BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 23, Number 2, October 1990

EXPONENTIAL SUMS OVER FINITE FIELDS AND DIFFERENTIAL EQUATIONS OVER THE COMPLEX NUMBERS: SOME INTERACTIONS

NICHOLAS M. KATZ

First Lecture

In these lectures, I will try to explain some interactions between the classic theory of linear differential equations in one complex variable with polynomial coefficients, and the theory of one-parameter families of exponential sums over finite fields.

What kind of exponential sums are we talking about? Suppose we start with some polynomial over $\mathbf{Z}$ in some number n of variables

$$f(X_1, X_2, \ldots, X_n) \in \mathbf{Z}[X_1, X_2, \ldots, X_n].$$

A fundamental question (perhaps the fundamental question) we can ask about such an f is: Does $f = 0$ have a solution in integers? Clearly, if there exists an integer solution, then for any prime p there exists a solution in the finite field $\mathbf{Z}/p\mathbf{Z}$; simply reduce mod p any integer solution. (For any ring R, f maps R^n to R. When we say "a solution of $f = 0$ in R," we mean an n-tuple of elements $(a_1, \ldots, a_n)$ of R such that $f(a) = 0$ in R.) Thus, for example, equations of the form

$$Y^a - Y^b = X^c - X^d - 1,$$

with any exponents a, b, c, d all ≥ 1, have no integer solutions, because they have no solutions mod 2. Similarly, the equation

Received by the editors September 13, 1989.

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 10G05, 11L40, 14F99, 32C38, 22E47.

Except for the correction of some typographical errors, the addition of some examples and parenthetical remarks, and the addition of an Appendix and References, this text is identical with the lecture notes distributed for the 1989 Amer. Math. Soc. Colloquium Lectures at Phoenix, and is very close to the oral lectures themselves. I have decided not to "expand" or "revise" them, in the hope that their brevity and informality will attract more readers than their attendant imprecision repels.

0273-0979/90 $1.00 + $.25 per page

$Y^2 = 4X^3 - X - 1$ has no integer solutions, because it has none mod 3. And for any prime $p \geq 5$, the equation

$$X^{p-1} + Y^{p-1} = 3, \quad \text{or } = 4, \quad \text{or } \ldots, \quad \text{or } = p - 1$$

has no integer solutions, because it has none mod p (the left-hand side is always 0, 1, or 2 mod p). So it is always a good idea to check first for the existence of mod p solutions of any equation we hope to solve in integers.

Staying with the same f, a second question that one often asks is: Does $f = N$ have an integer solution for every integer N? Or the variant: Does $f = N$ have an integer solution for every sufficiently large N? If we try checking for mod p solutions of all the equations $f = N$, for all N sufficiently large, we are really looking at only p distinct equations mod p, since all the sufficiently large N's have as reductions mod p only the elements of $\mathbf{Z}/p\mathbf{Z}$. The result of counting the number of mod p solutions of all these equations, then, is the $\mathbf{Z}$-valued function on $\mathbf{Z}/p\mathbf{Z}\, t \mapsto \mathrm{Sol}(f, p, t) :=$ the number of mod p solutions of $f = t$.

Now $\mathbf{Z}/p\mathbf{Z}$ is a finite abelian group under addition, and as such has a Pontrayagin dual group, the group of all the $\mathbf{C}$-valued additive characters $\psi: \mathbf{Z}/p\mathbf{Z} \to \mathbf{C}^\times$. One particular ψ, which we will denote ψ_p, is given by

$$\psi_p(x) := \exp(2\pi i x/p).$$

Any ψ is of the form $x \mapsto \psi_p(tx)$ for some unique t in $\mathbf{Z}/p\mathbf{Z}$, and the pairing $(t, x) \mapsto \psi_p(tx)$ makes $\mathbf{Z}/p\mathbf{Z}$ into its own Pontrayagin dual.

Now let us return to the $\mathbf{Z}$-valued function on $\mathbf{Z}/p\mathbf{Z}$

$$t \mapsto \mathrm{Sol}(f, p, t).$$

We lose no information if we view it as $\mathbf{C}$-valued rather than $\mathbf{Z}$-valued, and we lose no information if we pass to its Fourier transform, namely to the function on the character group

$$\psi \mapsto \sum_{t \bmod p} \psi(t)\, \mathrm{Sol}(f, p, t).$$

An alternate expression for the value at ψ of the Fourier transform is

$$\psi \mapsto \sum_{x_1, \ldots, x_n \bmod p} \psi(f(x_1, \ldots, x_n)).$$

If we use ψ_p to identify $\mathbf{Z}/p\mathbf{Z}$ with its dual, then we may rewrite the Fourier transform as the function on $\mathbf{Z}/p\mathbf{Z}$

$$a \mapsto \sum_{x_1, \ldots, x_n \bmod p} \exp(2\pi i a f(x_1, \ldots, x_n)/p).$$

The exponential sums we have in mind are precisely the values of this Fourier transform, i.e., the sums

$$\sum_{x_1, \ldots, x_n \bmod p} \psi(f(x_1, \ldots, x_n))$$

for each additive character ψ of $\mathbf{Z}/p\mathbf{Z}$.

Here is a mild generalization. Suppose instead of "just" wanting solutions of $f = N$ for all $N \gg 0$, we wanted solutions which were also solutions of a finite list of other equations $G_1 = 0$, $G_2 = 0, \ldots, G_r = 0$. Then we would consider f not as a function on the entire affine space $\mathbf{A}^n$, but rather on the affine variety $V \subset \mathbf{A}^n$ which is defined by the simultaneous vanishing of the G_i's. Repeating the above steps for this modified problem, we would find ourselves dealing with the sums

$$\sum_{x \in V(\mathbf{Z}/p\mathbf{Z})} \psi(f(x)).$$

(For any ring R, $V(R)$ is the set of n-tuples (x) in R^n on which all the $G_i(x) = 0$ in R.)

So the situation now is this: We start with an affine variety V over $\mathbf{Z}$, and a function f on V. For every prime p, and every additive character ψ of $\mathbf{Z}/p\mathbf{Z}$, we have the exponential sum

$$\sum_{x \in V(\mathbf{Z}/p\mathbf{Z})} \psi(f(x)).$$

A modest generalization of this situation is to replace $\mathbf{Z}$ by any ring R which is finitely generated as a $\mathbf{Z}$-algebra (e.g., a finite field, $\mathbf{Z}[\frac{1}{m}]$, or the ring of integers in a finite extension of $\mathbf{Q}$, or ...), and to look at any finite field k, any ring homomorphism $\varphi: R \to k$, any additive character ψ of k, and the exponential sum

$$\sum_{x \in V_\varphi(k)} \psi(f(x)),$$

where we have written V_φ for the variety over k obtained by applying φ to the coefficients of the defining equations of V. For $R = \mathbf{Z}$, the only thing that has changed is that we form the exponential sum over any finite field, not just over a prime field. (In

order to get a "standard" nontrivial character ψ_k of an arbitrary finite field k, it is enough to compose ψ_p ($p :=$ the characteristic of k) with the trace map from k to its prime field $\mathbf{Z}/p\mathbf{Z}$; then every character of k is of the form $x \mapsto \psi_k(tx)$ for a unique t in k, and the pairing $\psi_k(tx)$ makes k self-dual.)

There is a big a priori distinction between what happens when ψ is the trivial character, and when it is not.

Let us first discuss the (highly nontrivial) case when ψ is trivial. Then each term in the sum is 1 (independently of the function f), so the sum is just the number of k-valued points of V_φ, the variety over k deduced from V by applying φ to the coefficients of the defining equations of V. Here are two typical examples of what we are talking about.

Example 1. Take $R = \mathbf{Z}[\frac{1}{26}]$, V the affine curve $y^2 = 4x^3 - x - 1$. This is (the complement of the origin in) an elliptic curve over $\mathbf{Z}[\frac{1}{26}]$. For each prime $p \neq 2, 13$, and each integer $n \geq 1$, let us denote by $N(V; p^n) :=$ the number of $\mathbf{F}_{p^n}$-valued points on V. It has been known since the 1920s (Artin's thesis [Ar], F. K. Schmidt [Sch]) that for each $p \neq 2, 13$, there exist two complex numbers α_p and β_p such that $\alpha_p \beta_p = p$ and such that, for every $n \geq 1$, we have

$$p^n - N(V; p^n) = (\alpha_p)^n + (\beta_p)^n.$$

Thus α_p and β_p are the two roots of the polynomial in $\mathbf{Z}[T]$

$$T^2 - (p - N(V; p))T + p.$$

Hasse [Ha] proved in 1933 the "Riemann Hypothesis for elliptic curves over finite fields":

$$|p - N(V; p)| \leq 2\sqrt{p},$$

or what is the same (by the quadratic formula!), that

$$|\alpha_p| = |\beta_p| = \sqrt{p}.$$

So in this example, we need "only" know the $N(V; p)$'s to determine all the $N(V; p^n)$'s. (And for $p \geq 17$, Hasse's inequality shows that we know the integer $N(V; p)$ if we know it mod p (since all the "candidates" for the integer $N(V, p)$ are in a strip of width $4\sqrt{p}$).

How do the integers $N(V; p)$ vary with p? Again by Hasse, there is a unique "angle" $\vartheta_p \in [0, \pi]$ for which

$$p - N(V; p) = (2\sqrt{p})\cos(\vartheta_p).$$

So we are asking how the angles ϑ_p depend on p.

What we can say about this question depends a great deal on whether or not the elliptic curve in question has "complex multiplication" ["CM" for short] or not. Recall that over $\mathbf{C}$, the $\mathbf{C}$-valued points $E(\mathbf{C})$ of an elliptic curve E form a complex torus $\mathbf{C}/L$, where L is a lattice in $\mathbf{C}$, say $L = \mathbf{Z}\omega_1 + \mathbf{Z}\omega_2$ with $\tau := \omega_2/\omega_1$ having $\mathrm{Im}(\tau) > 0$. This τ is uniquely determined by E up to the fractional linear action of $\mathrm{SL}(2, \mathbf{Z})$ on the upper half-plane. In these terms, CM means that there exist complex numbers α, with α not in $\mathbf{Z}$ such that the "complex multiplication by α" map $z \mapsto \alpha z$ of $\mathbf{C}$ to itself maps the lattice L to itself, and thus defines a complex analytic endomorphism of $\mathbf{C}/L$. In terms of τ, CM means precisely that τ is quadratic over $\mathbf{Q}$. Intrinsically, CM means that E as a commutative algebraic group has more endomorphisms than just multiplication by integers. For example, the curve $y^2 = x^3 - 1$ has CM (cube roots of unity act, by $(x, y) \mapsto (\zeta_3 x, y)$, and τ is $\exp(2\pi i/3)$), and the curve $y^2 = x^3 - x$ has CM (fourth roots of unity act, by $((x, y) \mapsto (-x, iy)$, and τ is $i)$.

For a CM elliptic curve, this problem was solved, first by Weil [We] in some special cases, and then in general by Deuring [Deu], in the early 1950s. Our example curve, however, is not a CM curve (its j invariant, 1728/(-26), blows up at 13, whereas the j invariant of a CM curve is always an algebraic integer). For non-CM elliptic curves, even the **distribution** of the angles $\vartheta_p \in [0, \pi]$ is the subject of a long-standing conjecture, the Sato–Tate conjecture.

To explain what this conjecture says, let us recall first the general notion of equidistribution. Suppose X is a compact space, and μ is a positive $\mathbf{R}$-valued Borel measure on X of total mass 1. A sequence of points $\{x_n\}_{n \geq 1}$ in X is said to be equidistributed for μ if for every continuous $\mathbf{C}$-valued function f on X, we can integrate f against μ by averaging it over more and more of the points, i.e., if

$$\int f d\mu = \lim_{N \to \infty} \left(\frac{1}{N}\right) \sum_{n \leq N} f(x_n).$$

To check equidistribution, it suffices to check f's running over a set of functions whose $\mathbf{C}$-span is uniformly dense in all continuous functions on X. For example, if X is the circle $|z| = 1$, one can test only the functions $z \mapsto z^n$ for all $n \in \mathbf{Z}$. If X is $[0, \pi]$, one

could test any of the following three sequences of functions:

$$\{\cos(\vartheta)^n\}_{n\geq 0}, \qquad \{\cos(n\vartheta)\}_{n\geq 0}, \qquad \{\sin(n\vartheta)/\sin(\vartheta)\}_{n\geq 1}.$$

If X is a finite abelian group G, one can check either by using all the $\mathbf{C}$-valued characters of G, or by using the characteristic functions of all the elements of G.

The best-known and oldest equidistribution theorem is Dirichlet's theorem on primes in arithmetic progression. To put it into our setting, fix an integer $N \geq 2$, take for X the finite abelian group $(\mathbf{Z}/N\mathbf{Z})^\times$ of units mod N, and consider the sequence of elements of G, indexed by the primes p which don't divide N, given by $\{p \bmod N\}_p$. To say that this sequence is equidistributed in $(\mathbf{Z}/N\mathbf{Z})^\times$, with respect to normalized Haar measure, is precisely to say that for each given value $a \in (\mathbf{Z}/N\mathbf{Z})^\times$, the density of those p with $p \equiv a \bmod N$ is $1/\varphi(N)$, as one sees by testing equidistribution using characteristic functions.

Another example of equidistribution, this time due to Hecke (cf. [La], p. 318, Example 2), concerns the angles of Gaussian primes (i.e., of primes in the ring $\mathbf{Z}[i]$ of Gaussian integers). Every odd Gaussian prime has a unique generator π which satisfies $\pi \equiv 1 \bmod (2+2i)$. We will refer to the ratio $\pi/|\pi|$ on the unit circle as the "angle" of the prime. The assertion is that the angles of the odd Gaussian primes are equidistributed in the unit circle, with respect to its normalized Haar measure.

Yet another example of equidistribution, due to Heath-Brown and Patterson [HBP], concerns the angles of the cubic gauss sums. For each prime p, consider the sum

$$g(p) := \sum_{x \bmod p} \exp(2\pi i x^3/p).$$

This sum is real, and it vanishes trivially unless $p \equiv 1 \bmod 3$ (otherwise $x \mapsto x^3$ is bijective on $\mathbf{Z}/p\mathbf{Z}$). For $p \equiv 1 \bmod 3$, one knows that $|g(p)| \leq 2\sqrt{p}$. Therefore there is a unique "angle" $\vartheta_p \in [0, \pi]$ for which $g(p) = (2\sqrt{p})\cos(\vartheta_p)$. The assertion (cf. [Pa, Appendix II]) is that the sequence of angles $\{\vartheta_p\}_{p\equiv 1 \bmod 3}$ is equidistributed in $[0, \pi]$ for the uniform measure $(1/\pi)\,d\vartheta$.

The Sato–Tate conjecture is that for a non-CM elliptic curve over $\mathbf{Q}$, so with good reduction over some $\mathbf{Z}[\frac{1}{m}]$, the sequence of angles $\{\vartheta_p\}_p$ is equidistributed in $[0, \pi]$ for the "Sato–Tate measure" $\mu_{\mathrm{ST}} := (2/\pi)\sin^2\vartheta\, d\vartheta$. **It is not known in a single case.** However, the "function field analogue" of it is known (Yoshida

[Yo]). Moreover, in the function field case we have, thanks to Delignes Weil II [De 7], a very good understanding of quite general phenomena of this type. We will try to explain the general setup in the next lecture.

