Milan J. Math. Vol. 78 (2010) 179–220
DOI 10.1007/s00032-010-0111-x
Published online March 11, 2010

Milan Journal of Mathematics

Some Aspects and Applications of the Riemann Hypothesis over Finite Fields

E. Kowalski

Abstract. We give a survey of some aspects of the Riemann Hypothesis over finite fields, as it was proved by Deligne, and its applications to analytic number theory. In particular, we concentrate on the formalism leading to Deligne's Equidistribution Theorem.

Mathematics Subject Classification (2010). 11T23, 11L05, 14G15, 14F20.

Keywords. Exponential sums over finite fields, Deligne equidistribution theorem.

1. Introduction

The goal of this survey is to present some aspects of the Riemann Hypothesis over finite fields. The context is Deligne's celebrated work ([5], [6]) and its applications, and the text is roughly split in two parts. In the first part, we try to introduce and motivate the framework in which the powerful formalism of étale cohomology and the Riemann Hypothesis operate, emphasizing aspects leading to Deligne's remarkable Equidistribution Theorem. The second part (starting in Section 5) is a discussion of this theorem, which involves naturally "families" of exponential sums and L-functions over finite fields, and of some (mostly) recent applications of the Riemann Hypothesis and Deligne's theorem, concluding with a short list of open problems (emphasizing general, "philosophical" issues, rather than specific questions).

This is written with a target audience of readers who are not experts in algebraic geometry, in particular analytic number theorists. We use a few basic examples as references, notably Gauss sums, Kloosterman sums (and their average Sato-Tate distribution) and the very simple – but enlightening – case of finite (zero-dimensional) algebraic varieties. The emphasis is throughout in situations which, at least at first sight, are not immediately or obviously analogue to the classical Riemann Hypothesis for the Riemann zeta and Dirichlet L-functions.

We try to be as accessible as possible to a large audience; however, due to the author's bias, much of the examples, applications and problems are directly or indirectly related to analytic number theory. In some sense, this survey is a follow-up to Chapter 11, Section 11 of [11], where the cohomological approach to exponential

sums over finite fields was also surveyed, but with little attention paid to situations involving families and Deligne's equidistribution theorem.

Acknowledgments. I wish to thank warmly the organizers of the RISM conference for inviting me to participate and for the excellent organization. The written version was prepared at the Institute for Advanced Study (Princeton, NJ) during the fall semester 2009, while the author was on sabbatical leave. Thanks to this institution for its support.

Thanks also to M. Ibraimi, F. Jouve and D. Zywina for comments and corrections.

Notation. As usual, $|X|$ denotes the cardinality of a set. By $f \ll g$ for $x \in X$, or $f = O(g)$ for $x \in X$, where X is an arbitrary set on which f is defined, we mean synonymously that there exists a constant $C \geqslant 0$ such that $|f(x)| \leqslant Cg(x)$ for all $x \in X$. The "implied constant" refers to any value of C for which this holds. It may depend on the set X, which is usually specified explicitly, or clearly determined by the context.

We write $e(z) = e^{2i\pi z}$ for $z \in \mathbf{C}$. For any $q \neq 1$ which is a power of a prime, we write $\mathbf{F}_q$ for a finite field with q elements, in particular $\mathbf{F}_p = \mathbf{Z}/p\mathbf{Z}$. The letter p will always be used to refer to prime numbers.

2. Setting the stage

The Riemann Hypothesis was initially stated as a problem concerning the location of the zeros of a certain meromorphic function, and was generalized to Dirichlet L-functions in the same terms. It is possible to present the Riemann Hypothesis over finite fields in very close analogy (and we will recall this below, see Example 14).

However, applications often appeal to alternate statements, which may look quite different. For instance, two early occurrences of the Riemann Hypothesis over finite fields, historically, are the following results of Gauss: (1) for any odd prime number p, and any integer a coprime with p, we have

$$\Big|\sum_{x=1}^{p} e\Big(\frac{ax^2}{p}\Big)\Big| = \sqrt{p} \tag{1}$$

(recall we put $e(z) = e^{2i\pi z}$); (2) for any odd prime p with $p \equiv 1 \pmod 4$, the number of solutions in $\mathbf{Z}/p\mathbf{Z} \times \mathbf{Z}/p\mathbf{Z}$ of the equation

$$y^2 = x^4 - 1$$

is $p - 2a$, where a is the unique odd integer such that p may be written $p = a^2 + b^2$ with the sign of a fixed by the complex congruence $a + ib \equiv 1 \pmod{2(1+i)}$ (see the introduction to Weil's article [24] and his comments [25] for more historical perspective).

In these results (the first of which is elementary, while the second remains somewhat challenging), the clue to the Riemann Hypothesis is the exponent $1/2$

hidden in $\sqrt{p}$, in plain sight for (1), and disguised in the bound

$$|2a| \leqslant 2\sqrt{p}$$

which immediately follows from the recipe for a in the second example.

The best reference for purposes of comparison with the classical Riemann Hypothesis is then seen to be the statement

$$\Big|\sum_{p\leqslant x}\chi(p)\Big| \leqslant 2x^{1/2}(\log qx)^2, \qquad \text{for } x \geqslant 2,$$

for all primitive Dirichlet characters modulo $q \geqslant 1$, a concrete estimate which is known to be equivalent with the Generalized Riemann Hypothesis for the L-functions of such characters.

The essence of this statement, for our purposes, is that it shows that the values $\chi(p)$, when p ranges over primes (in increasing order) vary extremely randomly – recall that for randomly chosen, independent, arguments $\theta_p \in [0,1]$, $p \leqslant x$, the mean square of the sums

$$\sum_{p\leqslant x} e(\theta_p)$$

is precisely $\pi(x)$, by a simple application of the orthogonality of additive characters.

There are two important aspects that we want to emphasize for this survey: (1) the result – as far as the exponent of x – is best possible, because "there are zeros on the critical line"; (2) it is completely uniform with respect to the modulus. Both of these facts are important in applications of the Riemann Hypothesis, to the distribution of primes for instance, and in both respects, the current unconditional knowledge is quite poor. And both of these are also already visible in the simplest example (1) we gave of the Riemann Hypothesis over finite fields.

In the remainder of this survey, we will look at the Riemann Hypothesis from this point of view, and will explain how it provides not only excellent, often very explicit, estimates for certain sums

$$\sum_{x\in V(\mathbf{F}_q)} \Lambda(x)$$

over points of algebraic varieties over finite fields, where Λ is typically an oscillating factor of "algebraic" origin,[1] but also does so through a general framework, and a very powerful formalism.

Before going to the general case, here are a few additional examples that will reappear many times below as illustrations of the general theory.

Example 1 (Hasse bound). This generalizes the result of Gauss concerning the curve $y^2 = x^4 - 1$ (though a change of variable is required for this to be obvious): for any prime $p \neq 2, 3$, any integers a, b with $4a^3 + 27b^2$ not divisible by p, we have

$$\Big||\{(x,y) \in (\mathbf{Z}/p\mathbf{Z})^2 \mid y^2 = x^3 + ax + b\}| - p\Big| \leqslant 2\sqrt{p}$$

[1] We introduce these more precisely in the next section.

note again the uniformity in this estimate, where a and b do not occur on the right-hand side.

Example 2 (Hyper-Kloosterman sums). For any prime p, any $n \geqslant 1$, any a not divisible by p, let

$$HK(n;a,p) = \sum_{\substack{1\leqslant x_1,\dots,x_n\leqslant p-1\\ x_1x_2\cdots x_n=a \,(\mathrm{mod}\, p)}} e\Big(\frac{x_1+\cdots+x_n}{p}\Big) \tag{2}$$

(this is called a Hyper-Kloosterman sum, and is a sum over $n-1$ variables, since x_n can be recovered uniquely from $x_1, \dots, x_{n-1}$). Then we have the estimate

$$|HK(n;a,p)| \leqslant np^{(n-1)/2} \tag{3}$$

This bound was proved by Weil for $n = 2$, which is the case of the classical Kloosterman sums which are of crucial importance in analytic number theory (because they occur in Kloosterman's refinement of the circle method, and in the Fourier expansion of Poincaré series; see, e.g., [11, §20.3, §16]); the general case $n \geqslant 3$ was proved by Deligne [5]. Note again the sharpness of the statement: the square-root cancellation showing in the exponent $(n-1)/2$, and the explicit constant n.

Before presenting the next example, we recall an important definition:

Definition. Let X be a locally compact metric space, and μ a probability measure[2] on X, so that $\mu(X) = 1$. If (X_n) is a sequence of finite sets such that $X_n \subset X$, or more generally such that X_n is given with a map $X_n \xrightarrow{\theta_n} X$, not necessarily injective, then the sets (X_n) *become equidistributed with respect to* μ if, for all continuous and bounded functions $f : X \to \mathbf{C}$, we have

$$\frac{1}{|X_n|}\sum_{x\in X_n} f(\theta_n(x)) \longrightarrow \int_X f(x)d\mu(x).$$

The concept of equidistribution turns out to be extremely useful and ubiquitous in number theory. The link with oscillating sums is given by the well-known *Weyl criterion*:

Proposition 3. *Let X be a compact space, and let μ be a probability measure on X. Let $(\varphi_j)_j$ be continuous functions on X which form an orthonormal basis of the orthogonal complement of the constant function 1 in $L^2(X,\mu)$, so that in particular*

$$\int_X \varphi_j(x)d\mu(x) = 0, \qquad \text{for all } j.$$

Then a sequence of finite sets X_n, either $X_n \subset X$ or with $X_n \xrightarrow{\theta_n} X$, becomes equidistributed with respect to μ if and only if

$$\frac{1}{|X_n|}\sum_{x\in X_n} \varphi_j(\theta_n(x)) \longrightarrow 0$$

for all j.

[2] For the standard σ-algebra of Borel sets in X; we always assume μ is finite on compact sets.

We illustrate with two further examples:

Example 4 ("Average" Sato-Tate law). Example 2, with $n = 2$, and the fact that the Hyper-Kloosterman sums with n even are all real numbers, shows that for p prime and $a \in (\mathbf{Z}/p\mathbf{Z})^\times$, there is a unique angle $\theta_p(a) \in [0, \pi]$ such that

$$\sum_{x=1}^{p-1} e\Big(\frac{x + ax^{-1}}{p}\Big) = 2\sqrt{p}\ \cos\theta_p(a).$$

We are interested in the distribution of the angles $(\theta_p(a))_{a\in\mathbf{F}^\times}$, as p grows (in other words, in the distribution of values of Kloosterman sums, in the scale $\sqrt{p}$ corresponding roughly to their maximal size); there might be coincidences among those angles, so we let $X_p = \mathbf{F}_p^\times$ and use

$$\theta_p : \begin{cases} X_p \to [0, \pi] \\ a \mapsto \theta_p(a) \end{cases}$$

to be in the situation of the definition. N. Katz [14] showed that this sequence of sets of angles becomes equidistributed as $p \to +\infty$ with respect to the so-called Sato-Tate measure

$$\mu = \frac{2}{\pi}\sin^2\theta d\theta \qquad \text{on } [0, \pi].$$

Example 5 (Angles of Gauss sums). Here is a classical example, where already a number of the previous examples interact: consider an odd prime p, a non-trivial Dirichlet character χ modulo p and an element $a \in (\mathbf{Z}/p\mathbf{Z})^\times$; the corresponding Gauss sums are defined by

$$\tau_a(\chi) = \sum_{x\in\mathbf{F}_p^\times} \chi(x)e\Big(\frac{ax}{p}\Big) \tag{4}$$

(the link with (1) is that, if χ is the real non-trivial character modulo an odd prime p, it is easy to show that

$$\tau_1(\chi) = \sum_{x\in\mathbf{F}_p} e\Big(\frac{x^2}{p}\Big),$$

using the fact that the number of $x \in \mathbf{F}_p$ with $x^2 = y$ is given by $1+\chi(y)$ for $y \in \mathbf{F}_p$). Similarly to (1), one shows that $|\tau_a(\chi)| = \sqrt{p}$, so (taking $a = 1$ for simplicity), there is a unique angle $\theta_p(\chi) \in [0, 1[$ such that

$$\tau_1(\chi) = \sqrt{p}\ e(\theta_p(\chi)).$$

It turns out that, as $p \to +\infty$, the finite sets[3]

$$\{\theta_p(\chi) \mid \chi \text{ non-trivial } (\mathrm{mod}\, p)\} \subset [0, 1]$$

[3] Again, we are really taking $X_p = \{\chi \neq 1\}$ and use θ_p to map to $[0, 1]$. However, we will allow this abuse of notation here and elsewhere.

become equidistributed to the Lebesgue measure dx. At a very high level, the proof is as follows: one applies the Weyl criterion with the functions given by the additive characters

$$\varphi_j(x) = e(jx), \qquad j \in \mathbf{Z} - \{0\},$$

and then

$$\frac{1}{p-2} \sum_{\substack{\chi \,(\mathrm{mod}\, p)\\ \chi \neq 1}} \varphi_j(\theta_p(\chi)) = \frac{1}{p-2} \sum_{\substack{\chi \,(\mathrm{mod}\, p)\\ \chi \neq 1}} \Big(\frac{\tau_1(\chi)}{\sqrt{p}}\Big)^j$$

and if $j \geqslant 1$ (the case $j \leqslant -1$ being dealt using symmetry), we can expand the definition of $\tau_1(\chi)^j$ and use orthogonality of characters to obtain

$$\begin{aligned}\sum_{\substack{\chi \,(\mathrm{mod}\, p)\\ \chi \neq 1}} \tau_1(\chi)^j &= \sum_{x_1,\dots,x_j} e\Big(\frac{x_1 + \cdots + x_j}{p}\Big) \sum_{\chi \neq 1} \chi(x_1 \dots x_j)\\ &= (p-1) \sum_{\substack{x_1,\dots,x_j \in \mathbf{F}_p\\ x_1 x_2 \cdots x_j = 1}} e\Big(\frac{x_1 + \cdots + x_j}{p}\Big)\end{aligned}$$

in which we recognize a Hyper-Kloosterman sum in $j-1$ variables. Applying Deligne's estimate (3), we get

$$\Big|\frac{1}{p-2} \sum_{\substack{\chi \,(\mathrm{mod}\, p)\\ \chi \neq 1}} \varphi_j(\theta_p(\chi))\Big| \leqslant j \frac{p-1}{p-2} \frac{1}{p^{1/2}} \longrightarrow 0$$

as $p \to +\infty$, verifying the Weyl criterion. But note that although a weaker bound than (3) would suffice, it would still need to be extremely strong: the exponent $(j-1)/2$ can not be replaced by $j/2$, although the latter would already be quite good for large j (on the other hand, the leading multiplicative constant j does not play a big role here, since we apply the bounds for fixed j).

