \(K\) such that \(K'/\mathbb{Q}\) is Galois. Since \(\rho'\) is then potentially automorphic, after passing to a possibly larger totally real field \(K''\) over \(K'\), Galois over \(\mathbb{Q}\), one gets a compatible family of associated Galois representations,

\[
\rho''_{\ell} : G_{K''} \to \text{GL}_d(\bar{Q}_\ell),
\]

for all primes \(\ell\). This alone is a powerful enough application.

**Stage two:** But for many purposes, the fact that we now have a compatible family of representations \(\rho''_{\ell}\) allows us to pass to a residual representation of \(\rho''\) with respect

---

\(^{30}\)Some Galois deformation theorems require, for example, that \(K'/\mathbb{Q}\) be Galois with the prime \(p\) split. Some variants hypothesize that the automorphic form invoked in this hypothesis satisfies certain local conditions and then obtain that the automorphic form invoked in the conclusion will satisfy the analogous local conditions.

\(^{31}\)See [28], where the [LC(v)] condition for \(v|p > 2\) is that the local representation is Barsotti-Tate; [29] deals with \(p = 2\); and [30] covers the case where the local representations for \(v|p\) become semi-stable over an abelian extension of \(\mathbb{Q}_p\). Consult Kisin’s webpage for more: http://www.math.uchicago.edu/~kisin/preprints.html.
to a prime number $\tilde{\ell}$ different from $\ell$; i.e.,

$$\rho''_{\tilde{\ell}} : G_{K''} \rightarrow \text{GL}_d(\overline{\mathbb{F}}_{\tilde{\ell}}),$$

which, if $\rho''_{\tilde{\ell}}$ satisfies $[\text{RC}]$, would again be a candidate for a further application of the Galois deformation theorems, since it lifts to a potentially automorphic Galois representation.

4.9. A rich source of potentially automorphic Galois representations. To summarize our discussion up to this point, we have that the error term of our sample problem

$$p \mapsto e^{\pm i\theta_p}$$

will be shown to satisfy the Sato-Tate distribution if the compatible family of Galois representations attached to the $n$-th symmetric power of $V$ is shown to be potentially automorphic, for all odd values of $n$. This, in turn, could be demonstrated if the following list of requirements is met:

- if we have a good Galois deformation theorem requiring residual, local and determinantal conditions, appropriate for what will be required of it below; and
- if, for each odd positive integer $n$, we can find a prime number $\ell$ for which the residual representation attached to $\text{Symm}^n(V_\ell)$ can be shown to satisfy the residual condition, the Galois representations themselves and the characteristic zero Galois representations $\text{Symm}^n(V_\ell)$ satisfy the local and determinantal conditions; and
- if, for each of the residual representations in the previous bullet, we can find a lifting to characteristic zero satisfying the local and determinantal conditions that is potentially automorphic.

It is at this point that the important family of hypersurfaces

$$Y_t : \quad X_0^{n+1} + X_1^{n+1} + \cdots + X_0^{n+1} = (n + 1)tX_0X_1 \ldots X_n$$

($n$ even) comes to play its role ([18]). This family has the property that for appropriate values $t_o$ of $t$ in totally real number fields $F_o$, Galois over $\mathbb{Q}$, there is a compatible family of $n + 1$ dimensional representations

$$W^{(n)}_{t_o} = \{W^{(n)}_{t_o,\ell}\} \text{ (for all primes } \ell)$$

of $G_{F_o}$ occurring as subquotients of the compatible family of Galois representations attached to the middle dimensional cohomology of $Y_{t_o}$ such that there is a choice of primes $\ell_1$ and $\ell_2$ and a Galois deformation theorem with these properties:

- the representations $W^{(n)}_{t_o,\ell_1}$, $W^{(n)}_{t_o,\ell_2}$ and $\text{Symm}^n(V_{E,\ell_1})$ (the latter when restricted to $G_{F_o}$) all satisfy the determinantal and local conditions of the Galois deformation theorem,
- the residual $G_{F_o}$ Galois representations attached to $W^{(n)}_{t_o,\ell_1}$ and $W^{(n)}_{t_o,\ell_2}$ satisfy the residual conditions of the Galois deformation theorem,
- the residual $G_{F_o}$ Galois representations attached to $W^{(n)}_{t_o,\ell_1}$ are equivalent to the residual $G_{F_o}$ Galois representation obtained from $\text{Symm}^n(V_{E,\ell_1})$,
- the residual $G_{F_o}$ Galois representation attached to $W^{(n)}_{t_o,\ell_2}$ has a lifting to characteristic zero that is potentially automorphic and that satisfies the local and determinantal conditions of the Galois deformation theorem.
One cannot get something for nothing (or at least, for absolutely nothing) in mathematics, and it is the last bullet above that reminds us that although the output of our theorem may give us a wealth of Galois representations that are potentially automorphic, the input requires that we prime the pump with some small supply at least of Galois representations that we know to be potentially automorphic. This small supply consists of certain representations induced from one-dimensional characters (see Theorem 4.4.4 of [3] and Theorem 4.2 of [1]).