Just to show how very little is known, let us derive an elementary-sounding consequence of the Sato–Tate conjecture which is itself unknown. Thus let E be a non-CM elliptic curve over $\mathbf{Q}$, so with good reduction over some $\mathbf{Z}[\frac{1}{m}]$. For each p prime to m, we denote by N_p the number of mod p points (x, y) on E (i.e., not including the point ∞), so that the angle ϑ_p in $[0, \pi]$ is defined by $p - N_p = 2\sqrt{p}\cos(\vartheta_p)$. Since the Sato–Tate measure on $[0, \pi]$ is symmetric about the midpoint $\frac{\pi}{2}$, and $\cos(\vartheta)$ is positive in $[0, \frac{\pi}{2})$ and negative in $(\frac{\pi}{2}, \pi]$, it would follow that the set of primes p for which $p > N_p$ has density $\frac{1}{2}$, as does the set of primes p for which $p < N_p$. This is not known. (It is known (cf. [Se, IV-13, Exercise 2]) that the set of primes p for which $p = N_p$, or indeed for which ϑ_p is any given angle, has density zero, for any non-CM E over $\mathbf{Q}$.)

Example 2. Let p be a prime with $p \geq 5$,

$$R = \mathbf{F}_p[T, 1/(T(T-27))],$$

V the affine curve $y^2 = 4x^3 - Tx - T$. This is (the complement of the origin in) an elliptic curve over R. Its j invariant, $(12)^3 T/(T-27)$, is nonconstant (we took $p \geq 5$ to assure this). For each power q of p, and each $t \in \mathbf{F}_q$(:= the finite field with q elements) with $t \neq 0, 27$, this equation over $\mathbf{F}_q$, with $T \mapsto t$, is an elliptic curve. Just as above, it has an associated angle $\vartheta_{q,t}$ in $[0, \pi]$ such that $q - N(V; q, t) = (2\sqrt{q})\cos(\vartheta_{q,t})$.

So now we have, for each power q of p, a finite collection of angles $\{\vartheta_{q,t}\}_t$. Using this collection, we form the measure on μ_q on $[0, \pi]$, which consists of averaging over these angles. By the (known) function-field case of Sato–Tate, as q increases archimedeanly to ∞, the sequences of measures μ_q converges weakly to μ_{ST}, in the sense that for any continuous function f on $[0, \pi]$, we have

$$\int f d\mu_{\mathrm{ST}} = \lim_{q \to \infty} \int f \, d\mu_q.$$

This result is very effective. Consider the functions

$$f_n(\vartheta) := \sin(n\vartheta)/\sin(\vartheta), \qquad n = 1, 2, \ldots.$$

These are an orthonormal basis of $L^2([0, \pi], \mu_{ST})$; in particular,

$$f_n d\mu_{\mathrm{ST}} = 0 \qquad \text{unless } n = 1,\ f_1\, d\mu_{\mathrm{ST}} = 1.$$

Since each μ_q has total mass one, and f_1 is the constant function 1, we have $\int f_1\, d\mu_{\mathrm{ST}} = \int f_1\, d\mu_q$ for every q. For $n \geq 2$, we have the estimate (which depends heavily on Weil II [De 7])

$$\left|\int f_n\, d\mu_{\mathrm{ST}} - \int f_n\, d\mu_q\right| = \left|\int f_n\, d\mu_q\right| \leq n\sqrt{q}/(q-2).$$

(The constant "1" in front of the "n" in the estimate above is, in the general case, $(\pm 1)\times$ the topological Euler characteristic of the parameter curve, which in our example is $\mathbf{A}^1$-{2 points}.)

Example 3. Take $R = \mathbf{Z}[T, 1/(6T(T-27))]$, V the affine curve $y^2 = 4x^3 - x - T$. This is (the complement of the origin in) an elliptic curve over R whose j invariant is $(12)^3 T/(T-27)$. Now we have angles $\vartheta_{q,t}$ for every prime power $q = p^n$, $p \neq 2, 3$, and every t in $\mathbf{F}_q$ with $t \neq 0, 27$ in $\mathbf{F}_q$. So in some sense we have a "two-parameter family" of angles $\vartheta_{q,t}$, albeit with a slightly strange indexing set. Intrinsically, we should consider the scheme $\mathscr{M} := \mathrm{Spec}(R)$. By definition, for any ring A, an A-valued point of $\mathscr{M} := \mathrm{Spec}(R)$ is nothing more or less than a ring homomorphism from R to A:

$$\mathscr{M}(A) := \mathrm{Hom}_{\mathrm{ring}}(R, A).$$

Thus "t in $\mathbf{F}_q$ with $t \neq 0, 27$ in $\mathbf{F}_q$" is just an $\mathbf{F}_q$-valued point of $\mathscr{M}$, i.e.,

$$\mathscr{M}(\mathbf{F}_q) := \{t \text{ in } \mathbf{F}_q \text{ with } t \neq 0, 27 \text{ in } \mathbf{F}_q\}.$$

It will clarify matters if we consider only the prime fields themselves. For each p we have the packet of angles $\{\vartheta_{p,t}\}$ indexed by t in $\mathbf{F}_p - \{0, 27\}$, and the associated measures μ_p on $[0, \pi]$. From the estimate

$$\left|\int f_n\, d\mu_{\mathrm{ST}} - \int f_n\, d\mu_p\right| \leq n\sqrt{p}/(p-2),$$

we see that as the prime p increases to ∞, the measures μ_p converge weakly to μ_{ST}, a fact that was established already by Birch [Bi] in the late sixties.

To "see" what is going on, think of all the angles $\vartheta_{p,t}$ arranged on graph paper, with p along the x-axis and t along the y-axis. Then in successive vertical columns, the angles in each column

get "more and more equidistributed" with respect to Sato–Tate measure (by which we mean that the measures μ_p converge weakly to μ_{ST}, and that the above estimate holds). We will summarize this by saying that we have "**vertical equidistribution.**" From this point of view, the Sato–Tate conjecture for the curve $y^2 = 4x^3 - x - 1$ over $\mathbf{Z}[\frac{1}{26}]$ is the statement that the **horizontal sequence** of angles $\{\vartheta_{p,1}\}_{p\geq 17}$ (we may leave out any finite chunk of angles without affecting equidistribution) is equidistributed in $[0, \pi]$ for μ_{ST}.

Similarly, if we fix any rational number $\alpha \neq 0, 27$, then the equation $y^2 = 4x^3 - \alpha x - \alpha$ defines an elliptic curve E_α over $\mathbf{Q}$ with good reduction outside some finite set of primes. So it makes sense to speak of the angles $\vartheta_{p,\alpha}$ for all p where E_α has good reduction. (Notice that if α is a true fraction, then the α in $\vartheta_{p,\alpha}$ means "α mod p." When we go to graph the level set "$t = \alpha$," it is quite a mess. Of course if α is a positive integer, then as soon as $p > \alpha$ this level set is really a horizontal line of height α.)

Since there are some values of α for which the curve E_α has CM, we cannot hope that every horizontal sequence of angles $\{\vartheta_{p,\alpha}\}_p$ will be equidistributed with respect to μ_{ST}, since we know that this will fail if E_α is CM (see CM Note at the end of this section). On the other hand, there is a lot of computer computation which is compatible with the Sato–Tate conjecture. However, such evidence by itself is not too convincing, because a random horizontal sequence of angles will be equidistributed with respect to Sato–Tate.

Here is a precise statement. Take any infinite set S of primes $p \geq 5$. Denote by X the product space $X := \prod_{p\in S} \mathscr{M}(\mathbf{F}_p)$, recall $\mathscr{M}(\mathbf{F}_p) := \mathbf{F}_p - \{0, 27\}$, and endow each factor $\mathbf{F}_p - \{0, 27\}$ with the measure which gives each point mass $1/(p-2)$. Denote by μ_{prod} the product measure on X. Now a point x of the space X is a sequence $\{(p, \alpha_p)\}_p$ where each α_p is an element of $\mathbf{F}_p - \{0, 27\}$. Attached to each element x in X is the sequence of angles $\{\vartheta_{p,\alpha_p}\}_p$ indexed by the primes p in S.

Theorem (joint with O. Gabber). *The set of points x in X for which the associated sequence of angles is equidistributed with respect to μ_{ST} has measure 1 in X with respect to the measure μ_{prod} on X.*

Proof. Once properly set up, this is an easy consequence of the strong law of large numbers. One shows that for each integer

$n \geq 1$, the set of x in X such that successively averaging over its sequence of angles correctly integrates the function $f_n(\vartheta) := \sin(n\vartheta)/\sin(\vartheta)$ has measure one. This is done by considering, for $n \geq 2$ fixed (the case $n = 1$ being trivial), the sequence of independent functions $\{f_{n,p}\}_{p \in S}$ on X given by

$$x = \{(p, \alpha_p)\} \mapsto f_n(\vartheta_{p,\alpha_p}),$$

and applying to this sequence of independent functions the strong law of large numbers. These functions are all bounded (by n), so their variances are bounded (by n^2). Their expectations over X are tautologically given by $E(f_{n,p}) := \int f_n \, d\mu_p$. So from the above estimate for $\int f_n \, d\mu_p$, we see that for fixed n, the $E(f_{n,p})$ tend to zero (rather fast, being $O(1/\sqrt{p})$, but this is not needed) as p grows. By the Toeplitz lemma ($y_i \to 0 \Rightarrow (1/N)\sum_{i \leq N} y_i \to 0$), the sequence of numbers

$$(\tfrac{1}{N}) \sum_{\text{first } N \text{ primes } p \text{ in } S} E(f_{n,p})$$

has limit 0. So by the strong law of large numbers, the sequence of functions on X

$$(\tfrac{1}{N}) \sum_{\text{first } N \text{ primes } p \text{ in } S} f_{n,p}(x)$$

converges almost everywhere to zero, which is precisely the statement that the sequence of angles attached to x correctly (since $\int f_n \, d\mu_{ST} = 0$ for $n \geq 2$) integrates f_n for almost all x.

One then uses the fact that a countable union of sets of measure zero is of measure zero, so with probability one our x correctly integrates all the functions f_n. Q.E.D.

In other words, as soon as we know "vertical equidistribution" with a tiny bit of uniformity in p, then with probability one we have horizontal equidistribution. Of course this result tells us nothing about any **particular** horizontal sequence; it could even be the case that the entire countable set of horizontal sequences we are interested in (the angles of E_α for $\alpha \in \mathbf{Q} - \{0, 27\}$) lies in the exceptional set of measure zero where Sato–Tate does not hold!

CM Note. Here is what happens if we start with an elliptic curve E_α over $\mathbf{Q}$ which is CM, say with CM by the ring of integers in a quadratic imaginary field K. For those primes $p \geq 5$ of good

reduction which do not split in K, the angle $\vartheta_{p,\alpha}$ is always $\frac{\pi}{2}$ (because E mod p is then a supersingular elliptic curve over $\mathbf{F}_p$). If we look at the primes $p \geq 5$ of good reduction which do split in K, the sequence of angles $\{\vartheta_{p,\alpha}\}_{p \text{ splits}}$ in $[0, \pi]$ is equidistributed not with respect to Sato–Tate measure $\mu_{ST} := (\frac{2}{\pi}) \sin^2 \vartheta \, d\vartheta$, but rather with respect to uniform measure $(\frac{1}{\pi}) d\vartheta$. (For by Deuring [Deu] we are dealing with the angles of a unitary grossencharacter of infinite order, and by Hecke these are equidistributed in angular sectors.) Here is an example, due essentially to Gauss (cf. the "last entry" in his mathematical diary): take the CM curve $y^2 = x^3 - x$ over $\mathbf{Q}$, which has CM by $\mathbf{Z}[i]$. Each rational prime $p \equiv 1 \bmod 4$ splits in $\mathbf{Z}[i]$ as $p = \pi\overline{\pi}$, where π and its complex conjugate $\overline{\pi}$ are odd Gaussian primes both $\equiv 1 \bmod (2 + 2i)$. For $p \equiv 1 \bmod 4$, the numbers $\{\alpha_p, \beta_p\}$ for E are $\{\pi, \overline{\pi}\}$, so the angle ϑ_p in $[0, \pi]$ is given by

$$\cos(\vartheta_p) = \mathrm{Re}(\pi/|\pi|) = \mathrm{Re}(\overline{\pi}/|\pi|).$$

Second Lecture

In today's lecture, I want to try to explain the modern approach to equidistribution, and to show you "where the Sato–Tate measure comes from." This is a pretty long story, which starts in the early 1920s and culminates in the middle 1970s with Deligne's Weil II [De 7], so we will have to leave out some of the details.

For motivation, let us begin with a smooth connected complex algebraic variety X. Sometimes we will want to think of it algebraically, i.e., to think of the equations which define it, and sometimes we want to think of its $\mathbf{C}$-valued points $X(\mathbf{C})$ as a complex manifold "X^{an}," so that we can draw pictures and apply our topological intuition. For instance, sometimes we think of an elliptic curve as given by an equation, and sometimes we want to think of it as a complex torus.

When we think about X^{an}, we can apply to it the full arsenal of standard algebraic topology, and in particular we can talk about its integral cohomology groups $H^i(X^{an}, \mathbf{Z})$ and its integral cohomology groups with compact support $H^i_c(X^{an}, \mathbf{Z})$; we know that these are $\mathbf{Z}$-modules of finite type with a plethora of functorial properties. How much of this can be defined purely algebraically (i.e., only using the equations which define X, but without using the

topology of $\mathbf{C}$)? It turns out that what one can define purely algebraically are the corresponding groups with coefficients in $\mathbf{Z}/N\mathbf{Z}$ for any integer N; this is part of what Grothendieck, Artin et al. did in the early sixties in [Gro3, SGA4]. Picking a prime number ℓ and taking successively $N = \ell, \ell^2, \ell^3, \ldots$, then passing to an inverse limit, we find that the $\mathbf{Z}_\ell$-modules (here $\mathbf{Z}_\ell$ denotes the ring of ℓ-adic integers)

$$H^i(X^{an}, \mathbf{Z}) \otimes \mathbf{Z}_\ell, H^i_c(X^{an}, \mathbf{Z}) \otimes \mathbf{Z}_\ell,$$

have purely algebraic definitions; let us denote them

$$H^i(X, \mathbf{Z}_\ell), H^i_c(X, \mathbf{Z}_\ell)$$

to emphasize their purely algebraic nature. These are called the ℓ-adic cohomology groups of X (resp. ... with compact support).

What is even better for our purposes, the theory exists not just for complex varieties, but for arbitrary varieties X over arbitrary algebraically closed fields k; the only proviso is that if one is over a ground field of positive characteristic p, one must be sure to choose a prime $\ell \neq p$. In the following discussion, we will always assume that ℓ has been so chosen.