3. Algebraic exponential sums

We will now describe the statement of the Riemann Hypothesis for exponential sums

$$S(V, \Lambda; q) = \sum_{x \in V(\mathbf{F}_q)} \Lambda(x)$$

of a quite general type. But general does not mean arbitrary: the summation sets and summands must have a specific *algebraic structure* for the theory and formalism to be available,[4] and we will use the (non-standard) shorthand "algebraic exponential sums" to indicate this.

First, the summation sets are of the type $V(\mathbf{F}_q)$ where V is an algebraic variety defined over a finite field (either $\mathbf{F}_q$ or a subfield); the notation indicates, concretely, the set of points on V which have coordinates in $\mathbf{F}_q$. Quite often, very simple varieties

[4] We do not imply, of course, that other types of sums are not interesting; in fact, many exponential sums *not* of algebraic type occur in number theory; see, e.g., [11, §8, §13].

will do: the affine space $\mathbf{A}^d$ of dimension d, such that $\mathbf{A}^d(\mathbf{F}_q) = \mathbf{F}_q^d$ (i.e., summing over $\mathbf{A}^d(\mathbf{F}_q)$ is a "free summation" over d variables each in $\mathbf{F}_q$); the multiplicative group $\mathbf{G}_m$ with $\mathbf{G}_m(\mathbf{F}_q) = \mathbf{F}_q^\times$, and its powers $\mathbf{G}_m^d$ with $\mathbf{G}_m^d(\mathbf{F}_q) = (\mathbf{F}_q^\times)^d$. More generally but still very concretely, V can be any affine variety determined by the vanishing of finitely many polynomials in finitely many variables: in such a case, there exist (fixed, but not unique) integers m, $n \geqslant 0$ and polynomials

$$F_1(X_1, \dots, X_n), \dots, F_m(X_1, \dots, X_n) \in \mathbf{F}_q[X_1, \dots, X_n]$$

such that, for any extension field K of $\mathbf{F}_q$, we have

$$V(K) = \{(x_1, \dots, x_n) \in K^n \mid F_1(x_1, \dots, x_n) = \cdots = F_m(x_1, \dots, x_n) = 0\}.$$

Often (as is the case with $\mathbf{A}^d$ or $\mathbf{G}_m^d$ for any $d \geqslant 1$), such a description exists[5] with polynomials in $\mathbf{Z}[X_1, \dots, X_n]$, in which case we have a variety defined over $\mathbf{Z}$, and (by reducing modulo primes), we can consider its reduction modulo p for any primes; then $V(\mathbf{F}_p)$ makes sense for any p, and $V(\mathbf{F}_q)$ for any prime power $q \neq 1$.

Example 6. (1) The Hyper-Kloosterman sums (3) were defined by a summation over $(x_1, \dots, x_n)$ in $\mathbf{F}_p^n$ subject to the equation $x_1 \cdots x_n = a$ where $a \in \mathbf{F}_p^\times$ is fixed; this set is of the form $V_{n,a}(\mathbf{F}_p)$ for the variety $V_{n,a}/\mathbf{Z}$ defined by the corresponding polynomial $X_1 \cdots X_n - \tilde{a}$ (choosing $\tilde{a} \in \mathbf{Z}$ reducing to a). Note that if one writes

$$HK(n; a, p) = \sum_{(x_1, \dots, x_{n-1}) \in \mathbf{G}_m^{n-1}(\mathbf{F}_p)} e\Big(\frac{x_1 + \cdots + x_{n-1} + a/(x_1 \cdots x_{n-1})}{p}\Big),$$

then we are simply exhibiting at this concrete level an isomorphism

$$\left\{ \begin{array}{ccc} \mathbf{G}_m^{n-1} & \longrightarrow & V_{n,a} \\ (x_i) & \mapsto & (x_1, \dots, x_{n-1}, a \prod x_i^{-1}). \end{array} \right.$$

(2) Now suppose we tried to look instead at a sum like the following:

$$\sum_{\substack{x \in V_{n,a}(\mathbf{F}_p) \\ \sum x_i \text{ is a square}}} e\Big(\frac{x_1 + \cdots + x_n}{p}\Big),$$

where being a square refers to being a square in $\mathbf{F}_p$. Despite the algebraic appearance of the summation set, it is not of the type allowed, because the condition that $y \in \mathbf{F}_q$ is a square does not define an algebraic variety (it is not a stable condition under field extensions; any element $y \in \mathbf{F}_p$ is a square in the quadratic extension $\mathbf{F}_{p^2}$).

(3) There is one class of particularly simple examples which can sometimes be used to gain a minimal understanding of the algebraic issues involved: 0-dimensional varieties. This corresponds to equations where the total number of solutions in an algebraic closure of the ground field is finite. If we consider the one-variable case, this means that V is defined by the vanishing of a single polynomial $f \in \mathbf{F}_q[X]$, $f \neq 0$. Thus, $V(\bar{\mathbf{F}}_q)$ contains at most $\deg(f)$ points (because we allow multiple roots), and

[5] For $\mathbf{G}_m$ one must use the trick of introducing an extra variable and seeing $\mathbf{G}_m$ as the variety defined by the polynomial $X_1 X_2 - 1$.

$V(\mathbf{F}_{q^\nu})$ contains exactly those roots of f which generate a subfield of $\mathbf{F}_{q^\nu}$. In case f is irreducible over $\mathbf{F}_q$ and of degree $\geqslant 2$, in particular, we have $V(\mathbf{F}_q) = \emptyset$.

Thus the summation sets are quite simple. The definition of the right type of summands $\Lambda(x)$ is more delicate. When we look at a sum over $V(\mathbf{F}_p)$, the first examples that come to mind are those of the type

$$\Lambda_{f,g}(x) = \chi(g(x))e\Big(\frac{f(x)}{p}\Big) \tag{5}$$

where f, g are polynomial functions on V, with g not taking the value 0 and χ is a multiplicative character modulo p. These are suggested by the examples of Gauss sums, with

$$\Lambda_{aX,X}(x) = \chi(x)e\Big(\frac{ax}{p}\Big),$$

and Hyper-Kloosterman sums (seen as sums over $V = \mathbf{G}_m^{n-1}$, see (6, (1))), with

$$\Lambda_{X_1+\cdots+X_{n-1}+a/(X_1\cdots X_{n-1}),1}(x) = e\Big(\frac{x_1+\cdots+x_{n-1}+a/(x_1\cdots x_{n-1})}{p}\Big), \tag{6}$$

and indeed by many applications in analytic number theory (for instance, the circle method). Sums of this type are often called "character sums" or "mixed character sums"; if $g = 1$, one speaks of "additive character sums", and if $f = 0$, of "multiplicative character sums". Deligne's survey [4] of the cohomological techniques to study these sums is highly rewarding (but requires more knowledge of algebraic geometry).

We need to enlarge this class of summands in two ways, one of which is very easy to describe, but the other not so much.

We first define analogues of the sums with the summands (5) for $V(\mathbf{F}_q)$ where q is not necessarily prime. This is done by taking χ to be any multiplicative character of $\mathbf{F}_q^\times$, and replacing $e(f(x)/p)$ by $\psi(f(x))$, where

$$\psi : \mathbf{F}_q \to \mathbf{C}^\times$$

is any additive character of $\mathbf{F}_q$. It is very important for the theory that this construction also allows the formation of the "companion" sums

$$S_\nu(V, \Lambda_{f,g}; q) = \sum_{x\in V(\mathbf{F}_{q^\nu})} \chi(N_{\mathbf{F}_{q^\nu}/\mathbf{F}_q} g(x))\psi(\mathrm{Tr}_{\mathbf{F}_{q^\nu}/\mathbf{F}_q} f(x))$$

over $\mathbf{F}_{q^\nu}$, $\nu \geqslant 1$, which are an important part of the theory. Indeed, it will often be much easier to understand the behavior of the sums $S_\nu(V, \Lambda_{f,g}; q)$ in the limit where $\nu \to +\infty$, and this may give insight into more difficult situations (e.g., when $q = p$ is prime). It is customary to refer to this limit (fields of increasing size but fixed characteristic) as the "vertical direction" (or limit), and to refer to the case of increasing p as a "horizontal" direction.

Example 7. If we come back to the setting of the Hasse bound, we can see a multiplicative character sum in the background: indeed, let $f(x) = x^3 + ax + b$, with notation as in Example 1. The question is to count $|E(\mathbf{F}_p)|$, where E is the algebraic variety given by the equation $Y^2 - f(X)$. Recall that if p is an odd prime and χ

is the non-trivial real quadratic character of $\mathbf{F}_p^\times$, extended to $\mathbf{F}_p$ by $\chi(0) = 0$, the number of solutions of the equation

$$y^2 = f(x)$$

is equal to $1 + \chi(f(x))$; it follows that

$$|E(\mathbf{F}_p)| = \sum_{y\in\mathbf{F}_p} (1 + \chi(f(x))) = p + S(\mathbf{A}^1, \Lambda_\chi; p)$$

where $\Lambda_\chi(x) = \chi(f(x))$. Hence Hasse's result is equivalent with the upper bound

$$|S(\mathbf{A}^1, \Lambda_\chi; p)| \leqslant 2\sqrt{p}$$

for this multiplicative character sum. This is the situation of elliptic curves,[6] and if we allow f to be replaced by an arbitrary polynomial $f \in \mathbf{F}_q[X]$ with no repeated root, we will have a similar link between counting points on the so-called (affine) hyperelliptic curve with equation $Y^2 - f(X)$, and the corresponding multiplicative sum.

A good theory for (estimates of) character sums is already immensely useful in number theory. To cite just a few classical examples (before the 1960's), they were used extensively in the circle method and in estimates for Fourier coefficients of modular forms. Many mysteries remain about such sums, and even when good estimates exist in principle, it is not always easy to check that a concrete instance, encountered for some specific application, satisfies the assumptions of those results.

However, it is also the case that character sums are not sufficient for certain purposes, and more complicated summands are sometimes needed.

Example 8. In Example 5, we showed how Hyper-Kloosterman sums (which are additive character sums) are sufficient to describe the equidistribution properties of angles of Gauss sums. It may seem natural to try to do the same for the proof of the average Sato-Tate law of Katz (Example 4). The natural idea of computing the moments

$$\frac{1}{p-1}\sum_{a\in\mathbf{F}_p^\times}\Big(\frac{HK(2;a,p)}{2\sqrt{p}}\Big)^m = \frac{1}{p-1}\sum_{a\in\mathbf{F}_p^\times}(\cos\theta_{p,a})^m,$$

(as was done for Gauss sums) does not correspond to the Weyl criterion, as we defined it, because the functions $\theta \mapsto \cos(\theta)^m$, for $m \geqslant 0$ and $\theta \in [0,\pi]$, are not orthogonal for the target Sato-Tate measure (e.g., $\int (\cos\theta)^2 d\mu_{ST} = 1/4$).[7]

As it turns out (this will be justified in Example 25), the most natural orthonormal basis for $L^2([0,\pi],\mu_{ST})$ is the sequence $(U_m)_{m\geqslant 0}$ of Chebychev functions defined by

$$U_m(\theta) = \frac{\sin((m+1)\theta)}{\sin\theta}, \qquad m \geqslant 0,$$

[6] More precisely, we are looking here at the affine Weierstrass curve, with no point at infinity.

[7] Those moments can however be computed elementarily for the first few values of m; this is already enough (with $m = 2$) to check that the $(\theta_{p,a})_a$ are not uniformly distributed on $[0,\pi]$ as $p \to +\infty$. See also Example 31.

which are known to be of the form

$$U_m(\theta) = X_m(2\cos\theta),$$

where X_m is a polynomial in $\mathbf{Z}[X]$ of degree m (the Chebychev polynomials of the second kind).

Note that X_0 is the constant function 1, and thus the Weyl criterion indicates that the theorem of Katz is equivalent with the assertion that

$$\frac{1}{p-1}\sum_{a\in\mathbf{F}_p^\times} U_m(\theta_{p,a}) = \frac{1}{p-1}\sum_{a\in\mathbf{F}_p^\times} X_m\Big(\frac{HK(2;a,p)}{\sqrt{p}}\Big) \longrightarrow 0$$

as $p\to+\infty$, for $m\geqslant 1$. Since a ranges over $\mathbf{F}_p^\times = \mathbf{G}_m(\mathbf{F}_p)$, this suggests an algebraic framework where the summand

$$\Lambda(a) = X_m\Big(\frac{HK(2;a,p)}{\sqrt{p}}\Big)$$

is permitted. This can not be a character sum, for the simple reason that the summands are not of modulus 1 (as one can check easily numerically, if it does not seem clear enough).

To motivate the "black box" that we will need to introduce next, one may first start by reinterpreting additive character sums in a way that is generalizable to situations like that of the previous example. From a high-level arithmetic point of view, what is done is to replace analogues of *Dirichlet characters*[8] with analogues of *Galois-theoretic characters* or, in other words, it has something to do with reciprocity laws.

Many subsequent steps in the theory turn out to follow very naturally from this change of point of view, so to motivate it, we recall one formulation of the abelian reciprocity laws for the number field $\mathbf{Q}$. Let χ be a primitive Dirichlet character modulo $q\geqslant 1$, and let $K=K(\chi)$ be the cyclotomic field of q-th roots of unity, with ring of integer $\mathbf{Z}_K$. Then there exists a unique group homomorphism

$$\tilde{\chi} : \mathrm{Gal}(K/\mathbf{Q}) \longrightarrow \mathbf{C}^\times$$

which corresponds to χ as follows. For every prime p not dividing q, let $\mathfrak{p}\subset\mathbf{Z}_K$ be a prime ideal of $\mathbf{Z}_K$ containing $p\mathbf{Z}$, so that we have an extension $\mathbf{Z}/p\mathbf{Z}\subset\mathbf{Z}_K/\mathfrak{p}\mathbf{Z}_K$ of finite fields. There is then a well-defined Frobenius element $\mathrm{Fr}_p\in\mathrm{Gal}(K/\mathbf{Q})$, "lifting" the Frobenius automorphism $x\mapsto x^p$ acting on $\mathbf{Z}_K/\mathfrak{p}\mathbf{Z}_K$, and we have the reciprocity

$$\tilde{\chi}(\mathrm{Fr}_p) = \chi(p).$$

In fact, this character is easy to construct: since $\mathrm{Gal}(K/\mathbf{Q})\simeq(\mathbf{Z}/q\mathbf{Z})^\times$, with an isomorphism mapping Fr_p to $p\,(\mathrm{mod}\,q)$, one can define it by the composition

$$\mathrm{Gal}(K/\mathbf{Q})\simeq(\mathbf{Z}/q\mathbf{Z})^\times \xrightarrow{\chi} \mathbf{C}^\times$$

[8] Although those were only implicit in character sums; see [11, §11.5] for a concrete description of the Dirichlet character leading to Kloosterman sums.