The conclusion of this scenario is that for \( n \) an even positive integer, both \( W^{(n)}_{t_o} \) and \( \text{Symm}^n(V_E) \) are potentially automorphic.

4.10. **Concluding the theorem.** The direct consequence of the previous subsection is that for all even \( n \) the Galois representations \( \text{Symm}^n(V_E) \) are potentially automorphic. We also know (see section 4.2) that the Galois representations \( \text{Symm}^m(V_E) \) are automorphic—and hence, of course, potentially automorphic—for \( m \leq 4 \). It follows that we have the desired meromorphicity (and nonvanishing) behavior for the \( L \)-functions \( L_{m,n}(s) \) for \( m \leq 4 \) and even positive integers \( n \neq m \). Using merely the pairs \((0,n)\) and \((1,n)\) for even \( n \), we get, as consequence, that for every positive integer \( k \) there is a pair of distinct nonnegative integers \((n,m)\) with \( n + m = k \) and such that

\[
\lim_{C \to \infty} \frac{1}{\pi(C)} \sum_{p \leq C} s_m(\cos \theta_p)s_n(\cos \theta_p) = 0.
\]

Corollary 3.2 then tells us that the sought-for Theorem 2.1 follows.

The theorem proved in the articles I have been reporting on is established, of course, much more generally than only for our sample problem, the elliptic curve \( E : y^2 + y = x^3 - x^2 \) which has conductor 11. What is proved is that if \( E \) is any elliptic curve over \( \mathbb{Q} \) for which there is (at least) one prime number \( \ell \) dividing its conductor and such that \( \ell^2 \) doesn’t divide its conductor, its error terms (i.e., \( p \mapsto (1 + p) - \#E(F_p)) \) conform to the Sato-Tate distribution; moreover, there is a corresponding result for elliptic curves over totally real number fields.

Can one establish potential automorphicity for the Galois representations \( \text{Symm}^n(V_E) \) for all \( n \), even or odd, and all elliptic curves \( E \) over \( \mathbb{Q} \)?

If one has an affirmative answer to this, one will get the further corollary that the error term statistics for any two such elliptic curves that are nonisogenous are also noncorrelated. Recently Michael Harris has made significant progress towards that goal (see his [17]).

A major sticking point to generalize this result to other problems (for example: to the “sample problem” in Part I) is that the rich source of potentially automorphic Galois representations, i.e., the family \( Y_t \) of the previous subsection, has a Hodge structure that parallels the Hodge structure of symmetric powers of Galois representations associated to weight two modular forms, while our first sample problem is of weight 12 and so is not approachable in this manner. Nor would there be an easy replacement for \( Y_t \) that is suitable to tackle weights greater than 2, given the restrictions placed on variation of Hodge structure imposed by Griffiths transversality. We await progress here!

4.11. **Interpreting Sato-Tate as a statement about equidistribution.** We have not yet said a word about why we might expect the Sato-Tate distribution to be
the distribution that accounts for, say, the error term in the elliptic curve data

\[ p \mapsto N_E(p) = 1 + p - \sqrt{p}\{e^{i\theta_p} + e^{-i\theta_p}\} \]

attached to our elliptic curve \( E \). The relevant word here is \textit{equidistribution} as modeled by the Cebotarev Theorem for a finite Galois extension of number fields \( K/Q \), for example,\(^{32}\) that guarantees equidistribution of

\[ p \mapsto \{\text{conjugacy class of } \text{Frob}_p\} \subset \text{Gal}(K/Q), \]

where the finite Galois group \( \text{Gal}(K/Q) \) is given its natural (e.g. “Haar”) measure. In concrete terms this means that the probability that a fixed conjugacy class \( \mathcal{C} \subset \text{Gal}(K/Q) \) occurs as the conjugacy class of \( \text{Frob}_p \) is

\[ \frac{\#\mathcal{C}}{[K : Q]} \]

The specific example of the Sato-Tate Conjecture (now a theorem) that we have been dealing with can be expressed in a vocabulary analogous to the above formulation of the Cebotarev Theorem as follows. Recall that \( USP(2) \) is the unitary symplectic group of genus 1, i.e. the group of complex matrices

\[
\begin{pmatrix}
a & b \\
c & d
\end{pmatrix}
\]

doing determinant 1 and such that \( ac + bd = 0 \) and \( |a|^2 + |b|^2 = |c|^2 + |d|^2 = 1 \).