Now suppose we start with an algebraic variety X over a finite field $\mathbf{F}_q$. Then by extending scalars from $\mathbf{F}_q$ to an algebraic closure $\overline{\mathbf{F}}_q$, we get a variety $\overline{X} := X \otimes_{\mathbf{F}_q} \overline{\mathbf{F}}_q$ over an algebraically closed field. So we can speak of the finitely generated $\mathbf{Z}_\ell$-modules $H^i(\overline{X}, \mathbf{X}_\ell)$, $H^i_c(\overline{X}, \mathbf{Z}_\ell)$, and of the finite-dimensional $\mathbf{Q}_\ell$-vector spaces

$$H^i(\overline{X}, \mathbf{Q}_\ell) := H^i(\overline{X}, \mathbf{Z}_\ell) \otimes \mathbf{Q}_\ell, H^i_c(\overline{X}, \mathbf{Q}_\ell) := H^i_c(\overline{X}, \mathbf{Z}_\ell) \otimes \mathbf{Q}_\ell.$$

Because the theory is so functorial, the Galois group of $\overline{\mathbf{F}}_q$ over $\mathbf{F}_q$ operates on these cohomology groups. This Galois group is a free profinite group on one canonical generator, namely the qth power map $z \mapsto z^q$ on $\overline{\mathbf{F}}_q$, which is called the "arithmetic Frobenius." The **inverse** of this generator is called the "geometric Frobenius," and **denoted** Frob_q. **This geometric Frobenius, and its action on the compact cohomology of $\overline{X}$, entirely determines the Diophantine structure of** X. One has the Lefschetz trace formula (cf. [Gro 2]), according to which

$$\mathrm{Card}(X(\mathbf{F}_q)) = \sum_i (-1)^i \,\mathrm{trace}(\mathrm{Frob}_q \,|\, H^i_c(\overline{X}, \mathbf{Q}_\ell)),$$

and for every $n \geq 1$,

$$\mathrm{Card}(X(\mathbf{F}_{q^n})) = \sum_i (-1)^i \,\mathrm{trace}((\mathrm{Frob}_q)^n | H_c^i(\overline{X}, \mathbf{Q}_\ell)).$$

So we are in the strange situation of writing an ordinary whole number as an alternating sum of ℓ-adic numbers! (And even today it is not known in general that the individual traces are in $\mathbf{Z}$, and independent of the choice of ℓ; although it is known if X is both proper and smooth, as a consequence of Deligne's Weil II.) One can show, however, that all the eigenvalues of Frob_q on $H_c^i(\overline{X}, \mathbf{Q}_\ell)$ are algebraic over $\mathbf{Q}$. So by thinking of these algebraic numbers as complex numbers, we can ask about their complex absolute values. Of course, a given algebraic number α has many $\mathbf{Q}$-conjugates in $\mathbf{C}$, and in general the various conjugates have different absolute values (e.g., $1 \pm \sqrt{2}$). In Weil II [De 7], Deligne proves that for each eigenvalue α_i of Frob_q on $H_c^i(\overline{X}, \mathbf{Q}_\ell)$, there is an integer $w = w(\alpha_i)$, called its weight, such that

$$0 \leq w \leq i,$$

$$|\text{every conjugate of } \alpha_i \text{ as an algebraic number}| = (\sqrt{q})^w.$$

If, for a given i, all the α_i have the same weight $w(\alpha_i)$, we say that $H_c^i(\overline{X}, \mathbf{Q}_\ell)$ is pure of weight w. If all the weights in $H_c^i(\overline{X}, \mathbf{Q}_\ell)$ are at most some integer j, we say that $H_c^i(\overline{X}, \mathbf{Q}_\ell)$ is mixed of weight $\leq j$.

For X proper and smooth, a Poincare duality argument shows that $H_c^i(\overline{X}, \mathbf{Q}_\ell)$ is pure of weight i. In general, one will "only" have that $H_c^i(\overline{X}, \mathbf{Q}_\ell)$ is mixed of weight $\leq i$.

By the way, what happens if we start with a variety X over, say, $\mathbf{Z}[1/\ell]$? Then we can apply the theory to each of the fibers $X_p := X \otimes \mathbf{F}_p$, and obtain for each integer i and each prime $p \neq \ell$ a finite-dimensional $\mathbf{Q}_\ell$-vector space $H_c^i(\overline{X}_p, \mathbf{Q}_\ell)$ on which Frob_p acts. We would like to know that outside some finite set of exceptional primes p, these are all the same. The same as what? Well, the complex fiber $X_{\mathbf{C}}$ of X is a complex variety, so we can speak also of $H_c^i(X_{\mathbf{C}}, \mathbf{Q}_\ell)$. In fact, ℓ-adic cohomology is invariant under extension of algebraically closed field, so we have

$$H_c^i(X_{\overline{\mathbf{Q}}}, \mathbf{Q}_\ell) \approx H_c^i(X_{\mathbf{C}}, \mathbf{Q}_\ell).$$

By functoriality, the Galois group of $\overline{\mathbf{Q}}$ over $\mathbf{Q}$ operates on $H_c^i(\overline{X}_{\overline{\mathbf{Q}}}, \mathbf{Q}_\ell)$. It is a basic result of the theory (the "constructibility of higher direct images") that this action is unramified at all

but finitely many p, and that for any good prime p there is an isomorphism

$$H_c^i(\overline{X}_p, \mathbf{Q}_\ell) \approx H_c^i(X_{\overline{\mathbf{Q}}}, \mathbf{Q}_\ell)$$

which respects the action of Frob_p.

To understand this better, recall (cf. [Gro 3, SGA1]) that for any connected scheme X, one has its profinite π_1, which classifies the finite etale coverings of X. (Strictly speaking, one must pick base points, but we will ignore this difficulty and work up to conjugacy.) For variable X, π_1 is covariant, and for $X = \mathrm{Spec}(k)$ with k a field, π_1 is just the Galois group of $\overline{k}$ over k. For X a scheme, and k a field, we can interpret a k-valued point $x \in X(k)$ as a map of schemes $\varphi\colon \mathrm{Spec}(k) \to X$. By the covariance of π_1, we get a homomorphism $\mathrm{Gal}(\overline{k}/k) \to \pi_1(X)$ (well defined up to conjugacy in the target). So if k is a finite field $\mathbf{F}_q$, we can attach to each $\mathbf{F}_q$-valued point $x \in X(\mathbf{F}_q)$ the image of Frob_q in $\pi_1(X)$, which we denote $\mathrm{Frob}_{q,x}$; it is well defined as a conjugacy class in $\pi_1(X)$, and is called the Frobenius conjugacy class attached to the pair $(\mathbf{F}_q, x \text{ in } X(\mathbf{F}_q))$.

Example. For the scheme $X := \mathrm{Spec}(\mathbf{Z}[\frac{1}{m}])$, $\pi_1(X)$ is the quotient of the absolute Galois group of $\mathbf{Q}$ which classifies those finite extensions of $\mathbf{Q}$ which are unramified outside of primes dividing m. (For X the spec of any normal integral domain R, $\pi_1(X)$ is the quotient of the absolute Galois group of the fraction field K of R, which classifies those finite separable extensions E of K such the integral closure of R in E is finite etale over R.)

In the classical case, finite-dimensional linear representations of π_1 give us "local coefficient systems." Such local systems occur naturally as the "cohomology along the fibers" of nice fiber spaces over X. It makes sense to take cohomology with values in such a local system. And there is a general notion of a "constructible sheaf" such that a local system is a particularly nice kind of constructible sheaf. **In the ℓ-adic theory things work just the same.**

The main new aspect of the theory comes on schemes X of finite type over $\mathbf{Z}$, where the π_1 contains all the various Frobenius conjugacy classes $\mathrm{Frob}_{q,x}$. For $\mathscr{F}$ an ℓ-adic local system (i.e., a finite-dimensional $\mathbf{Q}_\ell$-representations of π_1, sometimes called a **lisse sheaf**), we can speak of the action of $\mathrm{Frob}_{q,x}$ on $\mathscr{F}$, its trace, its eigenvalues, etc. These notions also make sense for the

more general notion of a constructible ℓ-adic sheaf, a sheaf whose restriction to each piece of a stratification of X as a finite disjoint union of locally closed subvarieties is lisse.

For X over a finite field $\mathbf{F}_q$, and $\mathscr{F}$ on X, the Lefschetz trace formula asserts that

$$\sum_{x \text{ in } X(\mathbf{F}_q)} \text{trace}(\text{Frob}_{q,x} | \mathscr{F}_x) = \sum_i (-1)^i \text{trace}(\text{Frob}_q | H^i_c(\overline{X}, \mathscr{F})).$$

In the case of the constant sheaf $\mathscr{F} = \mathbf{Q}_\ell$, corresponding to the trivial one-dimensional representation of π_1, this is the earlier Lefschetz formula.

Now let Y be a scheme of finite type over $\mathbf{Z}$, $f: X \to Y$ a map of finite type, and $\mathscr{F}$ a constructible ℓ-adic sheaf on X. For every integer i, we can form the higher direct image with compact support $R^i f_! \mathscr{F}$, which will be a constructible ℓ-adic sheaf on Y. **Its formation commutes with passage to fibers** [Gro 3, SGA4]. Thus, for any finite field $\mathbf{F}_q$, and any point y in $Y(\mathbf{F}_q)$, the action of $\text{Frob}_{q,y}$ on the fiber $(R^i f_! \mathscr{F})_y$ is the action of Frob_q on the cohomology group $H^i_c(X_y \otimes_{\mathbf{F}_q} \overline{\mathbf{F}}_q, \mathscr{F})$. The constructibility of $R^i f_! \mathscr{F}$ entails that on some dense open set U of Y, the cohomology groups $H^i_c(X_y \otimes_{\mathbf{F}_q} \overline{\mathbf{F}}_q, \mathscr{F})$ with their actions of Frob_q all fit together into an ℓ-adic representation of $\pi_1(U)$. (In the special case when Y is $\text{Spec}(\mathbf{Z})$, the open dense sets are precisely the specs of the rings $\mathbf{Z}[\frac{1}{m}]$, and we recover the earlier discussion.)

In this context, we can state the fundamental estimate of Weil II in its full generality. We say that a constructible sheaf $\mathscr{F}$ on X is "punctually pure of weight w" if for every finite field $\mathbf{F}_q$ and every x in $X(\mathbf{F}_q)$, the eigenvalues of $\text{Frob}_{q,x}$ acting on $\mathscr{F}_x$ are algebraic numbers all of whose $\mathbf{Q}$-conjugates have complex absolute value $(\sqrt{q})^w$. We say that $\mathscr{F}$ is "mixed of weight $\leq w$" if it is a successive extension of constructible sheaves, each of which is punctually pure of some weight which is $\leq w$.

Theorem. (Deligne, Weil II [De 7]). *Let Y be a scheme of finite type over $\mathbf{Z}[1/\ell]$, $f: X \to Y$ a morphism of finite type, and $\mathscr{F}$ a constructible ℓ-adic sheaf on X which is mixed of weight $\leq w$. Then for every integer i, the* ith *higher direct image with compact supports $R^i f_! \mathscr{F}$ is a constructible sheaf on Y which is mixed of weight $\leq w + i$.*

Notice that in the case when Y is the spec of a finite field, and $\mathscr{F}$ is the constant sheaf $\mathbf{Q}_\ell$ on X, we recover the fact that $H^i_c(\overline{X}, \mathbf{Q}_l)$ is mixed of weight $\leq i$.

In general, even if $\mathscr{F}$ is both lisse and pure of some weight, the sheaf $R^i f_!\mathscr{F}$ will be neither lisse nor pure of any weight. However, there are many particular situations $(f\colon X \to Y, \mathscr{F})$ where, for one reason or another, one knows that $R^i f_!\mathscr{F}$ is both lisse, and pure of weight $w+i$. (This is true, for instance, if $\mathscr{F}$ is lisse and pure, and the map f is proper and smooth. But there are many other cases where suitable conditions on the ramification of $\mathscr{F}$ "at infinity" will also guarantee that the $R^i f_!\mathscr{F}$ are lisse, and pure of known weight. The study of exponential sums will produce a plethora of examples of this.)

The next thing I want to explain is the equidistribution theory of the sums attached to a lisse pure sheaf $\mathscr{G}$ (e.g., an $R^i f_!\mathscr{F}$) on a variety Y over a finite field $\mathscr{F}_q$. (I wish I could explain to you the theory when Y is an arbitrary scheme of finite type over $\mathbf{Z}$, or even when Y is just an open set $\mathrm{Spec}(\mathbf{Z}[\frac{1}{m}])$ of $\mathrm{Spec}(\mathbf{Z})$, but the theory does not yet exist.)

The essential case is when Y is a smooth, geometrically connected curve over $\mathbf{F}_q$. For instance, consider the situation of Example 2. $p \geq 5$, $R = \mathbf{F}_p[T, 1/(T(T-27))]$, V the affine curve

$$y^2 = 4x^3 - Tx - T.$$

We take

$$Y := \mathrm{Spec}(R) = \mathbf{A}^1 - \{0, 27\},$$
$$X := V, \mathscr{F} \text{ the constant sheaf } \mathbf{Q}_\ell \text{ on } V,$$
$$f\colon X \to Y \text{ the natural projection } (x, y, T) \mapsto T.$$

Then $\mathscr{G} := R^1 f_!\mathbf{Q}_\ell$ is lisse of rank two and pure of weight one on Y, and its traces of Frobenius are given by

$$\mathrm{trace}(\mathrm{Frob}_{q,t}|\mathscr{G}) = q - N(V; q, t) (= (2\sqrt{q})\cos(\vartheta_{q,t})).$$

In this example, the only other nonvanishing $R^i f_!\mathbf{Q}_\ell$ occurs for $i = 2$, and $R^2 f_!\mathbf{Q}_\ell$ is the Tate twist $\mathbf{Q}_\ell(-1)$. (In general, for any integer n, the Tate twist $\mathbf{Q}_\ell(-n)$ is the lisse rank one sheaf on which $\mathrm{Frob}_{q,t}$ acts by the scalar q^n. If we are willing to replace $\mathbf{Q}_\ell$ by a finite extension, it makes sense to speak of this Tate twist when n is a rational number.) So the above formula is just a rewriting of the Lefschetz trace formula

$$N(V; q, t) = \mathrm{trace}(\mathrm{Frob}_{q,t}|R^2 f_!\mathbf{Q}_\ell) - \mathrm{trace}(\mathrm{Frob}_{q,t}|R^1 f_!\mathbf{Q}_\ell).$$

Now if we Tate-twist $\mathscr{G}$ by $\frac{1}{2}$, i.e., pass to the sheaf

$$\mathscr{F} := \mathscr{G}(\tfrac{1}{2}) := \mathscr{G} \otimes \mathbf{Q}_\ell(\tfrac{1}{2}),$$

all we have done is to divide the traces by $\sqrt{q}$:

$$\operatorname{trace}(\mathrm{Frob}_{q,t}|\mathscr{F}) = (q - N(V; q, t))/\sqrt{q}(= 2\cos(\vartheta_{q,t})).$$

So we find that we have been speaking prose all along: The angles $\vartheta_{q,t}$ that we discussed at such length in the first lecture are the ones whose cosines are given by $((\frac{1}{2})\times)$ the traces of Frobenius of a lisse rank two sheaf on Y which is pure of weight zero! Believe it or not, this actually represents a tremendous conceptual advance!