(it is a much deeper fact that any Galois-character $\tilde{\chi}$ is obtained in this way from a Dirichlet character; indeed, this is a version of the Kronecker-Weber theorem).

Coming back to our setting of finite fields, if the base variety $V/\mathbf{F}_q$ is connected,[9] there is a certain group associated to V, the *algebraic fundamental group* of V, which is a compact topological group, denoted $\pi_1(V)$. This group was constructed by Grothendieck as a generalization of a Galois group, and its main property for our purpose is that it contains canonical *Frobenius conjugacy classes* Fr_{x,q^ν}, associated with any $\nu \geqslant 1$ and $x \in V(\mathbf{F}_{q^\nu})$, so that for *any* character sum $S(V, \Lambda_{f,g}; q)$ as above, there is a character

$$\chi_{f,g} : \pi_1(V) \to \mathbf{C}^\times$$

with the property that

$$\chi_{f,g}(\mathrm{Fr}_{x,q^\nu}) = \chi(N_{\mathbf{F}_{q^\nu}/\mathbf{F}_q} g(x))\psi(\mathrm{Tr}_{\mathbf{F}_{q^\nu}/\mathbf{F}_q} f(x)) \tag{7}$$

when $x \in V(\mathbf{F}_{q^\nu})$, and in particular, $\chi_{f,g}(\mathrm{Fr}_{x,q}) = \chi(g(x))\psi(f(x))$ if x is in $V(\mathbf{F}_q)$ (see, e.g., [11, p. 302] for a sketch of the construction when $g = 1$, which is related to the structure of Artin-Schreir extensions $y^q - y = f(x)$, in analogy with the cyclotomic extensions appearing for Dirichlet characters over $\mathbf{Q}$).

This gives an alternate form of the summand for character sums, and it is very natural from the point of view of harmonic analysis to generalize it by considering more general homomorphisms

$$\pi_1(V) \xrightarrow{\rho} GL(r, k),$$

where $r \geqslant 1$ and k is some field, and derive from them the summands of the type

$$\Lambda_\rho(x) = \mathrm{Tr}\, \rho(\mathrm{Fr}_{x,q}),$$

for $x \in V(\mathbf{F}_q)$ (taking the trace is justified by the fact that the Frobenius elements are only defined up to conjugacy: their trace *is* well-defined).

A minor difficulty is the choice of the coefficient field k we have introduced surreptitiously: in general, taking $k = \mathbf{C}$ leads to difficulties because a homomorphism with complex values is not continuous if the image of ρ is infinite (because the topology of $\pi_1(V)$ is totally disconnected). It is a fact[10] that the set of all Frobenius conjugacy classes

$$\mathrm{Fr}_{x,q^\nu}, \qquad \nu \geqslant 1, \qquad x \in V(\mathbf{F}_{q^\nu}),$$

is dense in $\pi_1(V)$, hence a *continuous* character is determined by the values taken on such classes. Therefore, the theory is developed with fields k equipped with a topology which is more compatible with that topology of $\pi_1(V)$. These fields depend on the choice of an auxiliary prime number ℓ distinct from the characteristic of $\mathbf{F}_q$ and are extensions of the ℓ-adic field $\mathbf{Q}_\ell$ (for instance, k could be an algebraic closure

[9] In a suitable algebraic sense; readers not familiar with the definition can restrict their attention to the following examples of connected algebraic varieties: (i) $\mathbf{A}^d$ or $\mathbf{G}_m^d$ for $d \geqslant 1$; (ii) the complement in $\mathbf{A}^d$ of a proper subvariety V.

[10] Which is closely related to the Chebotarev density theorem.

of $\mathbf{Q}_\ell$). For any $\ell \neq p$, the characters $\chi_{f,g}$ can be seen as a composition of a unique continuous homomorphism

$$\chi_{f,g} : \pi_1(V) \xrightarrow{\rho} k^\times$$

for some finite extension k of $\mathbf{Q}_\ell$, and an embedding $k^\times \xrightarrow{\iota} \mathbf{C}^\times$ which is the restriction of an injection $k \xrightarrow{\iota} \mathbf{C}$.

Such an injection (in fact, isomorphism) does exist for $k = \bar{\mathbf{Q}}_\ell$, but it is not unique (its construction requires the axiom of choice). However, for the discussion in this paper, this is not a very serious problem. (The reader may notice that the existence of ι means that one could, in principle, take $k = \mathbf{C}$ and just change the topology by means of ι).

Remark 9 (Geometric fundamental group). As we have hinted, the fundamental group can be considered as a type of Galois group (another analogy is with the topological fundamental group of a topological space, seen as a group of automorphisms of the universal cover).

Later on, it will also be important to use the *geometric fundamental group* $\pi_1(\bar{V})$, which is associated in a similar way with the "geometric" variety $\bar{V}$, obtained from V by forgetting its field of definition ($\bar{V}$ can most conveniently here be identified with the set $V(\bar{\mathbf{F}}_q)$ of points of V with coefficients in an algebraic closure of $\mathbf{F}_q$). There is a natural inclusion homomorphism

$$\pi_1(\bar{V}) \longrightarrow \pi_1(V),$$

and this makes $\pi_1(\bar{V})$ into a normal subgroup of $\pi_1(V)$. The quotient is well-understood (it is abelian, and topologically generated by a single element).

The Frobenius conjugacy classes do not lie in $\pi_1(\bar{V})$, since they are related to arithmetic properties of $V/\mathbf{F}_q$ and its rational points, with the field of definition taken into account.

Example 10. Let us immediately show that this generalization is, at least formally, able to handle the Weyl sums for the average Sato-Tate law (Example 8). We can write

$$\frac{HK(2;a,p)}{\sqrt{p}} = 2\cos\theta_{p,a} = e^{i\theta_{p,a}} + e^{-i\theta_{p,a}} = \mathrm{Tr}\begin{pmatrix} e^{i\theta_{p,a}} & 0 \\ 0 & e^{-i\theta_{p,a}} \end{pmatrix}, \tag{8}$$

which may suggest a 2-dimensional representation. Note that the left-hand side makes sense in any field of characteristic zero containing $\sqrt{p}$ and the p-th roots of unity (after identifying the exponential $e(1/p)$ with a primitive p-th root in the definition of $HK(2;a,p)$).

Moreover, there is a well-known interpretation of the Chebychev polynomials in similar terms:

$$X_m(2\cos\theta) = e^{mi\theta} + e^{(m-2)i\theta} + \cdots + e^{-(m-2)i\theta} + e^{-mi\theta} \tag{9}$$

(a sum of $m+1$ terms); this is the trace of the corresponding diagonal matrix, of course, but more pointedly, this matrix is the image of the 2-dimensional matrix

$$g(\theta) = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}$$

under the homomorphism

$$SL(2,k) \xrightarrow{\mathrm{Sym}^m} SL((m+1),k)$$

called the m-th symmetric power. This makes sense for k any algebraically closed field of characteristic zero. Concretely, one can think of this as follows: consider the $(m+1)$-dimensional k-vector space H_m of homogeneous forms of degree m in two variables, say X and Y. This space is spanned by the basis

$$(X^m, X^{m-1}Y, \dots, XY^{-m-1}, Y^m),$$

and for $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,k)$, one can define $\mathrm{Sym}^m(g)$ as the matrix, in this basis, of the linear map

$$P(X,Y) \mapsto P(aX+bY, cX+dY)$$

(e.g., for $g = g(\theta)$, we get $X^kY^{n-k} \mapsto e^{ik\theta+i(k-n)\theta}X^kY^{n-k}$, so that the trace of $\mathrm{Sym}^m(g)$ is indeed given by (9)).

Hence, *if* we can write

$$\sum_{a\in\mathbf{F}_p^\times} \frac{HK(2;a,p)}{\sqrt{p}} = \sum_{a\in\mathbf{G}_m(\mathbf{F}_p)} \mathrm{Tr}\,\mathcal{HK}(\mathrm{Fr}_{a,p}) \tag{10}$$

for some 2-dimensional representation $\pi_1(\mathbf{G}_m) \xrightarrow{\mathcal{HK}} GL(2,k)$, we get for free that

$$\sum_{a\in\mathbf{F}_p^\times} X_m\Big(\frac{HK(2;a,p)}{\sqrt{p}}\Big) = \sum_{a\in\mathbf{G}_m(\mathbf{F}_p)} \mathrm{Tr}(\mathrm{Sym}^m \circ \mathcal{HK})(\mathrm{Fr}_{a,p})$$

for any $m \geqslant 1$.

For later purposes, one remark is important: these sums (once they are known to actually exist) are one-parameter sums, over a very simple algebraic variety. Hence, their complexity resides almost entirely in the summand involved.

A last condition is needed before we can introduce formally a pretty good category of possible summands. It has to do with the necessity to ensure that, even if we need to apply an injection ι to pass from k to $\mathbf{C}$, the values $\Lambda_\rho(x)$ remain controlled. This is achieved by restricting them to numbers which are algebraic over $\mathbf{Q}$ (as the values $HK(2;a,p)/\sqrt{p}$ are, for instance), and indeed of a special type.

Definition (Weil number). Let k be a field of characteristic zero, $q \neq 1$ a power of a prime number p and $m \in \mathbf{Z}$ an integer. An element $\alpha \in k$ is a *q-Weil number of weight m* if and only if, α is algebraic over $\mathbf{Q}$ and for any embedding $\iota : \mathbf{Q}(\alpha) \hookrightarrow k \hookrightarrow \mathbf{C}$, we have

$$|\iota(\alpha)| = q^{m/2}.$$

Concretely, if k is itself a subfield of $\mathbf{C}$, this means that α satisfies an irreducible polynomial equation $P(\alpha) = 0$ with $P \in \mathbf{Q}[X]$ and that all the roots β of $P(X) = 0$ have the same modulus $q^{m/2}$.

Example 11. (1) The simplest example is given by $\alpha = 1$, a q-Weil number of weight 0 for any q; more generally, any root of unity ξ is a q-Weil number of weight 0 since all its conjugates are also roots of unity and hence lie on the unit circle when embedded in $\mathbf{C}$. Also, for any q, $\xi q^{m/2}$ is a q-Weil number of weight m.

(2) Let χ be a non-trivial Dirichlet character modulo p and $a \in (\mathbf{Z}/p\mathbf{Z})^\times$. The Gauss sums $\tau_a(\chi) \in \mathbf{C}$ (see (4)) are then p-Weil numbers of weight 1. Indeed, we recalled already that

$$|\tau_a(\chi)| = \sqrt{p},$$

and from the definition we note that $\tau_a(\chi)$ is an algebraic integer (a sum of $p-1$ roots of unity), and in fact $\tau(\chi) \in \mathbf{Q}(e(1/p), e(1/(p-1)))$, since the values of χ are roots of unity of order dividing $p-1$. The Galois conjugates of $\tau(\chi)$ are therefore obtained by applying all the automorphisms of this cyclotomic field; each of these (say, σ) acts on $e(1/p)$ by

$$\sigma(e(1/p)) = e(b/p)$$

for some $b \in (\mathbf{Z}/p\mathbf{Z})^\times$, and has the property that

$$\tilde{\chi}(x) = \sigma(\chi(x))$$

is itself a non-trivial Dirichlet character modulo p. So we have

$$\sigma(\tau_a(\chi)) = \sum_{x \in \mathbf{F}_p^\times} \tilde{\chi}(x) e\Big(\frac{abx}{p}\Big) = \tau_{ab}(\tilde{\chi}),$$

which is of modulus $\sqrt{p}$ again.

(3) For p prime and $a \in \mathbf{F}_p^\times$, we have

$$HK(2; a, p) = 2\sqrt{p}\cos\theta_{p,a} = \sqrt{p} \cdot e^{i\theta_{p,a}} + \sqrt{p} \cdot e^{-i\theta_{p,a}}$$

and each of $\sqrt{p}e^{\pm i\theta_{p,a}}$ is a p-Weil numbers of weight 1, while $e^{\pm i\theta_{p,a}}$ is a p-Weil number of weight 0 (which is *not* a root of unity). This can be checked by a Galois conjugation argument similar to (2).

Finally, we can define:

Definition (Lisse sheaves; summands for algebraic sums). Let $q \neq 1$ be a power of a prime p and let $V/\mathbf{F}_q$ be a connected algebraic variety. Let $\ell \neq p$ be a prime number and k an ℓ-adic field, for instance an algebraic closure of $\mathbf{Q}_\ell$.

(1) A lisse sheaf ρ, pointwise of weight $m \in \mathbf{Z}$, on $V/\mathbf{F}_q$ is a continuous homomorphism

$$\pi_1(V) \xrightarrow{\rho} GL(r, k)$$

for some $r \geqslant 1$, with the property that for any $\nu \geqslant 1$, and any $x \in V(\mathbf{F}_{q^\nu})$, all eigenvalues[11] of $\rho(\mathrm{Fr}_{x,q^\nu}) \in GL(r,k)$ are all q^ν-Weil numbers of weight m.

(2) An algebraic exponential sum $S(V, \Lambda; q)$ over $\mathbf{F}_q$ is a sum of the type

$$\sum_{x \in V(\mathbf{F}_q)} \Lambda(x)$$

[11] Those are well-defined, although the Frobenius is only well-defined as a conjugacy class.

where $V/\mathbf{F}_q$ is an algebraic variety, and $\Lambda(x) = \operatorname{Tr}\rho(\mathrm{Fr}_{x,q})$ is the trace function associated to a lisse sheaf on V, of some weight $m \in \mathbf{Z}$. We also denote

$$S_\nu(V, \Lambda; q) = \sum_{x \in V(\mathbf{F}_{q^\nu})} \operatorname{Tr}\rho(\mathrm{Fr}_{x,q^\nu})$$

for $\nu \geqslant 1$.

Although we have $S(V, \Lambda; q) \in k$, note that the sum

$$\sum_{x \in V(\mathbf{F}_q)} \Lambda(x)$$

is an algebraic number, and we usually tacitly see it as a complex number using any embedding of the field it generates to $\mathbf{C}$.