**Definition 4.4. The unitarized Frobenius conjugacy class at \( p \) of \( E \),** denoted \( \mathcal{C}(p) \subset USP(2) \) is the (unique) conjugacy class of elements in the Lie group \( USP(2) \) with eigenvalues \( e^{i\theta_p} \) and \( e^{-i\theta_p} \).

The Sato-Tate Conjecture for our elliptic curve \( E \) guarantees equidistribution of

\[ p \mapsto \mathcal{C}(p) \subset USP(2) \]

where the Lie group \( USP(2) \) is given its natural (e.g. Haar) measure. Since the trace of an element in \( USP(2) \) determines its conjugacy class, this natural measure on conjugacy classes can, more concretely, be viewed simply as the direct image of Haar measure on \( USP(2) \) under the trace mapping from \( USP(2) \) to the interval \([-2, +2]\). To follow up on this thread and on a discussion of equidistributional properties of a host of other number theoretic problems, see page 7 of [22] (and, indeed, the entire volume [22]).

I find—and I imagine that many people find—the definition we have just given raises more questions than it answers.

About a century ago, Hensel and others formulated the analogy between

- arithmetic (related to a finite prime \( \ell \)) in the field of \( \ell \)-adic numbers, the completion of \( \mathbb{Q} \) with respect to \( \ell \)-adic topology, and
- arithmetic (related to the so-called \textit{infinite prime}) in the field of real numbers, i.e., the completion of \( \mathbb{Q} \) with respect to usual topology.

But number theorists have been acquainted with an unsettling oddness in that analogy ever since. In the above definition we see an example of this “oddness”, as I’ll try to explain below.

\(^{32}\)And even more germane would be the field extensions that split the Galois representations over \( \mathbb{Q} \) acting on groups of \( n \)-torsion points of \( E \).
Our elliptic curve $E$ provides us with an elegant compatible family of $\ell$-adic Galois representations for all finite primes $\ell$. Fix such an $\ell$, and for any prime number $p$ different from 11 and $\ell$, the natural Galois action on $\ell$-power torsion points allows us to “naturally” associate to $p$ a conjugacy class of elements in the $\ell$-adic Lie group $GL_2(\mathbb{Z}_\ell)$ simply by choosing a Frobenius element at $p$ and allowing it to act naturally on the $\ell$-adic vector space $V_{E,\ell}$. Call that conjugacy class $C_{E,\ell}(p)$.

Now—as would seem to follow the format of the above analogy—we do obtain a similar structure relative to the infinite prime. Namely, for every finite $p$, we can pinpoint a conjugacy class, $C_{E,\infty}(p) = C(p)$, of elements in a real Lie group. But to get these conjugacy classes, on the one hand, we must invoke Hasse’s theorem, and on the other hand (at least at the present time), we get them only by executing a peculiarly formal gesture: we simply pick out the unique class with just the right eigenvalues; this is quite different from what we do when we work $\ell$-adically for $\ell$ a finite prime, where we actually find the relevant conjugacy classes selected in some natural way given the structure at hand. All this seems to cry out for some better understanding.

4.12. Expository accounts of this recent work. Different audiences benefit from different shapes of exposition. I wrote a brief article in the journal Nature (vol. 443, 7 September 2006) meant to give a hint of the nature of the Sato-Tate Conjecture and some related mathematical problems to scientists who are not necessarily familiar with much modern mathematics. For professional mathematicians, a number of excellent articles and videos—requiring different levels of prerequisites of their audiences—are devoted to exposing this material:

(1) Available through the MSRI website (http://www.msri.org/):
   (a) an introductory one-hour lecture by Nicholas Katz emphasizing the background and the historical perspective of the work;
   (b) a series of lectures for a number theory workshop by Richard Taylor, where an exposition of the proof itself is given;
   (c) two lectures by Michael Harris, one on some of the material in [3] and one on [18].

(2) Two hours of expository lectures by Laurent Clozel on this topic, aimed at a general mathematical audience in the conference on Current Developments in Mathematics at Harvard University. The notes for these should soon be available as well.

(3) An expository article by Michael Harris: “The Sato-Tate Conjecture: introduction to the proof”.

(4) A talk by Henri Carayol given in the Bourbaki seminar (June 17, 2007): “La conjecture de Sato-Tate [d’après Clozel, Harris, Shepherd-Barron, Taylor]”.

(5) The three articles by the principal authors, [3], [18], and [54], which can be obtained from Richard Taylor’s website (http://www.math.harvard.edu/~rtaylor/).

About the author
Barry Mazur is the Gerhard Gade University Professor at Harvard University.
REFERENCES


DEPARTMENT OF MATHEMATICS, HARVARD UNIVERSITY, ONE OXFORD STREET, CAMBRIDGE, MASSACHUSETTS 02138-2901