Third Lecture

In this lecture, I want to finish explaining the modern approach to equidistribution. So for today we take a finite field $\mathbf{F}_q$, an open smooth, geometrically connected curve Y over $\mathbf{F}_q$, and a lisse ℓ-adic sheaf $\mathscr{F}$ on Y which is pure of weight zero. Now $\mathscr{F}$ "is" an ℓ-adic representation, say ρ, of the fundamental group $\pi_1(Y)$ in some $\mathrm{GL}(n, \overline{\mathbf{Q}}_\ell)$, which is definable over some unspecified finite extension of $\mathbf{Q}_\ell$. (So now we are considering "lisse $\overline{\mathbf{Q}}_\ell$-sheaves"; this is an inessential change.) There is short exact sequence of fundamental groups

$$\begin{array}{c} 0 \to \pi_1(Y \otimes \overline{\mathbf{F}}_q) \to \pi_1(Y) \to \pi_1(\mathrm{Spec}(\mathbf{F}_q)) \to 0 \\ \qquad\qquad\qquad\qquad\qquad \| \\ \qquad\qquad\qquad\qquad\qquad \mathrm{Gal}(\overline{\mathbf{F}}_q/\mathbf{F}_q) \approx \widehat{\mathbf{Z}}, \quad \text{with generator } \mathrm{Frob}_q. \end{array}$$

The subgroup $\pi_1(Y \otimes \overline{\mathbf{F}}_q)$ of $\pi_1(Y)$ is called the geometric fundamental group of Y, and it is denoted $\pi_1^{\mathrm{geom}}(Y)$, or just π_1^{geom}; for emphasis the entire group $\pi_1(Y)$ is the denoted π_1^{arith}, the "arithmetic" fundamental group of Y. (The group π_1^{geom} is "geometric" in the sense that it is the π_1 of a variety over an algebraically closed field; the group π_1^{arith} is "arithmetic" in the sense that it contains lots of Frobenius conjugacy classes.)

The sheaf $\mathscr{F}$ is just a representation of the group π_1^{arith}, and its inverse image on $Y \otimes \overline{\mathbf{F}}_q$ is the restriction of that representation to π_1^{geom}. The cohomology groups $H_c^i(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})$ are representations of the quotient $\pi_1^{\mathrm{arith}}/\pi_1^{\mathrm{geom}} \approx \mathrm{Gal}(\overline{\mathbf{F}}_q/\mathbf{F}_q)$. To what extent

can we calculate the cohomology directly in terms of the representation? The answer is "half." Because Y is an open curve and $\mathscr{F}$ is lisse, we have

$$H_c^i(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F}) = 0 \quad \text{for } i \neq 1, 2.$$

The group $H_c^2(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})$ has an easy description in terms of the coinvariants $(\mathscr{F})_{\pi_1^{\text{geom}}}$:= the largest quotient where π_1^{geom} acts trivially. Namely

$$H_c^2(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F}) \approx (\mathscr{F})_{\pi_1^{\text{geom}}}(-1),$$

(-1) denoting the Tate twist, and the action of Frob_q on $H_c^2(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})$ is induced by the action on $(\mathscr{F})_{\pi_1^{\text{geom}}}(-1)$ of any element of π_1^{arith} which maps onto Frob_q in the homotopy short exact sequence above (e.g., if Y has an $\mathbf{F}_q$-point y, by the action of $\text{Frob}_{q,y}$). This description makes it clear that:

(1) If $\mathscr{F}$ is pure of weight zero, then $H_c^2(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})$ is pure of weight two.

(2) If π_1^{geom} acts irreducibly and nontrivially on $\mathscr{F}$, $H_c^2(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F}) = 0$.

What do we know about $H_c^1(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})$? There are two fundamental pieces of information. The first is Deligne's estimate: $H_c^1(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})$ is mixed of weight ≤ 1. The second is the Lefschetz trace formula: for every $n \geq 1$,

$$\begin{aligned}\text{trace}((\text{Frob}_q)^n | H_c^2(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})) - \text{trace}((\text{Frob}_q)^n | H_c^1(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})) \\ = \sum_{y \text{ in } Y(\mathbf{F}_{q^n})} \text{trace}(\text{Frob}_{q^n, y} | \mathscr{F}).\end{aligned}$$

Key Lemma. *Suppose that $\mathscr{F}$ is pure of weight zero and that π_1^{geom} acts irreducibly and nontrivially on $\mathscr{F}$. Then for any complex embedding, we have the estimate, for every $n \geq 1$,*

$$\left| \sum_{y \text{ in } Y(\mathbf{F}_{q^n})} \text{trace}(\text{Frob}_{q^n, y} | \mathscr{F}) \right| \leq (\dim H_c^1(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F}))(\sqrt{q})^n.$$

Proof. If π_1^{geom} acts irreducibly and nontrivially, the H_c^2 vanishes. As the H_c^1 is mixed of weight ≤ 1, the result is obvious from the Lefschetz trace formula. Q.E.D.

To complete the equidistribution story, there is one final actor we need to introduce, the "geometric monodromy group" G_{geom} of the sheaf $\mathscr{F}$. The definition of G_{geom} is very simple in terms of the representation ρ of π_1^{arith} in $\text{GL}(n, \overline{\mathbf{Q}}_\ell)$ which $\mathscr{F}$ "is":

G_{geom} := the Zariski closure of the image $\rho(\pi_1^{\text{geom}})$ in $\text{GL}(n, \overline{\mathbf{Q}}_\ell)$. One knows (by results of Grothendieck and Deligne, cf. [De 7]) that if $\mathscr{F}$ is pure, G_{geom} is a semisimple algebraic group. (For example, if $\mathscr{F}$ is the relative H^1 of a family of elliptic curves over Y with nonconstant j invariant, G_{geom} is the group $\text{SL}(2)$.)

To simplify the exposition, we will make the following hypothesis:

(Hyp)

Under the representation ρ, the image $\rho(\pi_1^{\text{arith}})$ lies is G_{geom}.

Now choose (!) an isomorphism $\overline{\mathbf{Q}}_\ell \approx \mathbf{C}$. This allows us to view G_{geom} as a semisimple group over $\mathbf{C}$, and to speak of the complex semisimple Lie group $G_{\text{geom}}(\mathbf{C})$. Let $K \subset G_{\text{geom}}(\mathbf{C})$ be a maximal compact subgroup of $G_{\text{geom}}(\mathbf{C})$. (For $G_{\text{geom}} = \text{SL}(2)$, we might take $K = \text{SU}(2)$.) Because we are under hypothesis (Hyp), it makes sense to consider the image of a Frobenius $\rho(\text{Frob}_{q^n, y}) \in G_{\text{geom}}(\mathbf{C})$. (Strictly speaking, it is a conjugacy class in $G_{\text{geom}}(\mathbf{C})$.) We know that this element has all its eigenvalues of absolute value one (this is the fact that $\mathscr{F}$ is pure of weight zero). We do not know that this element is semisimple (diagonalizable), but we can take the semisimple part of it Jordan decomposition $(\rho(\text{Frob}_{q^n, y}))^{ss}$, which is both semisimple and has unitary eigenvalues. So certainly $(\rho(\text{Frob}_{q^n, y}))^{ss}$ lies in a compact subgroup, even in a compact torus, of $G_{\text{geom}}(\mathbf{C})$. Since any compact subgroup of $G_{\text{geom}}(\mathbf{C})$ is conjugate to a subgroup of any chosen maximal compact subgroup K, we see that $(\rho(\text{Frob}_{q^n, y}))^{ss}$ is conjugate in $G_{\text{geom}}(\mathbf{C})$ to an element of K. However, it results easily from the Peter–Weyl theorem and the unitarian trick (the fact that the compact group K and the algebraic group G_{geom} have the "same" finite-dimensional representation theory) that if two elements of K are $G_{\text{geom}}(\mathbf{C})$-conjugate, then they are K-conjugate. What this means for us is that all possible methods of conjugating $(\rho(\text{Frob}_{q^n, y}))^{ss}$ into K lead to a single conjugacy class, which we denote $\vartheta(q^n, y)$, and think of as a "generalized angle of

Frobenius." (When G_{geom} is SL(2) and K is SU(2), the conjugacy classes in K are in bijective correspondence with the angles in $[0, \pi]$, by $\vartheta \mapsto \text{Diag}(e^{i\vartheta}, e^{-i\vartheta})$.)

Think of the space $K^\natural$ of conjugacy classes of K as the quotient space whose continuous functions are the continuous central functions on K. Denote by μ_{Haar} the normalized (total mass one) Haar measure on K, and by $\mu^\natural$ its direct image on $K^\natural$. The measure $\mu^\natural$ on $K^\natural$ is characterized as follows: For any continuous function f on $K^\natural$, viewed as a continuous central function $\tilde{f}$ on K, we have

$$\int_{K^\natural} f\, d\mu^\natural := \int_K \tilde{f}\, d\mu_{\text{Haar}}.$$

By Peter–Weyl, the $\mathbf{C}$-span of the characters f_Λ of the irreducible representations Λ of K are uniformly dense in the space of continuous central functions on K, and their integrals against Haar measures are given by

$$\int_{K^\natural} f_\Lambda\, d\mu^\natural = \begin{cases} 1, & \text{if } \Lambda \text{ is the trivial representation } \mathbf{1}, \\ 0, & \text{if } \Lambda \text{ is irreducible nontrivial.} \end{cases}$$

(For $K = \text{SU}(2)$, if we view $K^\natural$ as $[0, \pi]$, then $\mu_\natural$ is the Sato–Tate measure, and the functions $f_n(\vartheta) := \sin(n\vartheta)/\sin(\vartheta)$ are the characters of the irreducible representations of SU(2).)

Equidistribution theorem. (Deligne, Weil II). *Suppose $\mathscr{F}$ is a lisse $\overline{\mathbf{Q}}_\ell$-sheaf, pure of weight zero, on an open smooth geometrically connected curve Y over a finite field $\mathbf{F}_q$. Suppose that the hypothesis (Hyp) holds. Then the generalized angles of Frobenius $\{\vartheta(q^n, y)\}_y$ are equidistributed in $K^\natural$, with respect to the measure $\mu^\natural$, as n tends to ∞. More precisely, if we denote by μ_{q^n} the measure on $K^\natural$ which is "average over all the generalized angles of Frobenius $\vartheta(q^n, y)$ as y runs over $Y(\mathbf{F}_{q^n})$," then for every irreducible representation Λ of K we have the estimate*

$$\left| \int_{K^\natural} f_\Lambda\, d\mu_{q^n} - \int_{K^\natural} f_\Lambda\, d\mu^\natural \right| \leq c(\dim \Lambda)(\sqrt{q})^n / \text{Card}(Y(\mathbf{F}_{q^n})),$$

with c the constant

$$c := \text{rank}\,\mathscr{F} + |\chi_c(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})| \qquad \text{if } Y \text{ is } \mathbf{A}^1,$$

$$c := |\chi_c(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})| \qquad \text{for any } Y \neq \mathbf{A}^1.$$

Proof. Since $\mathrm{Card}(Y(\mathbf{F}_{q^n}))$ is about q^n, this an estimate certainly implies the equidistribution.

For Λ the trivial representation $\mathbf{1}$, this estimate is trivial (the left side vanishes identically for every n). Suppose now that Λ is irreducible nontrivial. Then the asserted estimate may be rewritten

$$\left| \sum_{y \text{ in } Y(\mathbf{F}_{q^n})} f_\Lambda(\vartheta(q^n, y)) \right| \leq c(\dim \Lambda)(\sqrt{q})^n.$$

We will prove it in this form. By the unitarian trick, Λ is the restriction to K of a unique representation Λ of G_{geom}. Via the chosen isomorphism $\overline{\mathbf{Q}}_\ell \approx \mathbf{C}$, Λ becomes ℓ-adic. Since $\rho(\pi_1^{\mathrm{arith}}) \subset G_{\mathrm{geom}}$, we can view the composite $\Lambda \circ \rho$ as a representation of π_1^{arith}, which then tautologically corresponds to a lisse sheaf "$\Lambda(\mathscr{F})$" on Y. This $\Lambda(\mathscr{F})$ is pure of weight zero (it occurs in some tensor power of $\mathscr{F}$), and π_1^{geom} acts irreducibly and nontrivially. Tautologically we have

$$f_\Lambda(\vartheta(q^n, y)) = \mathrm{trace}(\mathrm{Frob}_{q^n, y} | \Lambda(\mathscr{F})).$$

So the asserted estimate is just the Key Lemma, applied to $\Lambda(\mathscr{F})$, combined with an easy bound for $|\chi_c(Y \otimes \overline{\mathbf{F}}_q, \Lambda(\mathscr{F}))|$ in terms of $|\chi_c(Y \otimes \overline{\mathbf{F}}_q, \mathscr{F})|$. Q.E.D.

The moral of all this is that for a lisse pure sheaf on a curve over a finite field, there is always an equidistribution theorem, but we have to compute the group G_{geom} to know what it says!

Now we know what happens when we have a lisse pure sheaf on a curve over a finite field. But what happens when we have a smooth curve Y over, say, an open set $\mathrm{Spec}(\mathbf{Z}[\frac{1}{m}])$ of $\mathrm{Spec}(\mathbf{Z})$, and a lisse, pure of weight zero, sheaf $\mathscr{F}$ on Y. For each good prime p (i.e., $(p, m) = 1$), we get a situation of the type we have just considered on the fibers $Y \otimes_{\mathbf{Z}} \mathbf{F}_p$. So for each good prime p, there is an equidistribution theorem for the angles $\vartheta(p^n, y)$ in terms of the group $G_{\mathrm{geom}, p}$ for that characteristic, and there are explicit estimates which involve the constant $|\chi_c(Y \otimes_{\mathbf{Z}} \overline{\mathbf{F}}_p, \mathscr{F})|$. The reason we care about this is: If we knew that both the group $G_{\mathrm{geom}, p}$ and the constant $|\chi_c(Y \otimes_{\mathbf{Z}} \overline{\mathbf{F}}_p, \mathscr{F})|$ were independent of p, then (ignoring the question of the hypothesis (Hyp)), the probabilistic theorem of the first lecture would show (see Appendix) that a random horizontal sequence of angles $\{\vartheta(p, y_p)\}_p$ is equidistributed in $K^\natural$, with respect to the measure $\mu^\natural$. While this tells us

nothing about any particular horizontal sequence, and in particular nothing about those where we start with a $\mathbf{Z}[\frac{1}{m}]$-valued point y on Y and take $y_p := y \bmod p$, it does at least tell us what we should expect to see as the results of numerical experiments.

For example, consider the situation of Example 3 of the first lecture:

$$R = \mathbf{Z}[T, 1/(6T(T-27))],$$
$$Y := \mathrm{Spec}(R) = \mathbf{A}^1 - \{0, 27\} \text{ is smooth over } \mathbf{Z}[\tfrac{1}{6}].$$

Over Y, we have the affine curve V of equation

$$y^2 = 4x^3 - Tx - T,$$

and the natural projection map

$$f\colon V \to Y,\ (x, y, T) \mapsto T.$$

The sheaf $\mathscr{F} := R^1 f_! \mathbf{Q}_\ell(\frac{1}{2})$ is lisse of rank two and pure of weight zero on $Y[1/\ell]$, and its traces of Frobenius are given by

$$\mathrm{trace}(\mathrm{Frob}_{q,t} | \mathscr{F}) = 2\cos(\vartheta_{q,t}).$$

Here we know ("by inspection") that for each $p \neq 2, 3, \ell$ we have $G_{\mathrm{geom},p} = \mathrm{SL}(2)$, and $|\chi_c(Y \otimes_{\mathbf{Z}} \overline{\mathbf{F}}_p, \mathscr{F})| = 2$.