Example 12. (1) Mixed character sums are all of this type, with weight 0, since the only eigenvalue of the characters $\chi_{f,g}$ of $\pi_1(V)$ at $x \in V(\mathbf{F}_{q^\nu})$ is the value

$$\chi_{f,g}(\mathrm{Fr}_{x,q^\nu}) = \chi(N_{\mathbf{F}_{q^\nu}/\mathbf{F}_q} g(x))\psi(\operatorname{Tr}_{\mathbf{F}_{q^\nu}/\mathbf{F}_q} f(x))$$

itself, which is a root of unity.

A special case, which is historically and practically quite important, is when the character sum is trivial, i.e., $\Lambda(x) = 1$. Then the sum reduces to

$$\sum_{x \in V(\mathbf{F}_q)} 1 = |V(\mathbf{F}_q)|$$

and the properties of these numbers of rational points on varieties over finite fields were the subject of Weil's original conjectures. The case (Example 1) of the Hasse bound is one of the simplest non-trivial cases.

(2) Part of the proof of the average Sato-Tate conjecture is the fact, due to Deligne, that – as was hypothesized in Example 10 – for any p, there exists a lisse sheaf $\mathcal{HK}_1$ on $\mathbf{G}_m/\mathbf{F}_q$, of weight 1 and rank $r = 2$, such that

$$\operatorname{Tr}\mathcal{HK}_1(\mathrm{Fr}_{a,p}) = HK(2; a, p) \tag{11}$$

for any $a \in \mathbf{F}_p^\times$, and in fact

$$\operatorname{Tr}\mathcal{HK}_1(\mathrm{Fr}_{a,p^\nu}) = \sum_{x \in \mathbf{F}_{p^\nu}^\times} e\Big(\frac{\operatorname{Tr}_{\mathbf{F}_{p^\nu}/\mathbf{F}_p}(x + a/x)}{p}\Big)$$

for $\nu \geqslant 1$ and $a \in \mathbf{G}(\mathbf{F}_{p^\nu})$. (Properly speaking, such sheaves exist for every choice of the auxiliary prime ℓ used to define the coefficient field k, which must contain the p-th roots of unity; moreover, the value in k of the sum is obtained by fixing the meaning of $e(1/p)$ to be one of these primitive p-th roots of unity).

(3) As the rest of Example 10 suggests, one can create new algebraic exponential sums out of old ones using composites

$$\pi_1(V) \xrightarrow{\rho} GL(r, k) \xrightarrow{\pi} GL(s, k)$$

where π is an algebraic homomorphism (i.e., all entries of $\pi(g)$, for $g \in GL(r, k)$, are polynomial functions of the coordinates of g). It is a fact that if ρ is a lisse sheaf

pointwise of weight m, the composite $\pi \circ \rho$ is also pointwise of some weight (see [18, Proof of Th. 9.2.6]).

Remark 13. More general settings exist for algebraic exponential sums (see, e.g., the first chapter of [12]). The restriction to the setting above is made for simplicity, and is also due partly to the author's restricted knowledge. If $V/\mathbf{F}_q$ is not connected, the fundamental group is not sufficient to capture "independent" sums on the connected components, but of course these can be treated separately in most applications. More significantly, N. Katz (unpublished) has found a way to treat some sums where the summation sets are, for instance, the sets of multiplicative characters of $\mathbf{F}_{q^\nu}$ (which are not of the type $V(\mathbf{F}_{q^\nu})$ for an algebraic variety $V/\mathbf{F}_q$ which is *independent* of ν, as one can see by comparing the number of elements $\varphi(q^\nu - 1)$ with the asymptotic for $|V(\mathbf{F}_q)|$ that follow from the Riemann Hypothesis...)

The next section will describe the fundamental results of Grothendieck and Deligne concerning the structural properties and bounds for algebraic exponential sums.

4. The cohomological formalism

The apparent complexity of the definition of algebraic exponential sums is richly rewarded by the powerful formalism that becomes available to handle such sums.

The first tool is the *Lefschetz-Grothendieck trace formula*, which, after application of the Riemann Hypothesis, reinterprets the sums $S(V, \Lambda; q)$ as combinations of suitable Weil numbers (of higher weight than that of the summand $\Lambda(x) = \operatorname{Tr}\rho(\operatorname{Fr}_{x,q})$) which can be thought of, intuitively, as isolating not only a "main term", but also lower frequencies (in a certain sense).

More precisely, given $V/\mathbf{F}_q$ (connected) and a lisse sheaf $\rho : \pi_1(V) \to GL(r, k)$, pointwise of weight m, Grothendieck and his collaborators first constructed a sequence of finite-dimensional k-vector spaces

$$H_c^i(\bar{V}, \rho), \qquad i \geqslant 0,$$

called the i-th cohomology "group" of the geometric variety[12] $\bar{V}$ with compact support and coefficients in ρ. The fact that V is defined with equations with coefficients in $\mathbf{F}_q$ can be recovered by keeping track of the Frobenius automorphism $\varphi : x \mapsto x^q$ which acts on $\bar{V}$ (as well as its inverse F, which is called the "geometric" Frobenius). In fact, for $\nu \geqslant 1$, we have

$$V(\mathbf{F}_{q^\nu}) = |\{x \in \bar{V} \mid F^\nu(x) = x\}| = \bar{V}^{F^\nu},$$

i.e., $V(\mathbf{F}_q)$ is the subset of $\bar{V}$ consisting of the fixed points of F.

The cohomology groups turn out to encode the algebraic sums in a rather remarkable way. First, it is a general principle that any natural construction performed on $\bar{V}$ will lead to an object where the action of F remains visible in some "induced"

[12] Recall (Remark 9) that $\bar{V}$ can be interpreted as the set of all points of V with coordinates in an algebraic closure of $\mathbf{F}_q$.

way. For the cohomology group $H_c^i(\bar{V}, \rho)$, which is a k-vector space, this means that there exists a k-linear map

$$F \ : \ H_c^i(\bar{V}, \rho) \to H_c^i(\bar{V}, \rho)$$

"naturally induced" by F. Then, the trace formula states that

$$S(V, \Lambda; q) = \sum_{i \geqslant 0} (-1)^i \operatorname{Tr}(F \mid H_c^i(\bar{V}, \rho)), \tag{12}$$

where this seemingly infinite sum contains in fact only finitely many terms because $H_c^i(\bar{V}, \rho)$ is zero for all i large enough – in fact, it is zero for $i > 2d$, where d is the dimension of $\bar{V}$. More generally, we get

$$S_\nu(V, \Lambda; q) = \sum_{i \geqslant 0} (-1)^i \operatorname{Tr}(F^\nu \mid H_c^i(\bar{V}, \rho))$$

(when passing from $\mathbf{F}_q$ to $\mathbf{F}_{q^\nu}$, $\bar{V}$ doesn't change, but the Frobenius becomes $x \mapsto x^{q^\nu}$, with inverse F^ν).

Example 14 (Counting points). In the particular case where $\Lambda(x) = 1$ (i.e., ρ is the trivial representation), we get a formula for the number of $\mathbf{F}_q$-rational points on V:

$$|V(\mathbf{F}_q)| = \sum_{i \geqslant 0} (-1)^i \operatorname{Tr}(F \mid H_c^i(\bar{V})), \qquad H_c^i(\bar{V}) = H_c^i(\bar{V}, \text{trivial}). \tag{13}$$

The existence of such a formula was already conjectured by Weil, based on various examples (in particular, curves, see Example 16). However, Weil did not know how to define the spaces $H_c^i(\bar{V})$ in general.

Weil showed how such a formula implied the rationality of the relevant zeta function. Indeed, coming back to the general case, let $[\bar{V}]$ denote the set of orbits of F acting on $\bar{V}$ (if $V = \mathbf{A}^1$, this can be identified with the set of irreducible monic polynomials in $\mathbf{F}_q[T]$, each polynomial π corresponding to the set of its roots, which is a single orbit because of irreducibility). For $[x] \in [\bar{V}]$, let $\deg(x)$ be the number of elements in this orbit and $|x| = q^{\deg(x)}$. Then, define

$$L(V, \rho) = \prod_{[x] \in [\bar{V}]} \det(1 - T^{\deg(x)} \rho(\operatorname{Fr}_{x,|x|}))^{-1}$$

which is a type of Euler product, seen here as a formal power series in $k[[T]]$ (or indeed in $\mathbf{Q}[[T]]$, if $\rho = 1$).

Now, a familiar computation leads to the alternate expression

$$L(V, \rho) = \exp\Big(\sum_{\nu \geqslant 1} \frac{S_\nu(V, \Lambda; q) T^\nu}{\nu}\Big),$$

(which can also be taken as the definition of the L-function of ρ). From this, we see that (as formal power series) we have

$$T \frac{L'}{L}(V, \rho) = \sum_{\nu \geqslant 1} S_\nu(V, \Lambda; q) T^\nu,$$

and an application of the trace formula (for each ν) gives

$$T\frac{L'}{L}(V,\rho) = \sum_{i=0}^{2d}(-1)^i \sum_{\nu\geqslant 1}\mathrm{Tr}(F^\nu \mid H_c^i(\bar{V},\rho))T^\nu$$
$$= -\sum_{i=0}^{2d}(-1)^i T\frac{d}{dT}\log\det(1-TF \mid H_c^i(\bar{V},\rho)),$$

and hence implies the rationality

$$L(V,\rho) = \frac{P_1(T)\cdots P_{2d-1}(T)}{P_0(T)\cdots P_{2d}(T)}$$

where

$$P_i(T) = \det(1-TF \mid H_c^i(\bar{V},\rho)). \tag{14}$$

Note also that, for each factor $P_i(T)$ of the L-function, we have a *spectral interpretation* of its zeros, as related to the eigenvalues of the Frobenius acting on $H_c^i(\bar{V},\rho)$. At this point one must however still be careful that P_i is, a priori, in $k[T]$, and its eigenvalues may be arbitrary elements of k (in particular, they could be non-algebraic, and the algebraicity of $S(V,\Lambda;q)$ might arise by cancellation of transcendental terms).

Example 15 (0-dimensional case). Only one example of the Lefschetz trace formula can be presented completely elementarily, namely the case where V is 0-dimensional and $\rho = 1$ (although V is not connected then in general, the situation is simple enough to analyze). If $V/\mathbf{F}_q$ is defined as the zero set of a non-zero polynomial $f \in \mathbf{F}_q[X]$, the set $\bar{V}$ is just the collection of all zeros in an algebraic closure (note that multiplicity is allowed), permuted by the Frobenius F. We therefore get

$$|V(\mathbf{F}_q)| = |\{x\in\mathbf{F}_q \mid f(x)=0\}| = \mathrm{Tr}(F \mid H_c^0(\text{zeros of } f))$$

and this is interpreted as follows: the space H_c^0 is here d-dimensional, where $d = |\bar{V}|$ is the number of distinct zeros of f in an algebraic closure; moreover, one can find a natural basis $(e_x)_{x\in\bar{V}}$ of H_c^0 in such a way that the induced action of Frobenius is determined by

$$F(e_x) = e_{F(x)}.$$

In other words, F is a permutation matrix in the basis (e_x), associated to the permutation of the roots induced by the Frobenius. Then the trace formula becomes the well-known fact that the trace of a permutation matrix is the number of fixed points of the permutation.

The zeta function identity

$$L(V,1) = \prod_{[x]\in[\bar{V}]}(1-T^{\deg(x)})^{-1} = \frac{1}{\det(1-TF \mid H_c^0(\bar{V}))} \tag{15}$$

is also clear, since the product on the left-hand side is finite: it represents the factorization of the characteristic polynomial of a permutation matrix as a product of cyclotomic factors $(1-T^a)$ where a runs over the lengths of the cycles occurring in a representation of a permutation as a product of disjoint cycles.

Example 16 (Algebraic curves). For a smooth projective geometrically connected curve $C/\mathbf{F}_q$ (e.g., a curve $y^2 = f(x)$, with f squarefree of odd degree, with the addition of a point at infinity), we get

$$\exp\Big(\sum_{\nu\geqslant 1}\frac{|C(\mathbf{F}_{q^\nu})|T^\nu}{\nu}\Big) = \frac{P_1(T)}{P_0(T)P_2(T)},$$

where $\deg(P_1) = 2g$, $g \geqslant 0$ being an important invariant called the *genus* of the curve. One can show easily that $P_0(T) = 1 - T$ and $P_2(T) = 1 - qT$ (the latter is related to (17)). The polynomial

$$P_1(T) = \det(1 - TF \mid H^1(\bar{C}, \text{trivial})) = \prod_{1\leqslant j\leqslant 2g}(1 - \alpha_j T), \tag{16}$$

which is often called the L-function of C, satisfies the functional equation

$$P_1(T) = q^g T^{2g} P_1\Big(\frac{1}{qT}\Big),$$

which amounts to saying that one can order the roots α_j in pairs so that $\alpha_j\alpha_{2g+1-j} = q$ for $1 \leqslant j \leqslant 2g$.

The Riemann Hypothesis, in this case, is the statement that $|\alpha_j| = \sqrt{q}$ for all j, but notice that, although we have a spectral interpretation via the Frobenius automorphism acting on cohomology, this is not sufficient to obtain this!

This case of the Riemann Hypothesis was first proved by Weil, who used constructions related to the Jacobian variety of the curve and torsion points to – in effect – construct the dual of the necessary cohomology groups.

Remark 17. The definition and construction of the étale cohomology groups $H^i_c(\bar{V}, \rho)$ is a deep and subtle achievement which is unfortunately hard to explain in a few words, even to an audience familiar with classical algebraic topology. The proof of the trace formula, even given this construction, is also difficult. The simplest case (after the 0-dimensional one) is that of counting points on elliptic curves, in which case the elementary theory of isogenies and of the Weil pairing can lead to a relatively elementary proof (see [23, §V.2]).

Here is a very simple observation that can at least give a first impression of what is happening:

- Localizing close to a point x in the Zariski topology amounts to allowing "new" functions $1/f$, where f is regular at x, but may vanish elsewhere;
- "Localizing" close to x in the étale topology allows new functions $g(x)$ which satisfy (separable) *algebraic equations*, e.g., $g(x) = \sqrt{x}$ on $\mathbf{G}_m$, "defined" via the second projection

$$R = \{(x, y) \mid x = y^2\} \longrightarrow \mathbf{G}_m.$$

It may be noticed that the second case is very close to the idea of Riemann for defining Riemann surfaces of algebraic functions.

From the point of view of finding estimates for $S(V,\Lambda;q)$, nothing is achieved purely from applying the trace formula to get (12), despite its fundamental nature, because the right-hand side is not yet under control: each individual term in the alternating sum might be extremely large (not to mention the fact that, in principle, these terms are in the field k, and not in $\mathbf{C}$, and might not be algebraic). However, it was part of Weil's conjecture that the situation should in fact be much better than this: under appropriate circumstances, for $\Lambda = 1$, Weil conjectured that the Frobenius acting on each cohomology group H_c^i should have eigenvalues which are q-Weil numbers of weight exactly i.

This conjecture was proved by Deligne [5], who then succeeded in [6] in finding a much more general statement which encompasses all the algebraic sums we have considered – well beyond the original Weil conjectures:

Theorem 18 (Deligne). *Let q be a power of a prime p, $m \in \mathbf{Z}$, and consider an algebraic exponential sum over $\mathbf{F}_q$,*

$$S(V,\Lambda;q) = \sum_{x\in V(\mathbf{F}_q)} \Lambda(x)$$

where $V/\mathbf{F}_q$ is a connected algebraic variety of dimension d, and $\Lambda(x) = \operatorname{Tr}\rho(\operatorname{Fr}_{x,q})$ is the trace function associated to a lisse sheaf of weight m on V.

Then, for every i, $0 \leqslant i \leqslant 2d$, every eigenvalue of F acting on $H_c^i(\bar{V},\rho)$ is a q-Weil number – in fact, also an algebraic integer – of some weight, which is $\leqslant m+i$.

Under nice circumstances (related to smoothness and compactness and usually involving some form of Poincaré duality to transform upper bounds into lower bounds, or in analytic terms, related to the existence of a good functional equation for the L-functions), this can be refined to an equality: the eigenvalues α for H_c^i are then q-Weil numbers exactly of weight $m+i$. This was the case of the original Weil conjectures, but is not true in general.

Example 19. (1) Consider counting points on $\mathbf{G}_m/\mathbf{F}_q$. Of course, the answer is known: $|\mathbf{G}_m(\mathbf{F}_q)| = q-1$. In the cohomological interpretation, we have $H_c^2(\bar{\mathbf{G}}_m)$ of dimension 1 with F acting by multiplication by q, $H_c^1(\bar{\mathbf{G}}_m)$ of dimension 1, with F acting by multiplication by 1, and $H_c^0(\bar{\mathbf{G}}_m)$ of dimension 0.

More generally, if V is of dimension d, the formula (13) and Theorem 18 lead to

$$|V(\mathbf{F}_{q^\nu})| = \operatorname{Tr}(F^\nu \mid H_c^{2d}(\bar{V},\text{trivial})) + O(q^{\nu(d-1/2)})$$

for $\nu \geqslant 1$, where the implied constant depends on $\bar{V}$. Intuitively, we expect this number of points to be of order of magnitude $q^{\nu d}$ (by comparison with affine space of dimension d), and this can be confirmed using one of the few general formulas for computing cohomology: for $V/\mathbf{F}_q$ smooth,[13] and for any lisse sheaf $\rho : \pi_1(V) \longrightarrow GL(r,k)$ on V, we have

$$H_c^{2d}(\bar{V},\rho) \simeq \rho_{\pi_1(\bar{V})} \tag{17}$$

[13] In terms of equations (F_i) for V, this means that there is no point on V where all the partial derivatives $\partial_{x_j} F_i$ vanish.

where the right-hand side, the coinvariant space of ρ *under the action of the geometric fundamental group* (see Remark 9), is defined as the quotient k-vector space

$$k^r/\langle \rho(g)v - v,\ v \in k^r,\ g \in \pi_1(\bar{V})\rangle$$

(the largest quotient of k^r on which $\pi_1(\bar{V})$ acts trivially).

For the trivial sheaf, $r = 1$ and $\rho(g)v = v$ for all v and g, so we find that $H_c^{2d}(\bar{V}, \text{trivial})$ is of dimension 1. The action of F on this space can also be computed to be multiplication by q^d, if V is geometrically irreducible, and it follows that we have

$$|V(\mathbf{F}_{q^\nu})| = q^{\nu d} + O(q^{\nu(d-1/2)}) \tag{18}$$

for $\nu \geqslant 1$ (the implied constant depending on $V/\mathbf{F}_q$). In fact, this asymptotic formula was first proved by Lang and Weil (by reducing to the case of curves), and it is *equivalent* with the assumption of geometric irreducibility.[14]

(2) One can easily illustrate on simple examples what happens if $V/\mathbf{F}_q$ is not geometrically irreducible. Take for instance the curve with equation

$$V\ :\ x^2 + y^2 = 0.$$

If -1 is not a square in $\mathbf{F}_q$, let $\varepsilon \in \mathbf{F}_{q^2}$ such that $-1 = \varepsilon^2$. Over $\mathbf{F}_{q^2}$, the equation of the variety splits as

$$(x + \varepsilon y)(x - \varepsilon y) = 0,$$

and so $\bar{V}$ is the union of two lines in the plane, but those two lines are *not* defined over $\mathbf{F}_q$, only over $\mathbf{F}_q(\varepsilon) = \mathbf{F}_{q^2}$ (and the Frobenius of $\mathbf{F}_q$ exchanges them since $F(\varepsilon) = -\varepsilon$). We get

$$|V(\mathbf{F}_{q^{2\nu}})| = 2q^{2\nu} - 1, \qquad |V(\mathbf{F}_{q^{2\nu+1}})| = 1,$$

for $\nu \geqslant 1$.

Thus if some components of $V/\mathbf{F}_q$ are only defined over finite extensions of $\mathbf{F}_q$, one may obtain different leading terms for $V(\mathbf{F}_{q^\nu})$ depending on the value of ν. This is also clear when V is of dimension 0, defined by the single equation $F(x) = 0$ in one variable: if there are n_i distinct irreducible factors of F of degree i, we have

$$|V(\mathbf{F}_{q^\nu})| = \sum_{i|\nu} n_i.$$

(3) In terms of L-functions, we can factor each polynomial P_i given by (14) as

$$P_i(T) = \prod_{1 \leqslant j \leqslant \dim H_c^i} (1 - \alpha_{j,i}T)$$

with $\alpha_{j,i}$ running over the eigenvalues of F on $H_c^i(\bar{V}, \rho)$. For the complex function $P_i(q^{-s})$, $s \in \mathbf{C}$, Deligne's Theorem translates to the following analogue of the classical Riemann Hypothesis: the zeros of $P_i(q^{-s})$ lie on a union of finitely many lines of the type $\text{Re}(s) = j/2$, where j is an integer such that $j \leqslant m + i$.

[14] For some analytic applications, this means it can sometimes be used to check the latter, instead of using more algebraic results.

(4) In cohomological terms, Deligne's proof of the bound (3) for Hyper-Kloosterman sums was obtained by showing that for the relevant additive character sum on $V = \mathbf{G}_m^{n-1}$ (see (6)), we have

$$H_c^i(\bar{V}, \Lambda_{\sum X_i + a/\prod X_i}) = 0$$

if $i \neq n-1$, while

$$\dim H_c^{n-1}(\bar{V}, \Lambda_{\sum X_i + a/\prod X_i}) = n.$$

Since, for character sums, the summand has weight 0, we get the upper bound (3). In fact, in that case, Deligne also showed that the eigenvalues on H_c^{n-1} are all of weight $n-1$.

(5) In [5], Deligne also proved the following remarkable estimate for additive character sums (which Weil had proved for $n = 1$): consider

$$\Lambda_F(x) = e\Big(\frac{F(x)}{p}\Big),$$

where $F \in \mathbf{F}_q[X_1, \dots, X_n]$ is a polynomial of degree d, with $(d, p) = 1$, in n variables such that the homogeneous part of degree d, say F_d, defines a smooth projective hypersurface (for instance, $F_d = X_1^d + \cdots + X_n^d$; the components of smaller degree can then be chosen arbitrarily). Then, considering the sums over $V = \mathbf{A}^n/\mathbf{F}_q$, Deligne proved that

$$H_c^i(\bar{V}, \Lambda_F) = 0 \text{ if } i \neq n, \qquad \dim H_c^n(\bar{V}, \Lambda_F) = (d-1)^n.$$

Hence we get the uniform upper bound

$$\Big| \sum_{x_1, \dots, x_n \in \mathbf{F}_p} e\Big(\frac{F(x)}{p}\Big) \Big| \leqslant (d-1)^n p^{n/2} \tag{19}$$

for such polynomials. Here also, Deligne proved that the eigenvalues in H_c^n are all of weight exactly n.

As these examples suggest, Deligne's Theorem 18 becomes very powerful when combined with computations of the dimension of the cohomology groups, in particular with results of *vanishing* of cohomology groups. Indeed, a rough general bound that can be obtained is

$$\Big| \sum_{x \in V(\mathbf{F}_q)} \Lambda(x) \Big| \leqslant C q^{m + k/2} \tag{20}$$

where

$$k = \max\{i \mid H_c^i(\bar{V}, \rho) \neq 0\}, \qquad C = \sum_i \dim H_c^i(\bar{V}, \rho). \tag{21}$$

The trivial bound, in view of the point counting formula (18), is $k = 2d$, where d is the dimension; if one can show that $k < 2d$, a non-trivial estimate immediately follows for the sums $S_\nu(V, \Lambda; q)$ as $\nu \to +\infty$ (i.e., for the vertical direction; in horizontal direction, one needs to control C, which depends on p). This basic goal is often easy to derive from the coinvariant formula (17).

Example 20 (Ubiquity of non-trivial bounds). Consider a character sum associated with the summand (5). In this case, the underlying lisse sheaf is the character $\chi_{f,g} : \pi_1(V) \to k^\times$ such that (7) holds, so the rank is 1. Consequently, the coinvariant space, which is $H_c^{2d}(\bar{V}, \chi_{f,g})$ can only be of dimension 0 (in which case we get that $k < 2d$) or 1, and the second case is only possible if $\chi_{f,g}$ is trivial on $\pi_1(\bar{V})$. This does not mean, however, that $\chi_{f,g}$ is entirely trivial – as we mentioned in Remark 9, the Frobenius conjugacy classes do not belong to $\pi_1(\bar{V})$. But it shows that $\chi_{f,g}(\gamma)$ only depends on the image of γ in the quotient group $\pi_1(V)/\pi_1(\bar{V})$. This group is known, by Grothendieck's theory, to be isomorphic to $\hat{\mathbf{Z}}$, and because we have $\mathrm{Fr}_{x,q^\nu} \mapsto -\nu$ in this isomorphism, this means concretely that in this situation, $\chi_{f,g}(\mathrm{Fr}_{x,q^\nu})$ depends only on ν. In particular, this implies that $\Lambda_{f,g}(x)$ is constant for every $x \in V(\mathbf{F}_q)$.

We can summarize this roughly as follows: for a character sum, *either* the summand is constant (and no cancellation can be expected!), *or* there is some non-trivial (vertical) estimate. Observe also how the underlying group structure was used to derive this conclusion.

Obtaining a non-trivial bound is, however, sometimes not sufficient. One often requires the best possible situation, in which we have $k = d$ (the dimension of V, as we saw in Example 5 for equidistribution of angles of Gauss sums), together with an explicit formula or upper bound for C. Hyper-Kloosterman sums and Deligne polynomials are of this particularly nice type, and it remains an active area of research to give convenient criteria to compute k and C for "concrete" exponential sums, and in particular to bound C uniformly with respect to various parameters that may occur.

Example 21 (Uniformity). We illustrate this last point in the case of character sums. Here, the most important dependency in analytic applications is that with respect to p. The basic problem is that even if the parameter variety V is defined uniformly (by reduction modulo primes of equations which are independent of p, the basic examples being affine spaces and $\mathbf{G}_m$), and even if the summands are defined (say) by (5) with $g = 1$ and a fixed polynomial f with integral coefficients (which can again be reduced modulo every primes to construct the corresponding sums over $V(\mathbf{F}_p)$), it remains a fact that the sheaves which encode those summands are, a priori, constructed for each p *separately*, hence the corresponding C depends on p.

Examples like Deligne polynomials suggest that C should be independent of p, or at least bounded independently of p. This was confirmed by Bombieri, using p-adic techniques to complement the ℓ-adic formalism, then extended by Adolphson and Sperber for all character sums, and the most general version is due to Katz [16].

Applied in the way we have sketched in this section, the formalism of the trace formula and the Riemann Hypothesis can be seen as a way of re-expressing exponential sums (in a highly non-trivial, indeed, non-combinatorial way) as another sum, where a usually non-trivial estimation is possible by using that bluntest of tools, the triangle inequality.

5. Families, monodromy and Deligne's equidistribution theorem

The last remark of the previous section strongly suggests that, although estimates like (19) might be best possible in vertical respect, some improvements should hold in cases where the cohomology groups involved have high dimension.

For instance, a character sum over $\mathbf{F}_p^n$ with a Deligne polynomial in $n \geqslant 2$ variables and of degree d is expressed as a sum of $(d-1)^n$ Weil numbers, each with modulus $p^{n/2}$, say

$$p^{n/2}e^{2i\pi\theta_{p,j}(F)}, \qquad 1 \leqslant j \leqslant (d-1)^n,$$

where the phases $\theta_{p,j}(F) \in [0,1]$ might be expected to be themselves quite random. Summing those Weil numbers could therefore be expected to lead to smaller values of the sums

$$\sum_{x_1,\dots,x_n \in \mathbf{F}_p} e\Big(\frac{F(x)}{p}\Big) = p^{n/2} \sum_{j=1}^{(d-1)^n} e^{2i\pi\theta_{p,j}(F)}$$

at least for most polynomials F, with respect to the factor $(d-1)^n$ (resulting from the triangle inequality), not the exponent $n/2$.

It is indeed possible to analyze this situation, and the technique used is very similar to the one leading to the proof of the average Sato-Tate law (Example 4), exploiting to the full the formalism of general algebraic sums. The first – and crucial – point is that, if we fix d and n and a prime $p \nmid d$, the set of relevant Deligne polynomials with coefficients in any extension field $\mathbf{F}_q/\mathbf{F}_p$ is *itself* the set of $\mathbf{F}_q$-points of some algebraic variety $D(d,n)/\mathbf{F}_p$. Indeed, Deligne polynomials of degree d in $\mathbf{F}_q[X_1,\dots,X_n]$ are in one-to-one correspondence with the tuples of coefficients $(\alpha_M)_M$ where M runs over the set of monomials $M = X_1^{m_1}\cdots X_n^{m_n}$ in n variables which are of degree $\leqslant d$, and those are subject only to the condition that the sum

$$\sum_{\deg(M)=d} \alpha_M M$$

defines a smooth projective hypersurface, which is a condition which can be expressed using finitely many polynomial conditions (vanishing or non-vanishing) among the coefficients α_M with $\deg(M) = d$.