Fortunately, the general case behaves in exactly the same way, as results from quite general semicontinuity theorems for ℓ-adic cohomology, and the theory of specialization for π_1. For simplicity, we will state it over $\mathbf{Z}$, but one could just as well replace $\mathbf{Z}$ by a subring of $\mathbf{C}$ which is finitely generated as a $\mathbf{Z}$-algebra.

Theorem (cf. [De 7, 1.11.5; Ka 7, 8.18]). *Let $m \geq 1$ be an integer, and let Y be smooth curve over $\mathbf{Z}[1/m\ell]$ with geometrically connected fibers. Let $\mathscr{F}$ be a lisse $\overline{\mathbf{Q}}_\ell$-adic sheaf on Y, say of rank n. Then for almost all primes p, we have an equality of geometric monodromy groups*

$$G_{\mathrm{geom},p} = G_{\mathrm{geom},\mathbf{C}} \text{ (up to conjugacy in } \mathrm{GL}(n)),$$

and an equality of Euler characteristics

$$\chi_c(Y \otimes_{\mathbf{Z}} \overline{\mathbf{F}}_p, \mathscr{F}) = \chi_c(Y \otimes_{\mathbf{Z}} \mathbf{C}, \mathscr{F}).$$

This concludes our discussion of the exponential sums

$$\sum_{x \in V_\varphi(k)} \psi(f(x))$$

in the case when ψ is the trivial character!!! In the next lecture, we will (finally) begin discussing the case when ψ is nontrivial.

FOURTH LECTURE

So far, in a series of lectures which purport to be about exponential sums and differential equations, we have seen precious few exponential sums, and not a single differential equation. We will now attempt to remedy the situation.

First of all, how can we put exponential sums into the context of ℓ-adic sheaves? Given the general formalism, what we need is this: Given an affine variety V over a finite field k, a function f on V, and a $\overline{\mathbf{Q}}_\ell$-valued nontrivial additive character ψ of k, we need a lisse, rank one $\overline{\mathbf{Q}}_\ell$-sheaf $\mathscr{L}_{\psi(f)}$ on V such that for any finite extension E of k, and any point v in $V(E)$, $\mathrm{Frob}_{E,v}$ acts on $\mathscr{L}_{\psi(f)}$ by the scalar $\psi_E(f(v)) := \psi(\mathrm{trace}_{E/k}(f(v)))$. The good news is that such an $\mathscr{L}_{\psi(f)}$ exists (cf. [Gro 2, De 2]). The bad news, from the point of view of "doing things over $\mathbf{Z}$," is that it is constructed out of the "Artin-Schreier covering" of V defined in $V \times \mathbf{A}^1$ by the equation $y^q - y = f(v)$, which has q appearing as an exponent.

So if we start with the data $(V/k, f, \psi)$, then the Lefschetz trace formula shows that for every finite extension E of k, we have

$$\sum_{v \text{ in } V(E)} \psi_E(f(v)) = \sum_i (-1)^i \text{ trace } (\mathrm{Frob}_{E,v} | H_c^i(V \otimes \overline{k}, \mathscr{L}_{\psi(f)})).$$

So to study a single exponential sum, it is enough to study the various cohomology groups $H_c^i(V \otimes \overline{k}, \mathscr{L}_{\psi(f)})$. But where are the promised one-parameter families?

We have already seen that changing the choice of nontrivial character ψ amounts to scaling the function f. If we want to consider all possible ψ's over all possible finite extensions E of k, we might as well introduce the parameter variety $\mathbf{G}_{m,k} := \mathrm{Spec}(k[T, \frac{1}{T}])$, the "multiplicative group $\mathbf{G}_m$ over k," consider the product variety $V \times_k \mathbf{G}_{m,k}$, and the function Tf on it:

$$\begin{array}{ccc} V \times_k \mathbf{G}_{m,k} & \xrightarrow{(t,v) \mapsto tf(v)} & \mathbf{A}^1_k \\ {\scriptstyle \mathrm{pr}_2}\downarrow & & \\ \mathbf{G}_{m,k}. & & \end{array}$$

The general formalism of ℓ-adic cohomology tells us that for each i, we simultaneously recover all the cohomology groups

$\{H_c^i(V \otimes \overline{k}, \mathscr{L}_{\psi(tf)})\}$ as the stalks of the direct image sheaf $R^i(\mathrm{pr}_2)_!(\mathscr{L}_{\psi(Tf)})$ on $\mathbf{G}_{m,k}$.

In "favorable" circumstances, we will be able to prove that the sheaves $R^i(\mathrm{pr}_2)_!(\mathscr{L}_{\psi(Tf)})$ vanish for all but one value, say n, of i, and that this remaining sheaf $R^n(pr_2)_!(\mathscr{L}_{\psi(Tf)})$ is lisse on $\mathbf{G}_{m,k}$, pure of weight n, and it is suitably dual to the same sheaf formed with $\overline{\psi}$. As you can imagine, there is a fair amount of "high technology" which goes into deciding exactly what constitutes favorable circumstances, so I will pass over that entirely, though later I will tell you some examples where we know circumstances to be favorable.

Another obvious way to introduce a parameter into the situation is to consider a second function g on V, and then to look at the family of functions $f + Tg$, where now T is an additive parameter. This will lead to looking at sheaves on $\mathbf{A}^1_k$ rather than on $\mathbf{G}_{m,k}$. Once again, there are lots of circumstances known to be favorable.

Now we come to the real point of these lectures. The (V, f) or (V, f, g) data that go into making these families make sense over $\mathbf{Z}$. And there are lots of cases of data (V, f) or (V, f, g) over $\mathbf{Z}$ for which circumstances are favorable for almost all p, and any nontrivial ψ. And as explained in the first lecture, one is interested in the equidistribution properties of these sums. As a minor variant on this, one might give on V some lisse sheaf $\mathscr{G}$ which is pure of weight zero, rather than the constant sheaf, and look at the families of sums

$$\sum_{v \text{ in } V(E)} \psi_E(tf(v)) \operatorname{trace}(\mathrm{Frob}_{E,v} | \mathscr{G}),$$

or

$$\sum_{v \text{ in } V(E)} \psi_E(f(v) + tg(v)) \operatorname{trace}(\mathrm{Frob}_{E,v} | \mathscr{G}).$$

There are many situations of this type $(V, \mathscr{G}, f)$ "over $\mathbf{Z}$" which are also "favorable" in the above sense. Let us suppose that we are in such a case. Then the equidistribution is governed, in each good characteristic p and for each nontrivial ψ, by a certain lisse sheaf

$$\mathscr{F}_{p,\psi} := R^n(pr_2)_!(\mathscr{G} \otimes \mathscr{L}_{\psi(Tf)})(n/2)$$

or

$$R^n(pr_2)_!(\mathscr{G} \otimes \mathscr{L}_{\psi(f+Tg)})(n/2)$$

on the parameter space, which is pure of weight zero and lisse of some rank $r(p, \psi)$, a certain semisimple algebraic group

$$G_{\text{geom}, p, \psi} \subset \mathrm{GL}(r(p, \psi)),$$

and the Euler characteristic of the parameter space with coefficients on the sheaf $\mathscr{F}_{p, \psi}$.

Again by using some high technology which I cannot go into here, one can compute all of these data in many particular cases (cf. [Ka 3, Ka 5, Ka 7]). Essentially every time one can carry out such a computation (if there is time later I will explain the proviso "essentially"), one finds that the ranks $r(p, \psi)$ of the sheaves $\mathscr{F}_{p, \psi}$, the semisimple algebraic groups

$$G_{\text{geom}, p, \psi} \subset \mathrm{GL}(r(p, \psi)),$$

and the Euler characteristic of the parameter space with coefficients on the sheaf $\mathscr{F}_{p, \psi}$ **are all equal**, provided that we exclude finitely many bad p.

Here is a table of the common values of these quantities in some very simple examples of favorable situations (cf. [Ka 5, Ka 7]). In it, $\mathscr{L}_{\chi_2(x)}$ is the lisse rank one Kummer sheaf $\mathscr{L}_{\chi_2(x)}$ on $\mathbf{G}_m$, on which $\mathrm{Frob}_{E, x}$ acts by the quadratic character $\chi_{2, E}(x)$.

V	$\mathscr{G}$ on V	Family of Functions	r	G_{geom}	
$\mathbf{A}^1$	$\overline{\mathbf{Q}}_\ell$	$x^n + tx$	$n-1$	$\mathrm{SL}(n-1)$	if n even
$\mathbf{A}^1$	$\overline{\mathbf{Q}}_\ell$	$x^n + tx$	$n-1$	$\mathrm{Sp}(n-1)$	if n odd
$\mathbf{G}_m$	$\mathscr{L}_{\chi_2(x)}$	$x^n + tx$	n	$\mathrm{SL}(n)$	if n even, ≥ 4
$\mathbf{G}_m$	$\mathscr{L}_{\chi_2(x)}$	$x^n + tx$	n	$\mathrm{SO}(n)$	if n odd, $n \neq 7$
$\mathbf{G}_m$	$\mathscr{L}_{\chi_2(x)}$	$x^7 + tx$	7	G_2 (inside SO(7))	
$\mathbf{G}_m$	$\overline{\mathbf{Q}}_\ell$	$t(x + 1/x)$	2	$\mathrm{SL}(2)$	

Challenge. *Explain why G_2 occurs for the sums*

$$\sum_{x \neq 0 \text{ in } \mathbf{F}_q} \chi_2(x) \psi(x^7 + tx)$$

in every characteristic $p \geq 17$.

Let us return now to the general setup. I am indebted to Bill Messing for asking me what in retrospect is the obvious question about the fact that (in all the computed examples) for p large, all the $r(p, \psi)$ are equal, all the $G_{\text{geom}, p, \psi}$ are equal, and all the

Euler characteristics are equal: **If they are equal, what are they all equal to?**

Once this question is posed, the conjectural answer leaps out. To explain what it is, let us begin with the case when $\mathscr{G}$ is the constant sheaf, and V is an open set of $\mathbf{A}^1$. The exponential sums we are looking at,

$$\sum_{v \text{ in } V(E)} \psi_E(tf(v)),$$

or

$$\sum_{v \text{ in } V(E)} \psi_E(f(v) + tg(v)),$$

can be regarded as finite field analogues of Fourier integrals over $V_{\mathbf{C}}$

$$\int e^{tf(v)}\,dv, \quad \int e^{f(v)+tg(v)}\,dv.$$

Neglecting questions of convergence (or even of definition!), such formal "integrals" can be differentiated with respect to the parameter t. There is an obvious calculus of formal integrals of the form

$$\int h(v)e^{tf(v)}\,dv, \quad \int h(v)e^{f(v)+tg(v)}\,dv,$$

where h is allowed to be any $\mathbf{C}[t, \frac{1}{t}]$ linear (resp. $\mathbf{C}[t]$-linear) combination of polynomial functions on $V_{\mathbf{C}}$, and the only relations imposed are $\mathbf{C}[t, \frac{1}{t}]$-linearity (resp. $\mathbf{C}[t]$-linearity) in h, and a formal Stokes formula. The set of all such formal integrals, modulo these relations, is a $\mathbf{C}[t, \frac{1}{t}]$-module (resp. $\mathbf{C}[t]$-module) on which $d := \frac{d}{dt}$ operates. In other words, it is an algebraic $\mathscr{D}$-module, say $\mathscr{M}$ on $Y = \mathbf{G}_{m,\mathbf{C}}$ (resp. on $Y = \mathbf{A}^1_{\mathbf{C}}$). (See [Ber] and [Bor] for the general theory of algebraic $\mathscr{D}$-modules.) Moreover, in the examples above, one can see by explicit calculation that:

(1) $\mathscr{M}$ is $\mathscr{O}_Y$-locally free, of finite rank equal to the common rank r.

(2) The Euler characteristic of Y with coefficients in $\mathscr{M}$, in the sense of algebraic $\mathscr{D}$-modules ("algebraic de Rham cohomology" in the old terminology) is equal to the common Euler characteristic.

What about the common value of the groups G_{geom}? Ever since Picard–Vessiot, there has been the notion of attaching an algebraic group G_{gal}, the differential Galois group, to a linear differential

equation, i.e., to an $\mathscr{O}$-locally free of finite rank $\mathscr{D}$-module $\mathscr{M}$. If $\mathscr{M}$ is of rank r, then G_{gal} is an algebraic subgroup of $\mathrm{GL}(r)$. It thus seems foreordained to make the

Conjecture. The common value of the G_{geom}'s is the differential Galois group G_{gal}.

In the last few minutes, I would like to tell what we know and what we do not know about the truth of this conjecture. Recall that there were two exponential sum situations which we considered:

$$\sum_{v \text{ in } V(E)} \psi_E(tf(v)) \text{ trace } (\text{Frob}_{E,v} | \mathscr{G}),$$

or

$$\sum_{v \text{ in } V(E)} \psi_E(f(v) + tg(v)) \text{ trace } (\text{Frob}_{E,v} | \mathscr{G}).$$

We will refer to these as the first case and the second case.

Theorem. ([Ka 7, 14.6]). *Suppose we are in a favorable instance of the first case. Denote by $\mathscr{M}$ the associated differential equation. If G_{gal} is semisimple, then for all p sufficiently large, and all nontrivial ψ, we have*

(1) $\text{rank}\,\mathscr{F}_{p,\psi} := r(p, \psi) = \text{rank}\,\mathscr{M}$.

(2) $\chi_c(Y \otimes \overline{\mathbf{F}}_p, \mathscr{F}_{p,\psi}) = \chi_{DR}(Y_{\mathbf{C}}, \mathscr{M})$.

(3) $G_{\text{geom},p,\psi} = G_{\text{gal}}$ *(more precisely, via the isomorphism $\overline{\mathbf{Q}}_\ell \approx \mathbf{C}$, $G_{\text{geom},p,\psi}$ is conjugate to G_{gal}).*

Sketch of proof. To fix ideas, suppose that $\mathscr{G}$ is the constant sheaf. Then $\mathscr{F}_{p,\psi}$ is the ℓ-adic Fourier transform of $Rf_!\overline{\mathbf{Q}}_\ell$, and $\mathscr{M}$ is the $\mathscr{D}$-module Fourier transform of the corresponding (by the "Riemann–Hilbert correspondence," the $\mathscr{D}$-module version of Hilbert's 21st Problem) object in the world of RS ("regular singular," the $\mathscr{D}$-module version of regular singularities in the sense of Fuchs for an ordinary differential equation) holonomic $\mathscr{D}$-modules. All the groups involved are semisimple, so they can be recovered if one knows their algebras of tensor invariants. But tensor product on the Fourier transform side is convolution on the other side. And "invariants" on the Fourier transform side is "stalk at the origin" on the other side. Both the notions of convolution and of "stalk at the origin" make sense "over $\mathbf{Z}$." Now apply the standard semicontinuity results of ℓ-adic life on schemes over $\mathbf{Z}$ to $Rf_!\overline{\mathbf{Q}}_\ell$ and to its multiple convolutions with itself, and use the

fact that on the C-fiber all of this has a Riemann–Hilbert translation in the world of RS holonomic $\mathscr{D}$-modules. Q.E.D.