Example 22 (Quadratic Deligne polynomials). Consider $n = d = 2$; then $D(2,2) \simeq U(2,2) \times \mathbf{A}^3$ where $U(2,2) \subset \mathbf{A}^3$ with coordinates A, B, C is defined by the discriminant equation

$$AC - B^2 \neq 0,$$

which can be represented as a subset of $\mathbf{A}^4$ with coordinates A, B, C, D defined by

$$D(AC - B^2) = 1.$$

Coming back to general Deligne-type sums, this particular structure of the family of Deligne polynomials can be used to prove that the assignments

$$\Lambda \begin{cases} D(n,d)(\mathbf{F}_q) & \xrightarrow{S} & k \\ F & \mapsto & S(\mathbf{A}^n, \Lambda_F; q) \end{cases}$$

(where k is a suitable extension of $\mathbf{Q}_\ell$, for $\ell \neq p$, containing the p-th roots of unity) are suitable summands for algebraic sums over $D(d,e)(\mathbf{F}_q)$, or in other words, there exists a lisse sheaf

$$\rho_{n,d} : \pi_1(D(n,d)) \to GL((d-1)^n, k)$$

such that

$$\operatorname{Tr} \rho_{n,d}(\operatorname{Fr}_{F,q}) = \sum_{x_1,\dots,x_n \in \mathbf{F}_q} e\Big(\frac{\operatorname{Tr}_{\mathbf{F}_q/\mathbf{F}_p} F(x)}{p}\Big) = q^{n/2} \sum_{j=1}^{(d-1)^n} e^{2i\pi\theta_{q,j}(F)}$$

for any q and $F \in D(n,d)(\mathbf{F}_q)$.

This is quite a general feature, and N. Katz has exploited such constructions in a virtuosic way in many works (see, e.g., [14], [17], [12], or his book with P. Sarnak [18]) to study distribution properties of families of algebraic sums and their associated L-functions. These constructions are again quite difficult, and depend on the general framework of algebraic geometry in a rather sophisticated way.

Example 23 (Kloosterman sheaf). Consider (once more) Hyper-Kloosterman sums

$$HK(n,a;q) = \sum_{\substack{x_1,\dots,x_n \in \mathbf{G}_m(\mathbf{F}_q) \\ x_1 \cdots x_n = a}} e\Big(\frac{\operatorname{Tr}_{\mathbf{F}_q/\mathbf{F}_p}(x_1 + \cdots + x_n)}{p}\Big).$$

Here, Deligne (and Katz) showed that, for every prime p, there exists a lisse sheaf (now called a Kloosterman sheaf) on $\mathbf{G}_m/\mathbf{F}_p$, which we denote $\mathcal{HK}(n;p)$, with the property that

$$\operatorname{Tr}(\operatorname{Fr}_{a,q} \mid \mathcal{HK}(n;p)) = \frac{HK(n,a;q)}{q^{(n-1)/2}} \tag{22}$$

for all $q = p^\nu$ and $a \in \mathbf{G}_m(\mathbf{F}_q)$. For $n = 2$, this means that the wild surmise (10) in Example 10 is indeed correct.

We now consider an arbitrary algebraic sum, and we look at the distribution of the conjugacy classes $\rho(\operatorname{Fr}_{x,q})$, for $x \in V(\mathbf{F}_q)$. Those lie in $GL(r,k)$, but since their eigenvalues are algebraic numbers (being q-Weil numbers), there exists matrices in $GL(r,\mathbf{C})$ with the same eigenvalues, which we denote $\rho(\operatorname{Fr}_{x,q})^{\mathbf{C}}$ temporarily. If ρ is of weight m, the conjugacy class

$$\Theta(\operatorname{Fr}_{x,q}) = q^{-m/2} \rho(\operatorname{Fr}_{x,q})^{\mathbf{C}}, \tag{23}$$

can now be interpreted as a conjugacy class in the unitary group $U(r,\mathbf{C})$ (indeed, all its eigenvalues are of modulus 1). In the example above, this is just the conjugacy class with eigenvalues $e^{2i\pi\theta_{q,j}(F)}$.

Deligne's Equidistribution Theorem does two things: first, it shows that these conjugacy classes *always* satisfy *some* form of equidistribution statement similar to

the average Sato-Tate law (in the vertical direction for $x \in V(\mathbf{F}_{q^\nu})$ where $\nu \to +\infty$, at least, and quite often also for "horizontal" limits); second, it gives an interpretation of the precise shape of this law. This description is in terms of the algebraic data, and is another illustration of the "expressiveness" of the group-theoretic framework (compare with the discussion in Example 20).

This second step is important, because it leads in fact to a much quicker proof, by suggesting the right set of test functions for an application of the Weyl Criterion for equidistribution. In principle, it is also very simple: intuitively, the statement is that the limiting measure is the natural probability measure on the set of conjugacy classes of the smallest *compact group* for which such a statement is possible.

To give a precise formulation requires some care because of technical issues, related partly to the problem of apparent incompatibility of the topology on $\pi_1(V)$ (in which Frobenius classes lie) and the unitary groups $U(r, \mathbf{C})$ (in which we aim to get some equidistribution), and partly to the fact that the unitary conjugacy class $\Theta(\mathrm{Fr}_{x,q})$ might be "too big": it might not reflect fully the Frobenius conjugacy class within $\pi_1(V)$.

This explains the slightly awkward statement of the following preliminary (imperfect) version of Deligne's theorem:

Theorem 24 (Deligne). *Let $V/\mathbf{F}_q$ be a smooth, connected, geometrically irreducible, algebraic variety, and let ρ be a lisse sheaf on V of weight m. For $\nu \geqslant 1$, $x \in V(\mathbf{F}_{q^\nu})$, let $\Theta(\mathrm{Fr}_{x,q^\nu}) = q^{-\nu m/2}\rho(\mathrm{Fr}_{x,q^\nu})^{\mathbf{C}}$ be as before.*

Then there exists a compact subgroup K of $U(r, \mathbf{C})$ such that, for every $\nu \geqslant 1$ and $x \in V(\mathbf{F}_{q^\nu})$, the unitary conjugacy class $\Theta(\mathrm{Fr}_{x,q^\nu})$ intersects K, and the conjugacy class of $\Theta(\mathrm{Fr}_{x,q^\nu})$ in K is uniquely defined, and such that, moreover, as $\nu \to +\infty$, the finite sets given by

$$\begin{cases} V(\mathbf{F}_{q^\nu}) & \longrightarrow & K^\sharp \\ x & \mapsto & \Theta(\mathrm{Fr}_{x,q^\nu}) \end{cases}$$

become equidistributed in $K^\sharp$ with respect to the natural probability measure, where $K^\sharp$ is the space of conjugacy classes of K.

Before giving some clarifying remarks and stating another (better) version of this theorem, we recall the definition of the limiting measure: for any compact group K, there exists a unique Borel measure μ_K (the probability Haar measure) on K which is normalized by $\mu_K(K) = 1$ and is *translation invariant*:

$$\int_K f(xy)d\mu_K(x) = \int_K f(yx)d\mu_K(x) = \int_K f(x)d\mu_K(x)$$

for any $f \in L^1(K)$. Then, the set $K^\sharp$ of conjugacy classes in K inherits a probability measure $\mu_K^\sharp$ through the quotient map $K \longrightarrow K^\sharp$; concretely, we have

$$\int_{K^\sharp} f(\theta)d\mu_K^\sharp = \int_K f(x)d\mu_K(x)$$

if $f : K \to \mathbf{C}$ is invariant under conjugation (e.g., if f is a symmetric function of the r eigenvalues, for $K = U(r, \mathbf{C})$).

Example 25 (Group-theoretic interpretation of the Sato-Tate law). In the case of the average Sato-Tate law, the relevant group is $SU(2, \mathbf{C})$, the special unitary group of size 2. The conjugacy classes in this group are all given by

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

with $\theta \in \mathbf{R}$, and if we restrict to $\theta \in [0, \pi]$, then each conjugacy class is represented uniquely (because the matrices corresponding to θ and $-\theta$ are conjugate in $SU(2, \mathbf{C})$). Note that the conjugacy classes in $U(2, \mathbf{C})$ associated to Kloosterman sums are indeed of this type (see (8)).

Since twice the cosine function is a bijection $[0, \pi] \xrightarrow{2\cos} [-2, 2]$, we can also represent the set of conjugacy classes by the closed interval $[-2, 2]$. It turns out that,[15] in terms of the coordinate θ, the measure $\mu^\sharp_{SU(2)}$ is given by

$$\int_{SU(2)^\sharp} f(\theta) d\mu^\sharp_{SU(2)} = \frac{2}{\pi} \int_0^\pi f(\theta)(\sin^2 \theta) d\theta.$$

This is the group-theoretic explanation of the Sato-Tate measure. In particular, if we compare with the normalized Lebesgue measure, the factor $\sin^2 \theta$ reflects the non-commutativity of the group $SU(2)$.

As stated, Theorem 24 is still unsatisfactory, mainly because it does not provide an a priori description of the group K starting merely from the algebraic data involved (the algebraic variety V and the lisse sheaf ρ), and especially does not give any hint of the way this group might be computed for a particular case (e.g., for the proof of average Sato-Tate conjecture).

We will now explain how this is remedied. Although the reformulation is more abstract, the outcome well repays the effort that may be involved. The basic observation is that since ρ is a homomorphism

$$\rho : \pi_1(V) \longrightarrow GL(r, k),$$

the most obvious group to look in for equidistribution of Frobenius classes (which we know are dense in $\pi_1(V)$) is the image group $\rho(\pi_1(V)) \subset GL(r, k)$. However, this group of matrices with entries in k is on the other side of the mirror from the unitary group $U(r, \mathbf{C})$, and we can not even hope to map to $\mathbf{C}$ in a reasonable way using a chosen embedding $\iota : \bar{k} \hookrightarrow \mathbf{C}$ (since ι is not continuous, the image $\iota(\rho(\pi_1(V)))$ will not be closed or compact or anything).

However, we can observe that ι (being a field homomorphism) does allow *algebraic relations* to transfer. So we can consider the collection I_ρ of all polynomial relations (with coefficients in $\bar{k}$) valid on $\rho(\pi_1(V))$ – the variables being the entries of the matrices, and the inverse of the determinant –; then $\iota(I_\rho)$ is a set of polynomials with coefficients in $\mathbf{C}$, and we can define a group, called the *arithmetic monodromy group*[16] of the lisse sheaf ρ, as the set of $g \in GL(r, \mathbf{C})$ for which all relations in

[15] This is a special case of the so-called Weyl Integration formula.

[16] Properly speaking, relative to ι; in fact, most texts define it as the subgroup of $GL(r, \bar{k})$ where all relations hold, and only map later to $\mathbf{C}$.

$\iota(I_\rho)$ are valid. We denote this group $G^{\mathrm{arith}}(\rho)$. Similarly, the *geometric monodromy group* of ρ, denoted $G^{\mathrm{geom}}(\rho)$, is the subgroup of $G^{\mathrm{arith}}(\rho)$ of matrices satisfying all relations in $\iota(I_\rho^{\mathrm{geom}})$, where I_ρ^{geom} is the set of polynomial relations satisfied by the subgroup $\rho(\pi_1(\bar{V})) \subset \rho(\pi_1(V))$.

Example 26. (1) Suppose $\det(\rho(g)) = 1$ for all $g \in \pi_1(V)$. This is an algebraic relation, and hence (since $\iota(1) = 1$) we have $G^{\mathrm{arith}}(\rho) \subset SL(r, \mathbf{C})$. Similarly, if we have $\det(\rho(g))^h = 1$ for all $g \in \pi_1(V)$, for some fixed $h \geqslant 1$, we have

$$G^{\mathrm{arith}}(\rho) \subset \{g \in GL(r, \mathbf{C}) \mid \det(g)^h = 1\}.$$

(2) Suppose $r = 2g$ is even and there is a non-degenerate alternating form $\langle \cdot, \cdot \rangle$ on k^{2g} such that the image of ρ lies in the group

$$Sp(\langle \cdot, \cdot \rangle, k) = \{g \in GL(r, k) \mid \langle g(v), g(w) \rangle = \langle v, w \rangle \text{ for all } v, w \in k^{2g}\}$$

of symplectic automorphisms of k^{2g} with respect to this pairing. Then, since these conditions can be phrased polynomially, one can transfer the pairing by ι to a non-degenerate alternating form on $\mathbf{C}^{2g}$ such that the arithmetic (and geometric) monodromy group is a subgroup of $Sp(2g, \mathbf{C})$ with respect to this pairing. Similarly for a symmetric pairing. And because all alternating (resp. symmetric) bilinear forms on $\mathbf{C}$ are equivalent, up to to a linear change of variable, it is usual in this situation to omit specific mention of the bilinear form.

The operation just performed[17] is summarized in the language of algebraic geometry by saying that $G^{\mathrm{arith}}(\rho)$ (resp. $G^{\mathrm{geom}}(\rho)$) is the image under ι of the *Zariski closure*[18] of $\rho(\pi_1(V))$ (resp. $\rho(\pi_1(\bar{V}))$); topologically, one exploits the fact that ι is continuous for the (rather weak) Zariski topology.

This abstract definition turns out to be remarkably useful. If it looks unfamiliar, it is important to realize that the step of taking the Zariski closure is *immensely simplifying*: because $G^{\mathrm{geom}}(\rho)$, by definition, is defined by polynomial equations as a subgroup of $GL(r, \mathbf{C})$, it is a much more rigid object than an arbitrary subgroup. This translates into the fact that there are fewer possibilities, and hence that it is often possible to classify the possible choices of $G^{\mathrm{geom}}(\rho)$ and compute it.

We will now present a better version of Deligne's Equidistribution Theorem. This will introduce the further restrictions on ρ that it be of weight $m = 0$, and that $\iota(\rho(\pi_1(V))) \subset G^{\mathrm{geom}}(\rho)$: these amount to a suitable normalization of the Frobenius

[17] Its relevance was essentially realized by Grothendieck, and it was of course exploited masterfully by Deligne, though the thought process involved was certainly different (historians will notice that [6, 1.1.15] defines the monodromy group to be $\rho(\pi_1(\bar{V}))$, and does not give a name to the Zariski closure, though it is used extensively from [6, 1.3.7] onward.)

[18] The Zariski closure D^{Zar} of any subset $D \subset k^n$, where k is algebraically closed, is the set of points in k^n which satisfy the same polynomial equations as the points of D:

$$D^{Zar} = \bigcap_{F \in I(D)} \{y \in k^n \mid F(y) = 0\},$$

where

$$I(D) = \{F \in k[X_1, \ldots, X_n] \mid F(x) = 0 \text{ for all } x \in D\}.$$

conjugacy classes (in other words, they can be achieved by a proper algebraic analogue of the division by $q^{m/2}$ in the definition of the naive unitary classes (23); for Kloosterman sums, this is the step that passes from the sheaf $\mathcal{HK}_1$ of weight 1 – see (11) –, to the actual Kloosterman sheaf $\mathcal{HK}$, which is of weight 0).

Under these conditions, the subgroup K can be identified as a *maximal compact subgroup*[19] of $G^{\mathrm{geom}}(\rho)$; the apparent ambiguity is resolved by the fact that, for the groups that occur, all such maximal compact subgroups are conjugate in $G^{\mathrm{geom}}(\rho)$, hence have "identical" spaces of conjugacy classes, which is where the Frobenius conjugacy classes lie.

Example 27 ("Big" symmetry groups). In practice, the following three groups are the most important: in many contexts, the group $G^{\mathrm{geom}}(\rho)$ turns out to be either one of them, or to contain one of them as a subgroup with finite index. In Example 33 below, we will give concrete instances of families leading to these groups.

1. The group $SL(r)$, for some $r \geqslant 1$, has (in its incarnation over $\mathbf{C}$) maximal compact subgroup $SU(r, \mathbf{C})$. It is, essentially, the largest possible subgroup of $U(r, \mathbf{C})$ that can occur as $G^{\mathrm{geom}}(\rho)$ for ρ of rank r. This is because there are general structural properties (due to Grothendieck and Deligne), which imply that the connected component of the identity G^0 in $G^{\mathrm{geom}}(\rho)$ is semisimple, under fairly general conditions ([6, Cor. 1.3.9]) – this means that it does not contain any connected abelian non-trivial normal subgroup, in particular the center is finite. It follows that if $G^{\mathrm{geom}}(\rho)$ *contains* $SL(r)$, it is of the form
$$\{g \in GL(r) \mid \det(g)^h = 1\}$$
for some integer $h \geqslant 1$, with maximal compact subgroup $U_h(r, \mathbf{C})$ consisting of unitary matrices with determinant an h-th root of unity.
2. The group $Sp(2r, \mathbf{C})$, defined as the group of elements in $GL(2r, \mathbf{C})$ preserving a given non-degenerate alternating bilinear form $\langle \cdot, \cdot \rangle$, has maximal compact subgroup $USp(2r, \mathbf{C}) = U(2r, \mathbf{C}) \cap Sp(2r, \mathbf{C})$. Concretely, a conjugacy class in $USp(2r, \mathbf{C})$ is determined by its reversed characteristic polynomial $\det(1 - Tg) \in \mathbf{C}[T]$, which is a polynomial with $2r$ roots which can be put into r pairs of inverses $(e^{i\theta_j}, e^{-i\theta_j})$, with $\theta_j \in [0, \pi]$, $1 \leqslant j \leqslant r$. This group occurs typically for families of L-functions of algebraic curves over finite fields (see Example 16); as we saw, the pairing of roots can be seen as a reflection of the functional equation of the L-function. Geometrically, it is known that there exists a non-degenerate alternating bilinear form on $H^1_c(\bar{C})$, which is preserved by the unitarized Frobenius.
3. The group $O(r, \mathbf{C})$, or its subgroup $SO(r, \mathbf{C})$ of index 2, is defined as the elements in $GL(r, \mathbf{C})$ (resp. $SL(r, \mathbf{C})$) which preserve a non-degenerate symmetric bilinear form. Its maximal compact subgroup is $O(r, \mathbf{R})$ (resp. $SO(r, \mathbf{R})$), the corresponding real groups. Note that for r even, the eigenvalues come again in r pairs of inverses, so a conjugacy class in $SO(2r, \mathbf{R})$ looks exactly like one in $USp(2r, \mathbf{C})$; however, the measures on $SO(2r, \mathbf{R})^\sharp$ and $USp(2r, \mathbf{C})^\sharp$ are distinct

[19] Maximal with respect to inclusion.

(see, e.g., [18, 5.0.4, 5.0.6] for the Weyl formula which shows this). The underlying orthogonal or symplectic symmetry of a sheaf is therefore not immediately visible by looking simply at the Frobenius conjugacy classes.

Orthogonal (and special orthogonal) symmetry is found, for instance, in families of elliptic curves over function fields.

A basic guideline is that, unless and until one has reason to think otherwise, one should expect that $G^{\mathrm{geom}}(\rho)$ will be of one of those three types. In Example 31, we will explain some recent ideas of Larsen and Katz that give quite simple criteria to (essentially) check if this holds, in very concrete (in fact, numerically testable) arithmetic ways.

With all this done, we can now state Deligne's equidistribution theorem:

Theorem 28 (Deligne). *Suppose $V/\mathbf{F}_q$ is smooth and geometrically connected,*[20] *and assume ρ is a lisse sheaf of weight 0 on V with the property that, for the fixed $\iota : \bar{k} \hookrightarrow \mathbf{C}$ as above, we have $\iota(\rho(\pi_1(V))) \subset G^{\mathrm{geom}}(\rho)$. Let K be a maximal compact subgroup of $G^{\mathrm{geom}}(\rho)$. For $\nu \geqslant 1$ and $x \in V(\mathbf{F}_{q^\nu})$, write*

$$\iota(\mathrm{Fr}_{x,q^\nu}) = \Theta_{x,q^\nu} U_{x,q^\nu} \tag{24}$$

where Θ_{x,q^ν} is diagonalizable, U_{x,q^ν} is unipotent,[21] *and they commute.*

Then, as ν tends to infinity, the finite sets given by

$$\begin{cases} V(\mathbf{F}_{q^\nu}) & \longrightarrow & K^\sharp \\ x & \mapsto & \Theta_{x,q^\nu} \end{cases}$$

become equidistributed in the space $K^\sharp$ of conjugacy classes of K, with respect to its natural probability measure.

This is a slightly simplified variant of [18, Th. 9.2.6]; it is not difficult to recover the preliminary statement of Theorem 24 from it, but it is probably best to simply forget that earlier version.

Remark 29 (Diagonalizability). It is conjectured that (in most cases at least) $\rho(\mathrm{Fr}_{x,q^\nu})$ is always diagonalizable (or semisimple, as the proper terminology has it), so that the diagonalization step (24) is not necessary. This is the case for curves (due to Weil), and also holds (for a given x) whenever the eigenvalues of $\rho(\mathrm{Fr}_{x,q^\nu})$ are all distinct.

This theorem is a powerful confirmation that the viewpoint on algebraic sums associated with representations of the group $\pi_1(V)$ is correct: it provides a clear explanation of the fact (which is not at all obvious and was partly discovered empirically) that *whenever there is some equidistribution in a family of exponential sums, the limiting measure is associated to some group K*. Note that such a direct explanation is not yet available for other conjectured equidistribution results (see the last section for examples).

[20] These are just for simplicity.

[21] This is the Jordan decomposition; it is known that Θ_{x,q^ν} and U_{x,q^ν} are in $G^{\mathrm{geom}}(\rho)$, and that Θ_{x,q^ν} is conjugate to an element of K.

Moreover, the proof is quite straightforwardly based on the Riemann Hypothesis and the general formalism of algebraic sums. Because of this it is highly effective and uniform, and this allows other variants of Deligne's Theorem to be proved. We sketch the argument, because its underlying simplicity is the best justification for the preparatory definitions (and it illuminates the necessity of the specific normalizing assumption on ρ).

Sketch of the proof. We use the Weyl Criterion and consider, as basis of functions on $K^\sharp$ with mean zero, the functions

$$g \mapsto \operatorname{Tr} \pi(g)$$

where $\pi : K \to GL(n_\pi, \mathbf{C})$ runs over the non-trivial continuous, irreducible representations of K (this means that π is a continuous homomorphism, not $= 1$, and that we have $\int_K |\operatorname{Tr} \pi(g)|^2 d\mu_K = 1$).

The main fact one needs to know is that such a representation π corresponds uniquely to an algebraic[22] (non-trivial) irreducible representation

$$G^{\mathrm{geom}}(\rho) \longrightarrow GL(n_\pi, \mathbf{C}),$$

and the latter (via ι^{-1}) to an algebraic irreducible representation

$$\iota^{-1} G^{\mathrm{geom}}(\rho) \xrightarrow{\tilde{\pi}} GL(n_\pi, \bar{k}),$$

in such a way that

$$\operatorname{Tr} \tilde{\pi}(\iota^{-1}(g)) = \operatorname{Tr} \pi(g), \qquad \text{if } g \in K \subset G^{\mathrm{geom}}(\rho)$$

(this correspondence is due essentially to H. Weyl, and is often called the *unitary trick*; it has to do with the fact that, for π as above, the trace is a symmetric, *polynomial function*, of the eigenvalues of $g \in K$: see (9) for an illustration).

This relation applies in particular to the conjugacy classes Θ_{x,q^ν} (*because* of the assumption $\iota(\rho(\pi_1(V))) \subset G^{\mathrm{geom}}(\rho)$), and therefore we find the basic formula

$$\sum_{x \in V(\mathbf{F}_{q^\nu})} \operatorname{Tr} \pi(\Theta_{x,q^\nu}) = \sum_{x \in V(\mathbf{F}_{q^\nu})} \operatorname{Tr}(\tilde{\pi} \circ \rho)(\mathrm{Fr}_{x,q^\nu}).$$

One then shows that $\tilde{\pi} \circ \rho$ is itself a lisse sheaf on V of weight 0; consequently, we can apply to the right-hand side the trace formula and then Deligne's Riemann Hypothesis to get the estimate

$$\sum_{x \in V(\mathbf{F}_{q^\nu})} \operatorname{Tr} \pi(\Theta_{x,q^\nu}) = \operatorname{Tr}(F^\nu \mid H_c^{2d}(\bar{V}, \tilde{\pi} \circ \rho)) + O(q^{\nu(d-1/2)})$$

for $\nu \geqslant 1$ (the implied constant is the sum of dimensions of the lower-index cohomology groups; compare with (20)). The idea for not looking beyond the topmost index is that we know that

$$|V(\mathbf{F}_{q^\nu}))| = q^{d\nu} + O(q^{\nu(d-1/2)})$$

[22] Where *algebraic* means that the coefficients of the matrices representing these homomorphisms are polynomials.

(because V is geometrically connected, as explained in (18)) and thus we obtain

$$\frac{1}{|V(\mathbf{F}_{q^\nu})|}\sum_{x\in V(\mathbf{F}_{q^\nu})}\operatorname{Tr}\pi(\Theta(\mathrm{Fr}_{x,q^\nu}))=\frac{\operatorname{Tr}(F^\nu\mid H_c^{2d}(\bar{V},\pi\circ\rho))}{q^{d\nu}}+o(1)$$

as $\nu\to+\infty$. To go further, we use the formula (17):

$$\dim_k H_c^{2d}(\bar{V},\tilde{\pi}\circ\rho)=\dim_k(\tilde{\pi}\circ\rho)_{\pi_1(\bar{V})}=(\tilde{\pi})_{\rho(\pi_1(\bar{V}))}.$$

Now we use the following remark: since $\tilde{\pi}$ is given by *polynomial* formulas, the coinvariant quotient of $\bar{k}^{n_\pi}$ for the action of $\rho(\pi_1(\bar{V}))$ is *the same* as the coinvariants for the action of its Zariski closure (or "the same" as those for $G^{\mathrm{geom}}(\rho)$ on $\mathbf{C}^{n_\pi}$). But since $\tilde{\pi}$ is irreducible and non-trivial, this coinvariant space is 0. So the cohomology group $H_c^{2d}(\bar{V},\tilde{\pi}\circ\rho)$ is in fact 0 (compare Example 20) and we deduce

$$\lim_{\nu\to+\infty}\frac{1}{|V(\mathbf{F}_{q^\nu})|}\sum_{x\in V(\mathbf{F}_{q^\nu})}\operatorname{Tr}\pi(\Theta(\mathrm{Fr}_{x,q^\nu}))=0,$$

as desired. □

Example 30. (1) Analytic number theorists should compare the proof with that of Dirichlet's Theorem on primes in arithmetic progressions: the basic strategy is identical, including the important point that one needs to use the "right harmonics" for the job.

(2) One goes from Deligne's Theorem to the average Sato-Tate law in two steps. One is fairly easy: from the identification of the Haar measure on $SU(2)^\sharp$ as the Sato-Tate measure, we see that to prove the "vertical direction" (with $a\in\mathbf{F}_{p^\nu}$ with $\nu\to+\infty$), it is sufficient to show that $K=SU(2)$ for the Kloosterman sheaf $\mathcal{HK}(2;p)$ of Example 10. In turn, this means that we must show that $G=G^{\mathrm{geom}}(\mathcal{HK}(2;p))$ is equal to $SL(2)$. Now, this monodromy group is given a priori as an algebraic subgroup of $GL(2)$, since $\mathcal{HK}(2;p)$ is of rank 2, and as stated in Example 27, its connected component of the identity is semisimple. One can show that these conditions only leave the possibilities $G\supset SL(2)$, or G finite. One can show that the second alternative does not hold,[23] and moreover check that the determinant of this Kloosterman sheaf is trivial (because the product of the two eigenvalues is 1), so that $G=SL(2)$, as desired.

Dealing with the horizontal direction of the average Sato-Tate law (as we stated it originally in Example 4) requires a more careful proof of Deligne's Theorem, leading to a uniform version over varying primes. The point is that, once we have computed that the geometric monodromy group of $\mathcal{HK}(2;p)$ is $SL(2)$ for *all* (odd) primes p, we can reproduce the argument in the proof while keeping track of the dependency on p: for any $m\geqslant 1$, one gets

$$\frac{1}{p-1}\Big|\sum_{a\in\mathbf{F}_p^\times}\operatorname{Tr}(\mathrm{Sym}^m\circ\mathcal{HK})(\mathrm{Fr}_{a,p})\Big|\leqslant\frac{\sqrt{p}}{p-1}\times(h^0(p)+h^1(p))$$

[23] This is not obvious, of course, but roughly it would imply that the Kloosterman sums are "much simpler than expected", and satisfy unrealistic properties – which can indeed be disproved.

where

$$h^i(p) = \dim H^i_c(\bar{\mathbf{G}}_m, \mathrm{Sym}^m \circ \mathcal{HK}(2;p))$$

(the main term is dealt with, uniformly, because the monodromy group turned out to be independent of p). The main point is the dependency on p, which is a non-trivial issue because the sheaves involved depend on p. We refer to [14, Ch. 11, 13] for statements bounding the dimensions[24] of these cohomology spaces (valid in the greater generality of families of Hyper-Kloosterman sums).