Remark. The hypothesis that G_{gal} be semisimple is essential. Here is an example to show why. Take for V the subvariety of $\mathbf{A}^1$ which consists of the point 1, i.e., $V := \text{Spec}(\mathbf{Z}[x]/(x-1))$, and f the function "x." Then our family of exponential sums is simply $t \mapsto \psi(t)$, and so the sheaf $\mathscr{F}_{p,\psi}$ is $\mathscr{L}_\psi$. Therefore $G_{\text{geom},p,\psi}$ is the finite subgroup μ_p of GL(1). The corresponding $\mathscr{D}$-module $\mathscr{M}$ is the one given by the function e^x, and its G_{gal} is GL(1). So the theorem is false, but G_{gal} is still reductive. It is the smallest reductive group which contains almost all the G_{geom}'s, and for $p \gg 0$ both G_{geom} and G_{gal} have the same $G^{0,\text{der}}$, and G_{geom} maps onto every semisimple quotient of G_{gal}. This can be shown to be the general pattern:

Theorem bis. ([Ka 7, 14.10, 14.11.2]. *Suppose we are in a favorable instance of the first case. Denote by $\mathscr{M}$ the associated differential equation. Then G_{gal} is reductive, and for all p sufficiently large, and all nontrivial ψ, we have:*

(1) $\text{rank}\,\mathscr{F}_{p,\psi} := r(p,\psi) = \text{rank}\,\mathscr{M}$.

(2) $\chi_c(Y \otimes \overline{\mathbf{F}}_p, \mathscr{F}_{p,\psi}) = \chi_{DR}(Y_{\mathbf{C}}, \mathscr{M})$.

(3) $G_{\text{geom},p,\psi} \subset G_{\text{gal}}$ (*more precisely, via the isomorphism $\overline{\mathbf{Q}}_\ell \approx \mathbf{C}$, $G_{\text{geom},p,\psi}$ is conjugate a subgroup of G_{gal}*), *both $G_{\text{geom},p,\psi}$ and G_{gal} have the same $G^{0,\text{der}}$, and $G_{\text{geom},p,\psi}$ maps onto every semisimple quotient of G_{gal}.*

Moreover, no proper reductive subgroup of G_{gal} contains almost all the groups G_{geom}.

(When G_{gal} is semisimple, this reduces to the theorem above.)

What about the second case? Here there is no theorem, just a big body of "experimental evidence." For instance, consider the G_2 example,

$$\sum_{x \neq 0 \text{ in } \mathbf{F}_q} \chi_2(x)\psi(x^7 + tx).$$

The corresponding integral is

$$\int (\sqrt{1/x}) \exp(x^7 + tx)\, dx,$$

which is killed by the seventh order differential operator in

$\partial := d/dt$

$$7\partial^7 + t\partial + \tfrac{1}{2}.$$

Once one is convinced by the analogy with exponential sums that this differential equation should have G_{gal} the subgroup G_2 of $\mathrm{SO}(7)$, it turns out to be not difficult to prove it (cf. [Ka 7, 2.10.5, 2.10.6]). But I do not know why it is true.

Thus our overall situation is this. We start with a conjecture of the form "$X = Y$." Since we do not know that it is true, we test cases by separately computing the two sides. Whenever we can compute both sides, we find that they are both equal to the same Z. In doing this, we learn more than that $X = Y$ is true: We learn that $X = Z = Y$ with an explicit Z. In our proof of the conjecture in the first case, we prove that $X = Y$ by proving, as it were, that $X - Y = 0$, without knowing either X or Y. But one also wants to know what the common value is. On the other hand, in the second case, we do not even know in what universe it could make sense to subtract $X - Y$.

This pretty much exhausts what I had hoped to explain here. In the remaining few minutes, I want to tie up a few loose ends.

First, there is the question of "unifying" the above discussion with what we did in the "ψ trivial" case, where we also saw that G_{geom} was independent of $p \gg 0$. There we had a lisse $\mathscr{F}$ on a smooth curve $Y/\mathbf{Z}$, and we were looking at the G_{geom}'s for the restrictions of $\mathscr{F}$ to the geometric fibers. By general ℓ-adic facts, we knew that for $p \gg 0$, all these G_{geom}'s were the same as that for the complex fiber, i.e., for $\mathscr{F}_{\mathbf{C}}$ on the complex curve $Y_{\mathbf{C}}$. Again by Riemann–Hilbert, as soon as we pick an isomorphism $\overline{\mathbf{Q}}_\ell \approx \mathbf{C}$, the local system $\mathscr{F}_{\mathbf{C}}$ on $Y_{\mathbf{C}}$ corresponds to an RS differential equation, say $\mathscr{M}$, on $Y_{\mathbf{C}}$. Because this $\mathscr{M}$ is RS, its G_{gal} is equal to the G_{mono} for $\mathscr{F}_{\mathbf{C}}$. So in the case of "ψ trivial" as well, the common value of the G_{geom} for $p \gg 0$ has a natural interpretation as the G_{gal} of an associated differential equation.

Second, there is the question of finding a general conceptual framework in which to think about the variation with p of exponential sums on arbitrary schemes of finite type over $\mathbf{Z}$. In the "ψ trivial" case, this framework is the theory of constructible ℓ-adic sheaves (for the full theory one needs the ℓ-adic derived category) on such schemes, and of Grothendieck's "six operations" operations ($Rf_!$, Rf_*, f^*, $f^!$, $\otimes$, $R\mathrm{Hom}$) on them. What we need now is an over-world of this "classical" world, which is stable by the

six operations, and which in addition contains one new object on $\mathbf{A}^1/\mathbf{Z}$ which gives rise to the sheaves $\mathscr{L}_{\psi(x)}$ on the characteristic p fibers, and which gives rise to the $\mathscr{D}$-module for e^x on the complex fiber. If we have such an object in such an over-world, call it $\mathbf{E}$, then for any scheme X of finite type over $\mathbf{Z}$, and any function f on X, viewed as a map to $\mathbf{A}^1$, we can use f to form the pullback $f^*\mathbf{E}$ on X. This pullback on X will incarnate the sheaves $\mathscr{L}_{\psi(f)}$ on the characteristic p fibers of $X/\mathbf{Z}$, and the $\mathscr{D}$-module for $e^{f(x)}$ on the $\mathbf{C}$-fiber. Given a usual ℓ-adic sheaf $\mathscr{F}$ (or a derived category object) on our X, and a morphism $\pi: X \to Y$ of schemes of finite type over $\mathbf{Z}$,

$$\begin{array}{lccc} \mathscr{F} \text{ on } & X & \xrightarrow{f} & \mathbf{A}^1 \\ & {\scriptstyle\pi}\downarrow & & \\ & Y, & & \end{array}$$

it would then make sense to form $R\pi_!(\mathscr{F} \otimes f^*\mathbf{E})$ on Y, which would incarnate $R\pi_!(\mathscr{F} \otimes \mathscr{L}_{\psi(f)})$ on the characteristic p fibers, and the $\mathscr{D}$-module derived category object $R\pi_!(\mathscr{F} \otimes e^f)$ on the $\mathbf{C}$-fiber. So our over-world would contain all objects on Y of the form $R\pi_!(\mathscr{F} \otimes f^*\mathbf{E})$ for all diagrams

$$\begin{array}{lccc} \mathscr{F} \text{ on } & X & \xrightarrow{f} & \mathbf{A}^1 \\ & {\scriptstyle\pi}\downarrow & & \\ & Y. & & \end{array}$$

There is a simplification possible in forming $R\pi_!(\mathscr{F} \otimes f^*\mathbf{E})$ on Y. Namely, we can factor the map π through $\pi \times f$, and the diagram becomes

$$\begin{array}{ccccc} & & X & & \\ & {\scriptstyle\pi\times f}\swarrow & & \searrow{\scriptstyle f} & \\ Y & & \xleftarrow{pr_2} & & \mathbf{A}^1, \mathbf{E} \\ & {\scriptstyle pr_1}\searrow & & & \\ & & Y. & & \end{array}$$

So we should have

$$\begin{aligned} R\pi_!(\mathscr{F} \otimes f^*\mathbf{E}) &= R(pr_1)_!R(\pi \times f)_!(\mathscr{F} \otimes f^*\mathbf{E}) \\ &= R(pr_1)_!R(\pi \times f)_!(\mathscr{F} \otimes (\pi \times f)^*(pr_2)^*\mathbf{E}), \end{aligned}$$

which by the projection formula is

$$R(pr_1)_!(R(\pi \times f)_!\mathscr{F} \otimes (pr_2)^*\mathbf{E}).$$

Since the object $\mathscr{G} := R(\pi \times f)_!\mathscr{F}$ on $Y \times \mathbf{A}^1$ makes perfect sense in the classical world, the objects we are looking at are all of the form

$$R(pr_1)_!(\mathscr{F} \otimes (pr_2)^*\mathbf{E})$$

for the single diagram

$$\begin{array}{ccc} Y \times \mathbf{A}^1 & \xrightarrow{pr_2} & \mathbf{A}^1, \mathbf{E} \\ {\scriptstyle pr_1}\downarrow & & \\ Y, & & \end{array}$$

and variable classical objects $\mathscr{F}$ on $Y \times \mathbf{A}^1$. Let us say that the over-world object $R(pr_1)_!(\mathscr{F} \otimes (pr_1)^*\mathbf{E})$ on Y is represented by the classical object $\mathscr{F}$ on $Y \times \mathbf{A}^1$.

As Richard Pink pointed out to me in February, 1988, the class of all these objects $R(pr_1)_!(\mathscr{F} \otimes (pr_2)^*\mathbf{E})$ on Y has very nice stability properties:

(0) For any function $f: Y \to \mathbf{A}^1$ on Y, the object $f^*\mathbf{E}$ on Y is represented by the constant sheaf $\overline{\mathbf{Q}}_\ell$ on the graph of f, extended by zero to all of $Y \times \mathbf{A}^1$.

(1) A classical object $\mathscr{F}$ on Y is represented by the "delta-object" on $Y \times \mathbf{A}^1$ which is "$\mathscr{F}$ on $Y \times (t = 0)$, extended by zero to all of $Y \times \mathbf{A}^1$."

(2) (f^* stability) for a morphism $f: X \to Y$ of schemes of finite type over $\mathbf{Z}$, and K an over-world object on Y represented by $\mathscr{F}$ on $Y \times \mathbf{A}^1$, f^*K on X is represented by $(f \times id)^*\mathscr{F}$ on $X \times \mathbf{A}^1$.

(3) ($f_!$ stability) for a morphism $f: Y \to Z$ of schemes of finite type over $\mathbf{Z}$, and K an over-world object on Y represented by $\mathscr{F}$ on $Y \times \mathbf{A}^1$, $Rf_!K$ on Z is represented by $R(f \times id)_!\mathscr{F}$ on $Z \times \mathbf{A}^1$.

(4) ($\otimes$ stability) given two objects on Y, represented by classical objects $\mathscr{F}$ and $\mathscr{G}$ on $Y \times \mathbf{A}^1$, their "tensor product" is represented by the classical object $\mathscr{F} * \mathscr{G}$ on $Y \times \mathbf{A}^1$ which is their additive convolution,

$$\mathscr{F} * \mathscr{G} := R(\mathrm{sum})_!(\mathscr{F} \times \mathscr{G}),$$

defined by viewing $Y \times \mathbf{A}^1$ as the Y-group scheme $\mathbf{A}^1_Y$.

This "calculus of the representing object" suggests a way we might begin to construct such an over-world. It should be a fibered category $\mathscr{E}$ over the category of schemes of finite type over $\mathbf{Z}$,

whose objects on a scheme Y are all the classical objects on $Y \times \mathbf{A}^1$:

$$\text{Objects}(\mathscr{E}_Y) = \text{Objects}(D_c^b(Y \times \mathbf{A}^1, \overline{\mathbf{Q}}_\ell)).$$

But what about the morphisms in $\mathscr{E}_Y$. These should be some sort of localization of the morphisms in $D_c^b(Y \times \mathbf{A}^1, \overline{\mathbf{Q}}_\ell)$, the idea being to invert (at least certain of) those morphisms which induce isomorphisms on all the "realizations" $R(pr_1)_!(\mathscr{F} \otimes (pr_2)^* \mathscr{L}_\psi)$, ψ nontrivial, on $Y \otimes \overline{\mathbf{F}}_p$ and $R(pr_1)_!(\mathscr{F} \otimes (pr_2)^* e^x)$ on $Y_{\mathbf{C}}$. At the very least one wants all constant-in-x objects (i.e., those on $Y \times \mathbf{A}^1$ of the form $(pr_1)^*(\mathscr{G})$) to be identified to the zero object; exactly this is done by applying the idempotent functor "convolution with $Rj_* \overline{\mathbf{Q}}_\ell[1](1)$", where j denotes the inclusion of $Y \times \mathbf{G}_m$ into $Y \times \mathbf{A}^1$. One could then define the $\mathscr{E}_Y$-homs between two objects $\mathscr{F}$, $\mathscr{G}$ of $D_c^b(Y \times \mathbf{A}^1, \overline{\mathbf{Q}}_\ell)$ to be the usual $D_c^b(Y \times \mathbf{A}^1, \overline{\mathbf{Q}}_\ell)$-homs between their convolutions with $Rj_* \overline{\mathbf{Q}}_\ell[1](1)$. Does this lead to a reasonable over-world? Need one invert more? Nothing is known!

Now let us return to the world of the concrete. Suppose we are given a scheme Y of finite type over $\mathbf{Z}$, and an ℓ-adic sheaf (or derived category object) $\mathscr{F}$ on $Y \times \mathbf{A}^1$, with coordinates (y, x). For each p and for each nontrivial character ψ of a finite subfield of $\overline{\mathbf{F}}_p$, we can form the ℓ-adic derived category object

$$K_p := R(pr_1)_!(\mathscr{F} \otimes \mathscr{L}_{\psi(x)}) \text{ on } Y \otimes \overline{\mathbf{F}}_p,$$

and the $\mathscr{D}$-module derived category object

$$K_{\mathbf{C}} := R(pr_1)_!(\mathscr{F} \otimes e^x) \text{ on } Y_{\mathbf{C}}.$$

In this generality, what can we hope for?

It is known that there exists an integer $N > 0$ and a dense open set U of $Y[\frac{1}{N}]$ such that:

(1) For $p \gg 0$, the cohomology sheaves $\mathscr{H}^i(K_p)$ are lisse on $U \otimes \overline{\mathbf{F}}_p$, of some common rank r_i.

(2) The $\mathscr{D}$-module $\mathscr{H}^i(K_{\mathbf{C}})|U_{\mathbf{C}}$ is a differential equation (i.e., $\mathscr{O}$-locally free) of some rank s_i on $U_{\mathbf{C}}$.

(3) $\sum(-1)^i r_i = \sum(-1)^i s_i$.

General Conjecture. For each i, we have

(1) $r_i = s_i$,

(2) for $p \gg 0$, $G_{\mathrm{geom},p}(:=$ the G_{geom} for $\mathscr{H}^i(K_p)|U \otimes \overline{\mathbf{F}}_p)$ is conjugate in $\mathrm{GL}(r_i)$ to a subgroup of

$$G_{\mathrm{gal}}(:=\text{ the } G_{\mathrm{gal}} \text{ for } \mathscr{H}^i(K_{\mathbf{C}})|U_{\mathbf{C}}),$$

and G_{gal} is minimal among algebraic subgroups of $\mathrm{GL}(r_i)$ with this property.

Much remains to be done.

Appendix: Horizontal versus vertical equidistribution

In this appendix, we consider the following general situation.

Fix an integer $r \geq 1$, a prime number ℓ, an algebraic closure $\overline{\mathbf{Q}}_\ell$ of $\mathbf{Q}_\ell$, and an isomorphism of fields $\iota: \overline{\mathbf{Q}}_\ell \approx \mathbf{C}$. Let S be a scheme of finite type over $\mathbf{Z}[1/\ell]$, and Y/S a smooth affine curve over S. We suppose that Y/S is of the form $\overline{Y} - D$, for some proper smooth curve $\overline{Y}/S$ with geometrically connected fibers and a divisor D in $\overline{Y}$ which is finite etale over S, of some constant degree $d \geq 1$. For each finite field $\mathbf{F}$, and each $\mathbf{F}$-valued point s in $S(\mathbf{F})$, we denote by $Y_s/\mathbf{F}$ the fiber of Y/S over the $\mathbf{F}$-valued point s, and we fix a geometric point ξ_s of Y_s. Suppose that for each finite field $\mathbf{F}$, and each $\mathbf{F}$-valued point s in $S(\mathbf{F})$, we are given a lisse of rank r $\overline{\mathbf{Q}}_\ell$-sheaf $\mathscr{F}_s$ on Y_s which is ι-pure of weight zero, and which satisfies the hypothesis (Hyp): under the representation

$$\rho_s: \pi_1(Y_s, \xi_s) \to \mathrm{GL}(\mathscr{F}_{\xi_s})$$

which $\mathscr{F}_s$ "is," $\rho_s(\pi_1(Y_s, \xi_s)) \subset G_{\mathrm{geom},s}$. (Here $G_{\mathrm{geom},s}$ denotes the Zariski closure in $\mathrm{GL}(\mathscr{F}_{\xi_s})$ of $\rho_s(\pi_1^{\mathrm{geom}}(Y_s, \xi_s))$.)

We suppose that there exists a (necessarily semisimple) algebraic subgroup G of $\mathrm{GL}(r, \overline{\mathbf{Q}}_\ell)$, an integer $M \geq 1$, and for each $(\mathbf{F}, s \text{ in } S(\mathbf{F}))$ as above an isomorphism $\mathscr{F}_{\xi_s} \approx (\overline{\mathbf{Q}}_\ell)^r$ such that

(1) $|\chi_c(Y_s \otimes_{\mathbf{F}} \overline{\mathbf{F}}, \mathscr{F}_s)| \leq M$.

(2) Via the given isomorphism $\mathscr{F}_{\xi_s} \approx (\overline{\mathbf{Q}}_\ell)^r$, $G_{\mathrm{geom},s} = G$ (equality inside $\mathrm{GL}(r, \overline{\mathbf{Q}}_\ell)$).

Choose any sequence $(\mathbf{F}_i, s_i \text{ in } S(\mathbf{F}_i))$ in which $q_i := \mathrm{Card}(\mathbf{F}_i)$ tends Archimedeanly to infinity and for which each finite set $Y_{s_i}(\mathbf{F}_i)$ is nonempty, and denote by $X := \Pi_i Y_{s_i}(\mathbf{F}_i)$ the corresponding product space. Endow each factor $Y_{s_i}(\mathbf{F}_i)$ with the probability measure which gives each point the same measure $1/\mathrm{Card}(Y_{s_i}(\mathbf{F}_i))$, and endow X with the product measure μ_{prod}.

Choose a maximal compact subgroup K of the complex semi-simple group $G(\mathbf{C})$, $\mathbf{C}$ viewed as a $\overline{\mathbf{Q}}_\ell$-algebra via ι, and denote by $K^\natural$ the space of conjugacy classes of K, and by $\mu^\natural$ the projection to $K^\natural$ of Haar measure on K. For each point y_i in each factor $Y_{s_i}(\mathbf{F}_i)$, denote by $\vartheta(y_i, s_i, \mathbf{F}_i)$ in $K^\natural$ the corresponding "angle" of Frobenius. (This "angle" in $K^\natural$ makes sense by the hypothesis (Hyp), and the hypothesis that all the groups $G_{\text{geom}, s}$ coincide with G.)

Theorem (joint with Ofer Gabber). *Notations and hypotheses as above, the set of points* $x := (y_i)_i$ *in* $X := \Pi_i Y_{s_i}(\mathbf{F}_i)$ *for which the corresponding sequence of "angles"* $(\vartheta(y_i, s_i, \mathbf{F}_i))_i$ *in* $K^\natural$ *is equidistributed with respect to* $\mu^\natural$ *is a set of measure one in* X *with respect to* μ_{prod}.

Proof. For each of the countably many isomorphism classes of irreducible representations Λ of the compact group K, we denote by $f_\Lambda(\vartheta)$ its trace function, viewed as a continuous $\mathbf{C}$-valued function on $K^\natural$. By Peter–Weyl, the $\mathbf{C}$-span of these functions is uniformly dense in the space of continuous functions on $K^\natural$. Since a countable union of sets of measure zero is of measure zero, it suffices to show that for each fixed Λ, the set of x in X such that successively averaging over its sequence of angles correctly integrates the function $f_\Lambda(\vartheta)$ has measure one. For Λ the trivial representation, there is nothing to prove: Every single x "works."

Suppose now that Λ is nontrivial (and fixed). We apply the strong law of large numbers to the sequence of independent functions f_i on X given by

$$f_i(x) := f_\Lambda(\vartheta(y_i, s_i, \mathbf{F}_i)),$$

where x is the sequence $(y_i)_i$. Since the function $f_\Lambda(\vartheta)$ on $K^\natural$ has integral zero, Λ being irreducible nontrivial, what we must show is that outside a set of measure zero in X, the sequence

$$(\tfrac{1}{N}) \sum_{1 \le i \le N} f_i(x)$$

tends to zero as the integer N tends to ∞.

The functions f_i are bounded, by $\dim(\Lambda)$, so their variances are bounded, by $(\dim(\Lambda))^2$. Just as in the main text, the Toeplitz lemma and the strong law of large numbers reduce us to showing that the expectations $E(f_i)$ tend to zero.

The expectations $E(f_i)$ over X are tautologically given by

$$E(f_i) := (1/\operatorname{Card}(Y_{s_i}(\mathbf{F}_i))) \sum_{y_i \text{ in } Y_{s_i}(\mathbf{F}_i)} f_\Lambda(\vartheta(y_i, s_i, \mathbf{F}_i)).$$

In terms of the sheaf $\Lambda(\mathscr{F}_{s_i})$ corresponding to the composite representation $\Lambda \circ \rho_{s_i}$, we may rewrite this as

$$E(f_i) := (1/\operatorname{Card}(Y_{s_i}(\mathbf{F}_i))) \sum_{y_i \text{ in } Y_{s_i}(\mathbf{F}_i)} \operatorname{trace}(\operatorname{Frob}_{y_i} | \Lambda(\mathscr{F}_{s_i})).$$

Since $\Lambda(\mathscr{F}_{s_i})$ is ι-pure of weight zero (being a subquotient of some tensor power of $\mathscr{F}_{s_i}$) and geometrically irreducible, the Key Lemma gives

$$|E(f_i)| \leq (1/\operatorname{Card}(Y_{s_i}(\mathbf{F}_i))) \cdot |\chi_c(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i, \Lambda(\mathscr{F}_{s_i}))| \cdot (\sqrt{q_i}).$$

Since $\operatorname{Card}(Y_{s_i}(\mathbf{F}_i)))$ is itself equal to q_i up to an error which is uniformly $O(\sqrt{q_i})$, we have, for some absolute constant A depending only on Y/S, the estimate

$$|E(f_i)| \leq \left(A \cdot |\chi_c\left(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i, \Lambda(\mathscr{F}_{s_i})\right)|\right) / (\sqrt{q_i}).$$

Since the q_i tend to infinity by hypothesis, it suffices to show that the integers $|\chi_c(Y_s \otimes_{\mathbf{F}} \overline{\mathbf{F}}, \Lambda(\mathscr{F}_s))|$ remain bounded, for each fixed Λ, as the data $(\mathbf{F}, s \text{ in } S(\mathbf{F}))$ vary. In fact, we will see that the ratio

$$|\chi_c\left(Y_s \otimes_{\mathbf{F}} \overline{\mathbf{F}}, \Lambda(\mathscr{F}_s)\right)| / \dim(\Lambda)$$

remains bounded as both Λ and the data $(\mathbf{F}, s \text{ in } S(\mathbf{F}))$ vary.

To see this, we argue as follows. By the Euler-Poincare formula, we have

$$\chi_c\left(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i, \Lambda(\mathscr{F}_{s_i})\right) = \chi_c\left(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i\right) \dim(\Lambda) - \sum_{\text{points } z_i \text{ at } \infty} \operatorname{Swan}_{z_i}(\Lambda(\mathscr{F}_{s_i})).$$

As Swan conductors are nonnegative, and are sums of breaks, we get

$$0 \leq \operatorname{Swan}_{z_i}(\Lambda(\mathscr{F}_{s_i})) \leq \dim(\Lambda)(\text{biggest } z_i\text{-break of } \Lambda(\mathscr{F}_{s_i})).$$

As $\Lambda(\mathscr{F}_{s_i})$ is a subquotient of some tensor power of $\mathscr{F}_{s_i}$, its biggest z_i-break is at most the biggest z_i-break of $\mathscr{F}_{s_i}$ itself, which is in turn trivially bounded by $\operatorname{Swan}_{z_i}(\mathscr{F}_{s_i})$, so we obtain the estimate

$$0 \leq \operatorname{Swan}_{z_i}(\Lambda(\mathscr{F}_{s_i})) \leq \dim(\Lambda) \operatorname{Swan}_{z_i}(\mathscr{F}_{s_i}).$$

Thus we obtain

$$\begin{aligned}0 \le \chi_c\left(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i\right)\dim(\Lambda) - \chi_c\left(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i, \Lambda(\mathscr{F}_{s_i})\right)\\ = \sum_{\text{points } z_i \text{ at } \infty} \operatorname{Swan}_{z_i}(\Lambda(\mathscr{F}_{s_i}))\\ \le \dim(\Lambda) \sum_{\text{points } z_i \text{ at } \infty} \operatorname{Swan}_{z_i}(\mathscr{F}_{s_i})\\ = \dim(\Lambda)\left[\chi_c\left(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i\right)\operatorname{rank}(\mathscr{F}_{s_i}) - \chi_c\left(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i, (\mathscr{F}_{s_i})\right)\right]\end{aligned}$$

But $\operatorname{rank}(\mathscr{F}_{s_i})$ is the constant r, $|\chi_c(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i)|$ is constant on connected components of S, say bounded by B, and by hypothesis we have $|\chi_c(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i, \mathscr{F}_{s_i}))| \le M$, so we get the uniform bound

$$|\chi_c\left(Y_{s_i} \otimes_{\mathbf{F}_i} \overline{\mathbf{F}}_i, \Lambda(\mathscr{F}_{s_i})\right)|/\dim(\Lambda) \le B + Br + M. \quad \text{Q.E.D.}$$

Here is a variant of the preceding result.

Fix an integer $r \ge 1$, a prime number ℓ, an algebraic closure $\overline{\mathbf{Q}}_\ell$ of $\mathbf{Q}_\ell$, and an isomorphism of fields $\iota\colon \overline{\mathbf{Q}}_\ell \approx \mathbf{C}$. Let K be a finite extension of $\mathbf{Q}$, $\mathscr{O}_K$ its ring of integers, $N \ge 1$ an integer, and S the spectrum of $\mathscr{O}_K[1/N\ell]$. Let Y/S be a smooth affine curve over S. We suppose that Y/S is of the form $\overline{Y} - D$, for some proper smooth curve $\overline{Y}/S$ with geometrically connected fibers and a divisor D in $\overline{Y}$ which is finite etale over S, of some constant degree $d \ge 1$. For each finite field $\mathbf{F}$, and each $\mathbf{F}$-valued point s in $S(\mathbf{F})$, we denote by $Y_s/\mathbf{F}$ the fiber of Y/S over the $\mathbf{F}$-valued point s, and we fix a geometric point ξ_s of Y_s. Suppose that for each finite field $\mathbf{F}$, and each $\mathbf{F}$-valued point s in $S(\mathbf{F})$, we are given a lisse of rank $r\overline{\mathbf{Q}}_\ell$-sheaf $\mathscr{F}_s$ on Y_s which is ι-pure of weight zero, and which satisfies the hypothesis (Hyp): Under the representation

$$\rho_s\colon \pi_1(Y_s, \xi_s) \to \mathrm{GL}(\mathscr{F}_{\xi_s})$$

which $\mathscr{F}_s$ "is," $\rho_s(\pi_1(Y_s, \xi_s)) \subset G_{\mathrm{geom}, s}$. (Here $G_{\mathrm{geom}, s}$ denotes the Zariski closure in $\mathrm{GL}(\mathscr{F}_{\xi_s})$ of $\rho_s(\pi_1^{\mathrm{geom}}(Y_s, \xi_s))$.)

We suppose that there exists a reductive algebraic subgroup G of $\mathrm{GL}(r, \overline{\mathbf{Q}}_\ell)$, an integer $M \ge 1$, and for each ($\mathbf{F}$, s in $S(\mathbf{F})$) as above, an isomorphism $\mathscr{F}_{\xi_s} \approx (\overline{\mathbf{Q}}_\ell)^r$ such that

(1) $|\chi_c(Y_s \otimes_{\mathbf{F}} \overline{\mathbf{F}}, \mathscr{F}_s)| \le M$.

(2) Via the given isomorphism $\mathscr{F}_{\xi_s} \approx (\overline{\mathbf{Q}}_\ell)^r$, $G_{\mathrm{geom}, s} \subset G$ (inclusion of subgroups of $\mathrm{GL}(r, \overline{\mathbf{Q}}_\ell)$).

(3) For any irreducible nontrivial representation Λ of G, there exists a dense open set U_Λ of S such that if $s \in U_\Lambda$, then the restriction of Λ to $G_{\text{geom},s}$ is an irreducible nontrivial representation of $G_{\text{geom},s}$.

For the sort of S we are considering, (3) may be rewritten:

(3bis) For any irreducible nontrivial representation Λ of G, there exists an integer M_Λ such that the restriction of Λ to $G_{\text{geom},s}$ is an irreducible nontrivial representation of $G_{\text{geom},s}$, provided that the characteristic of $\mathbf{F}$, i.e., the residue characteristic of s, is $\geq M_\Lambda$.

Choose any sequence $(\mathbf{F}_i, s_i \text{ in } S(\mathbf{F}_i))$ in which $p_i := \text{Char}(\mathbf{F}_i)$ tends Archimedeanly to infinity and for which each finite set $Y_{s_i}(\mathbf{F}_i)$ is nonempty, and denote by $X := \Pi_i Y_{s_i}(\mathbf{F}_i)$ the corresponding product space. Endow each factor $Y_{s_i}(\mathbf{F}_i)$ with the probability measure which gives each point the same measure $1/\text{Card}(Y_{s_i}(\mathbf{F}_i))$, and endow X with the product measure μ_{prod}.

Choose a maximal compact subgroup K of the complex reductive group $G(\mathbf{C})$, $\mathbf{C}$ viewed as a $\overline{\mathbf{Q}}_\ell$-algebra via ι, and denote by $K^\natural$ the space of conjugacy classes of K, and by $\mu^\natural$ the projection to $K^\natural$ of Haar measure on K. For each point y_i in each factor $Y_{s_i}(\mathbf{F}_i)$, denote by $\vartheta(y_i, s_i, \mathbf{F}_i)$ in $K^\natural$ the corresponding "angle" of Frobenius. (This "angle" in $K^\natural$ makes sense by the hypothesis (Hyp), and the hypothesis that all the groups $G_{\text{geom},s}$ are subgroups of G.)

Variant Theorem. (joint with Ofer Gabber). *Notations and hypotheses as above, the set of points* $x := (y_i)_i$ *in* $X := \Pi_i Y_{s_i}(\mathbf{F}_i)$ *for which the corresponding sequence of "angles"* $(\vartheta(y_i, s_i, \mathscr{F}_i))_i$ *in* $K^\natural$ *is equidistributed with respect to* $\mu^\natural$ *is a set of measure one in* X *with respect to* μ_{prod}.

Proof. The proof is essentially unchanged. One need only remark that for fixed Λ, the bounds for $E(f_i)$ established above remain valid for all but the finitely many values of i for which $p_i < M_\Lambda$, and consequently the $E(f_i)$ tend to zero. Q.E.D.

REFERENCES

[AdSp] A. Adolphson and S. Sperber, *Exponential sums and Newton polyhedra; cohomology and estimates*, Ann. of Math. **130** (1989), 367–406.

[Ar] E. Artin, *Quadratische Korper in Gebiete der hoheren Kongruenzen* I *and* II, Math. Z. **19** (1924), 153–246.

[Ax] J. Ax, *Zeros of polynomials over finite fields*, Amer. J. Math. **86** (1964), 255–261.

[Bei] A. A. Beilinson, On the derived category of the category of the category of perverse sheaves, in Yu. I. Manin (ed.), *K-Theorey, arithmetic and geometry*, Lecture Notes in Math., vol. 1289, Springer-Verlag, Berlin, 1987, pp. 27–41.

[BBD] A. A. Beilinson, I. N. Bernstein and P. Deligne, Faisceaux pervers, entire contents of *Analyse et Topologie sur les espaces singuliers I, Conference de Luminy*, Astérisque **100** (1982).

[Bel] G. V. Belyi, *On Galois extensions of a maximal cyclotomic field*, Math. USSR-Izv. **14**(2) (1980), 247–256.

[Ber] J. Bernstein, *Six lectures on the algebraic theory of $\mathscr{D}$-modules*, Xeroxed notes, E.T.H., Zurich, 1983.

[Bert] P. Berthelot, *Cohomologie cristalline des schémas de caractéristique* $p > 0$, Lecture Notes in Math., vol. 407, Springer-Verlag, Berlin, 1974.

[BH] F. Beukers and G. Heckman, *Monodromy for the hypergeometric function* ${}_nF_{n-1}$, Invent. Math. **95** (1989), 325–354.

[BBH] F. Beukers, D. Brownawell and G. Heckman, *Siegel normality*, Ann. of Math. **127** (1988), 279–308.

[Bi] B. Birch, *How the number of points of an elliptic curve over a fixed prime field varies*, J. London Math. Soc. **43** (1968), 57–60.

[Bor] A. Borel, et al. *Algebraic $\mathscr{D}$-Modules*, Academic Press, Boston, 1987.

[Bour1] N. Bourbaki, *Groupes et algebres de Lie*, Chapter 1, Diffusion CCLS, Paris, 1971.

[Bour2] ——, *Groupes et algebres de Lie*, Chapters 4, 5 and 6, Masson, Paris, 1981.

[Bour3] ——, *Groupes et algebres de Lie*, Chapters 7 and 8, Diffusion CCLS, Paris, 1975.

[Bour4] ——, *Groupes et algebres de Lie*, Chapter 9, Masson, Paris, 1982.

[Br] J.-L. Brylinski, *Transformations canoniques, dualité projective, théorie de Lefschetz, transformations de Fourier et sommes trigonométriques*, Astérisque **140–141** (1986), 3–134.

[CR1] C. W. Curtis and I. Reiner, *Representation theory of finite groups and associative algebras*, Interscience Publ., New York and London, 1962.

[CR2] ——, *Methods of representation theory with applications to finite groups and orders*, vol. I, John Wiley and Sons, New York, 1981.

[De1] P. Deligne, Rapport sur la formule des traces, in *Cohomologie Etale* (*SGA* $4\frac{1}{2}$), Lecture Notes in Math., vol. 569, Springer-Verlag, Berlin, 1977, pp. 76–109.

[De2] ——, Application de la formule des traces aux sommes trigonométriques, in *Cohomologie Etale* (*SGA* $4\frac{1}{2}$), Lecture Notes in Math., vol. 569, Springer-Verlag, Berlin, 1977, pp. 168–232.

[De3] ——, *Catégories tannakiennes*, Grothendieck Festschrift (to appear).

[De4] ____, *Equations differentielles à points singuliers reguliers*, Lecture Notes in Math., vol. 163, Springer-Verlag, Berlin, 1970.

[De5] ____, Theoremes de finitude en cohomologique ℓ-adique, in *Cohomologie Etale* (*SGA* $4\frac{1}{2}$), Lecture Notes in Math., vol. 569, Springer-Verlag, Berlin, 1977, pp. 233–251.

[De6] ____, *La conjecture de Weil* I, Publ. Math. I.H.E.S. **48** (1974), 273–308.

[De7] ____, *La conjecture de Weil* II, Publ. Math. I.H.E.S. **52** (1981), 313–428.

[DM] P. Deligne and J. Milne, Tannakian categories, in P. Deligne, J. Milne, A. Ogus and K.-Y. Shih(eds.), *Hodge cycles, motives, and Shimura varieties*, Lecture Notes in Math., vol. 900, Springer-Verlag, Berlin, 1982, pp. 101–228.

[Deu] Deuring, *Die Zetafunktion einer algebraischen Kurve von Geschlechte Eins, Drei Mitteilungen*, Nachr. Akad. Wiss. Göttingen, Math.-Phys. K1. II, 1953, pp. 85–94; 1955, pp. 13–43; 1956, pp. 37–76.

[Dw] B. Dwork, *Bessel functions as p-adic functions of the argument*, Duke Math. J. **41** (1974), 711–738.

[EGA] A. Grothendieck, Rédigé avec la collaboration de Dieudonné, J., *Elements de géométrie algébrique*, Publ. Math. I.H.E.S. 4, 8, 11, 17, 20, 24, 28, 32, (1960-1967).

[Ek] T. Ekedahl, *On the adic formalism*, Grothendieck Festschrift (to appear).

[Er] A. Erdelyi, *Higher transcendental functions*, vol. I (Bateman Manuscript Project), McGraw-Hill, New York, 1953.

[Ev] L. Evens, *A generalization of the transfer map in the cohomology of groups*, Trans. Amer. Math. Soc. **108** (1963), 54–65.

[Gel] I. M. Gel′fand, *The general theory of hypergeometric functions*, Dokl. Akad. Nauk SSSR **288** (1) (1986), 14–18.

[Gre] J. Greene, *Hypergeometric functions over finite fields*, Transl. Amer. Math. Soc. **301** (1987), 77–101.

[Gro1] A. Grothendieck, Crsytals and the De Rham cohomology of schemes, reprinted in *Dix exposés sur la cohomologie des schémas*, North-Holland, Amsterdam, 1968.

[Gro2] ____, Formule de Lefschetz et rationalité des fonctions L, Seminaire Bourbaki 1964-65, Exposé 279, reprinted in *Dix exposés sur la cohomologie des schémas*, North-Holland, Amsterdam, 1968.

[Gr3] A. Grothendieck, et al. *Séminaire de géométre algébrique du Bois-Marie*, SGA 1; SGA 4 Parts I, II, and III; SGA 41/2; SGA 5; SGA 7 Parts I and II, Lecture Notes in Math., Vols. 224, 269–270, 305, 569, 589, 288–340, Springer-Verlag, Berlin, 1971–1977.

[Ha] H. Hasse, *Beweis des Analogons der Riemannschen Vermutung fur die Artinsche und F. K. Schmidschen Kongruenz-zetafunktionen in gewissen elliptischen Fallen*, Ges. d. Wiss. Nach. Math-Phys. Klasse **3** (1933), 253–262.

[HBP] D. J. Heath-Brown and S. J. Patterson, *The distribution of Kummer sums at prime arguments*, J. Reine Angew. Math. **310** (1979), 111–130.

[Ill1] L. Illusie, *Appendix to Deligne, P., Theoremes de finitude en cohomologique ℓ-adique*, Cohomologie Etale (*SGA* $4\frac{1}{2}$), Lecture Notes in Math., vol. 569, Springer-Verlag, Berlin, 1977, 233–251.

[Ill2] ____, *Deligne's ℓ-adic Fourier transform*, Algebraic Geometry. Bowdoin 1985 (S.J. Bloch, ed.), Amer. Math. Soc., Providence, 1987.

[K] M. Kashiwara, *The Riemann-Hilbert problem for holonomic systems*, RIMS, Kyoto Univ., 1983.

[Ka1] N. Katz, *Algebraic solutions of differential equations; p-curvature and the Hodge filtration*, Invent. Math. **18** (1972), 1–118.

[Ka2] ——, *On the calculation of some differential Galois groups*, Invent. Math. **87** (1987), 13–61.

[Ka3] ——, *Gauss sums, Kloosterman sums, and monodromy groups*, Ann. of Math. Study, vol. 116, Princeton Univ. Press, Princeton, N.J., 1988.

[Ka4] ——, *Local to global extensions of representations of fundamental groups*, Ann. Inst. Fourier (Grenoble) **36** (4) (1986), 59–106.

[Ka5] ——, *On the monodromy groups attached to certain families of exponential sums*, Duke Math. J. **54** (1) (1987), 41–56.

[Ka6] ——, Perversity and exponential sums, in *Algebraic number theory—in honor of K. Iwasawa*, Adv. Studies Pure Math., vol. 17, North-Holland, Amsterdam, 1989, pp. 209–259.

[Ka7] ——, *Exponential sums and differential equations*, Ann. of Math. Study 124, Princeton University Press, Princeton, N.J., 1990.

[Ka8] ——, *A conjecture in the arithmetic theory of differential equations*, Bull. Soc. Math. France **110** (1982), 203–239.

[Ka9] ——, *Sommes exponentielles*, rédigé par G. Laumon, Astérisque **79** (1980).

[Ka10] ——, *Travaux de Laumon*, Séminaire Bourbaki 1987-88, Exposé 691, Astérisque (to appear).

[KL] N. Katz and G. Laumon, *Transformation de Fourier et majoration de sommes exponentielles*, Publ. Math. I.H.E.S. **62** (1986), 361–418.

[KaPi] N. Katz and R. Pink, *A note on pseudo-CM representations and differential Galois groups*, Duke Math. J. **54** (1) (1987), 57–65.

[Kob] N. Koblitz, *The number of points on certain families of hypersurfaces over finite fields*, Compositio Math. **48** (1983), 3–23.

[Kol] E. Kolchin, *Algebraic groups and algebraic independence*, Amer. J. Math. **90** (1968), 1151–1164.

[Kos] B. Kostant, *A characterization of the classical groups*, Duke Math. J. **25** (1958), 107–123.

[La] S. Lang, *Algebraic Number Theory*, Addison-Wesley, Reading, Mass., 1970.

[Lau1] G. Laumon, *Transformation de Fourier, constantes d'équations fonctionelles et conjecture de Weil*, Publ. Math. I.H.E.S. **65** (1987), 131–210.

[Lau2] ——, Semi-continuité du conducteur de Swan (d'après P. Deligne), in *Caractéristique d'Euler-Poincaré, Seminaire E.N.S. 1978-79*, Astérisque **82–83** (1981), 173–219.

[Lev1] A. H. M. Levelt, *Jordan decomposition for a class of singular differential operators*, Ark. Mat. **13** (1975), 1–27.

[Le2] ——, Hypergeometric functions, thesis, University of Amsterdam, 1961.

[Ma] J. Manin, *Moduli fuchsiani*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **19** (1965), 13–126.

[Me1] Z. Mebhkout, *Sur le problème de Hilbert-Riemann*, Lecture Notes in Phys., vol. 129, Springer-Verlag, Berlin, 1980, pp. 99–110.

[Me2] ——, *Une équivalence de catégories et une autre équivalence de catégories*, Compositio Math. **51** (1984), 55–69.

[Me3] ——, *Le formalisme des six operations de Grothendieck pour les coefficients de de Rham*, Seminaire de Plans-sur-Bex, mars 1984, Travaux en Cours, Hermann, 1986.

[Pa] S. J. Patterson, *On the distribution of Kummer sums*, J. Reine Angew. Math. **303** (1978), 126–143.

[Ri] K. Ribet, *Galois action on division points of abelian varieties with real multiplication*, Amer. J. Math. **98** (1976), 751–805.

[Saa] N. Saaveedra Rivano, *Catégories tannakiennes*, Lecture Notes in Math., vol. 265, Springer-Verlag, Berlin, 1972.

[Se] J.-P. Serre, *Abelian ℓ-adic representations and elliptic curves*, Addison-Wesley, Reading, Mass., 1989.

[SeTa] J.-P. Serre and J. Tate, *Good reduction of abelian varieties*, Ann. of Math. (2) **88** (1968), 492–517.

[Sch] F. K. Schmidt, *Analytische Zahlentheorie in Korporn der Characteristik* p. Math. Z. **33** (1931), 1–32.

[Ver] J.-L. Verdier, Specialization de faisceaux et monodromie moderé, in *Analyse et topologie sur les espaces singuliers*, Vols. II and III, Astérisque **101–102** (1983), 332–364.

[We] A. Weil, *Jacobi sums as Grossencharaktere*, Trans. Amer. Math. Soc. **73** (1952), 487–492.

[Yo] H. Yoshida, *On an analogue of the Sato conjecture*, Invent. Math. **19** (1973), 261–277.

[Za1] Yu. G. Zarkhin, *Weights of simple Lie algebras in the cohomology of algebraic varieties*, Izv. Akad. Nauk. SSSR, Ser. Mat. **48** (1984), 264–304; English translation in Math. USSR-Izv. **24** (1985), 245–281.

[Za2] ——, Linear simple Lie algebras and ranks of operators, Grothendieck Festschrift (to appear).

DEPARTMENT OF MATHEMATICS, PRINCETON UNIVERSITY, PRINCETON, NEW JERSEY 08544