6. Some recent applications

We present here a few fairly recent works involving Deligne's Equidistribution theorem and the Riemann Hypothesis.

Example 31 (The Larsen alternative). The main step in applying successfully Deligne's Equidistribution Theorem is often the computation (if possible!) of the geometric monodromy group. Thanks to fairly recent developments, there are now very concrete criteria to do this in some cases.

Consider a lisse sheaf ρ of rank r satisfying the assumptions of Theorem 28, and let $G = G^{\mathrm{geom}}(\rho)$. For integers $k \geqslant 1$, define

$$M_k(\rho,\nu) = \frac{1}{|V(\mathbf{F}_{q^\nu})|} \sum_{x \in V(\mathbf{F}_{q^\nu})} |\operatorname{Tr} \rho(\mathrm{Fr}_{x,q^\nu})|^k,$$

$$M_k(\rho) = \lim_{\nu \to +\infty} M_k(\rho,\nu).$$

By Deligne's Theorem and the definition of equidistribution, we know that the limit exists and is given by the average

$$M_k(\rho) = \int_K |\operatorname{Tr}(\theta)|^k d\mu_K(\theta),$$

in other words, it contains some basic information on the geometric monodromy group, through the subgroup K.

Larsen's Alternative is the following remarkable statement:

Theorem 32 (The Larsen Alternative). *Suppose one knows that $G = G^{\mathrm{geom}}(\rho)$ is* infinite. *Then the following holds:*

(1) *If $M_4(\rho) = 2$, then $SL(r) \subset G$, and $SU(r,\mathbf{C}) \subset K$.*

(2) *If $r = 2g$ is even, $r \geqslant 3$, and there exists a non-degenerate alternating pairing with respect to which $G \subset Sp(2g)$, and if $M_4(\rho) = 3$, then $G = Sp(2g)$ and $K = USp(2g,\mathbf{C})$.*

(3) *If $r \geqslant 3$ and there exists a non-degenerate symmetric pairing with respect to which $G \subset O(r)$, and if $M_4(\rho) = 3$, then $SO(r) \subset G$ and $SO(r,\mathbf{C}) \subset K$.*

[24] Dimensions of cohomology spaces are commonly called Betti numbers.

In other words, if $G^{\mathrm{geom}}(\rho)$ is not finite, one can check that it is one of the three "big" monodromy groups of Example 27 by "simply" computing $M_4(\rho)$. The point of this is that, if ρ was defined so that the summands $\mathrm{Tr}(\rho(\mathrm{Fr}_{x,q^\nu}))$ are *themselves* concrete exponential sums, it is sometimes possible to use this expression to perform this computation by the standard analytic tool of expanding the fourth power, and exchanging the order of summation.

It is equally remarkable that the proof of Theorem 32 is, in fact, not very difficult (see [13], and see [12] for very general contexts in which the computation of M_4 is possible).

Consider, as an example, the average Sato-Tate conjecture again. For $\rho = \mathcal{HK}(2;p)$, we find that by definition (using the fact that the Kloosterman sums are real) that

$$M_4(\rho,\nu) = \frac{1}{p^\nu-1}\frac{1}{p^{2\nu}}\sum_{a\in\mathbf{F}_{p^\nu}^\times}\Big(\sum_{x\in\mathbf{F}_{p^\nu}^\times} e\Big(\frac{\mathrm{Tr}(x+ax^{-1})}{p}\Big)\Big)^4$$

and it is a classical computation (due, in fact, to Kloosterman, see, e.g., [10, §4.4]) that

$$M_4(\rho,\nu) = \frac{2p^{3\nu}-3p^{2\nu}-p^\nu-1}{p^{3\nu}-p^\nu} \longrightarrow 2 = M_4(\rho),$$

confirming that G is *either* finite, or contains $SL(2,\mathbf{C})$.

Furthermore, part of the beauty of the Larsen alternative is that it is quite amenable to numerical check: in many cases, one can compute an "empirical" fourth moment $M_4(\rho,\nu)$ for a given family of (say) exponential sums and ν small; if this is found to be close to the expected value (2 or 3), there is strong reasons to believe that this identifies the relevant monodromy group (up to the finite indeterminacy noticed above). Of course, proving that this is so might be more difficult...

Example 33 (Symmetry examples). We give examples of families with each of the three "big" symmetry types (Example 27). These (and similar) examples are related to the conjectures relating L-functions and Random Matrix Theory, as explained in detail in [18]. In each case, we look for concreteness and simplicity, and select examples of 1-parameter families: there are many more (and more general) examples known!

– [Unitary symmetry] Examples of unitary monodromy are given by some of the Kloosterman sheaves $\mathcal{HK}(n;p)$ of rank n (Example 23). Indeed, Katz [14, 11.1] proved that the corresponding geometric monodromy group is $SL(n)$ if n and p are both odd. Concretely, by general facts about equidistribution, this means in particular (after the horizontality is taken care of, as Katz does) that

$$\Big\{\frac{HK(n;a,p)}{p^{(n-1)/2}} \mid a\in\mathbf{F}_p^\times\Big\}$$

becomes equidistributed on $[-n,n]$ with respect to the image by the trace $\mathrm{Tr} : SU(n,\mathbf{C}) \longrightarrow [-n,n]$ of the Haar measure of $SU(n,\mathbf{C})$.

– [Symplectic symmetry] Symplectic monodromy occurs in families of algebraic curves. For example, let q be odd and let $g \geqslant 1$ be given. Fix a polynomial $f \in \mathbf{F}_q[X]$ which is monic, squarefree, and of degree $2g$. Then consider the algebraic curves with equation

$$C_t \,:\, y^2 = f(x)(x-t)$$

where t is a parameter which is not a zero of f. This condition defines an algebraic variety $U/\mathbf{F}_q$ (the complement of the zeros of f; it is smooth, connected, geometrically irreducible). For each $\nu \geqslant 1$ and $t \in U(\mathbf{F}_{q^\nu})$, we obtain a smooth projective algebraic curve $\tilde{C}_t/\mathbf{F}_{q^\nu}$ of genus g (by adding a point at infinity to C_t), with an L-function as in Example 16. From the general machinery, Katz and Sarnak [18, §10.1] show that there exists a lisse sheaf $\rho_{f,1}$ of weight 1 on $U/\mathbf{F}_q$ such that

$$\det(1 - T\rho_{f,1}(\mathrm{Fr}_{t,\mathbf{F}_{q^\nu}})) = L(C_t, T) = \det(1 - TF \mid H^1(\bar{\tilde{C}}_t, \mathrm{trivial})).$$

The functional equation of the L-functions reflects the fact that there exists a non-degenerate alternating pairing on the cohomology group for which F acts as a symplectic similitude; after renormalizing, they obtain a sheaf ρ_f of weight 0 such that

$$\det(1 - T\rho_f(\mathrm{Fr}_{t,\mathbf{F}_{q^\nu}})) = L(C_t, q^{-\nu/2}T),$$

and then they show that $\rho_f(\pi_1(\bar{U})) \subset Sp(2g)$, and indeed they prove that

$$G^{\mathrm{geom}}(\rho_f) = Sp(2g). \tag{25}$$

As a corollary, for instance, one gets

$$\frac{1}{|U(\mathbf{F}_{q^\nu})|} \sum_{t \in U(\mathbf{F}_{q^\nu})} \det(1 - \rho_f(\mathrm{Fr}_{t,\mathbf{F}_{q^\nu}}))^k \longrightarrow \int_{USp(2g,\mathbf{C})} \det(1-g)^k d\mu(g)$$
$$= \prod_{j=1}^{k} \frac{1}{(2j-1)!!}, \tag{26}$$

as $\nu \to +\infty$, for any fixed integer $k \geqslant 1$ (the last formula being a result of Keating and Snaith; recall that $(2j-1)!! = 1 \cdot 3 \cdots (2j-3) \cdot (2j-1)$).

– [Orthogonal symmetry] The simplest examples of orthogonal symmetry are given by twists of elliptic curves over functions fields. The basic theory, which we illustrate here, is again due to Katz [17] (there is also a short survey in [20]).

For any odd prime power $q \geqslant 3$, any integer $d \geqslant 1$, consider the elliptic curves over the field $\mathbf{F}_q(t)$ which are given by the Weierstrass equation

$$E_z \,:\, y^2 = (t^d - dt - 1 - z)x(x+1)(x+t)$$

where $z \in \mathbf{F}_q$ is a parameter such that z is not a critical value of $t^d - dt - 1$, i.e., not a value of this polynomial at a root of the derivative; again this condition defines a parameter algebraic variety $U/\mathbf{F}_q$ (which depends on d).

Katz shows that there exists a lisse sheaf $\rho_{d,1}$ on $U/\mathbf{F}_q$, of rank $2d$ and weight 2, such that the associated L-function (which is defined by the "standard" Euler

product over prime ideals in $\mathbf{F}_q[t]$, with suitable ramified factors, as for L-functions of elliptic curves over $\mathbf{Q}$) is of the form

$$L(E_z, T) = \det(1 - T\rho_{d,1}(\mathrm{Fr}_{z,\mathbf{F}_{q^\nu}})).$$

After normalization, one obtains as before a sheaf ρ_d of weight 0, and the theory provides a non-degenerate symmetric pairing for which ρ_d takes value in $O(2d)$. Then, for $d \geqslant 146$ at least and provided $(p-1, d-1) = 1$ and $p \nmid d(d-1)(d+1)$, Katz proves that $G^{\mathrm{geom}}(\rho_d) = O(2d)$.

As a cautionary tale, here is an example with *finite* monodromy (see [12, Remark 3.8.3] for a few more). Consider $p = 5$ and exponential sums of the type

$$S(f) = \sum_{x \in \mathbf{F}_q} e(\mathrm{Tr}(f(x))/5)$$

where $q = 5^\nu$ and $f \in \mathbf{F}_q[X]$ is monic of degree 3. There is an algebraic variety $D_3/\mathbf{F}_5$ parametrizing the polynomials f, and a sheaf ρ_3 of weight 0 and degree 2 such that

$$\mathrm{Tr}(\rho_3(\mathrm{Fr}_{f,5^\nu})) = \frac{S(f)}{5^{\nu/2}}$$

for $\nu \geqslant 1$ and $f \in D_3(\mathbf{F}_{5^\nu})$. Katz shows that the corresponding $G^{\mathrm{geom}}(\rho_3)$ is a finite group (but has $M_4(\rho_3) = 2$).

Example 34 (Sieve and families of L-functions). A variant of Deligne's Equidistribution Theorem is a very general version of the Chebotarev density theorem. This corresponds to the study of the distribution of $\rho(\mathrm{Fr}_{x,q^\nu})$ for a homomorphism ρ of the type

$$\rho : \pi_1(V) \longrightarrow G,$$

where G is a now an abstract *finite* group (not necessarily seen as a subgroup of a matrix group). In that situation, there is no problem of continuity or difficulty with comparison of ℓ-adic and complex fields (all data involved involves only algebraic numbers). Because of the uniformity and control afforded by the Riemann Hypothesis, which is applied to sums of the type

$$\sum_{x \in V(\mathbf{F}_{q^\nu})} \mathrm{Tr}\, \pi\rho(\mathrm{Fr}_{x,q^\nu}),$$

one can prove very explicit and uniform results (in terms of the group G and even of $V/\mathbf{F}_q$, see, e.g., [20]).

This type of equidistribution results, in turn, can be combined with ideas of sieve theory to study certain arithmetic properties of families of L-functions over finite fields given by

$$\det(1 - T\rho(\mathrm{Fr}_{x,q^\nu}))$$

for some lisse sheaf ρ over $V/\mathbf{F}_q$. Those, in many cases (families of curves, for instance) are *integral* polynomials, and (following a question of Katz that was first solved qualitatively by N. Chavdarov), one may ask, for instance, whether they are irreducible? Another arithmetic question which has attracted some interest is

whether the order of the group of $\mathbf{F}_q$-rational points of the Jacobian is sometimes a prime number (or an almost prime)?

The basic tool is the existence (in some circumstances) of homomorphisms

$$\rho_\ell : \pi_1(V) \longrightarrow GL(r, \mathbf{F}_\ell)$$

for *every* prime $\ell \neq p$, such that

$$\det(1 - T\rho_\ell(\mathrm{Fr}_{x,q^\nu})) \equiv \det(1 - T\rho(\mathrm{Fr}_{x,q^\nu})) \,(\mathrm{mod}\, \ell),$$

for every $x \in V(\mathbf{F}_{q^\nu})$. Controlling the distribution of the Frobenius under ρ_ℓ, with sufficient uniformity with respect to ℓ, and applying various sieve techniques leads to many interesting applications.

We illustrate this with one particular result which is especially concrete; it is found (together with further discussion and applications) in [19, §8], and uses the families of curves in the symplectic symmetry example above.

Theorem 35. *Let $q \neq 1$ be a power of an odd prime. Let $f \in \mathbf{F}_q[X]$ be squarefree of degree $2g$ for some integer $g \geqslant 1$. Consider the family of curves of genus g given by $C_t : y^2 = f(x)(x - t)$, and its L-functions*

$$L(C_t) = P_1(C_t) = \prod_{1 \leqslant j \leqslant 2g} (1 - \alpha_{t,j} T), \qquad \textit{where } |\alpha_{t,j}| = \sqrt{q}.$$

Then

$$|\{t \in \mathbf{F}_q \mid P_1(C_t) \textit{ has ``small'' Galois group}\}| \ll g^2 q^{1-\gamma_g}$$

for some $\gamma_g \approx 1/4g^2$, and some absolute implied constant.

The meaning of "small" is the following: the existence of g pairs of roots α_j, α_k with $\alpha_j \alpha_k = g$ implies that the splitting field of $L(C_t)$ must have Galois group G_t isomorphic to a subgroup of the group W_{2g} of signed permutation matrices of size g (i.e., matrices in $GL(g, \mathbf{Z})$ where there is a single non-zero element in each row and column, and this element is either 1 or -1). To say that $L(C_t)$ has small Galois group means that G_t is a proper subgroup of W_{2g}.

A crucial input to this result is a deep fact, due to J-K. Yu (and recently reproved in greater generality by C. Hall [8]): for the relevant ρ_ℓ, the group $\rho_\ell(\pi_1(\bar{V}))$ is (isomorphic to) the group $Sp(2g, \mathbf{F}_\ell)$ for all $\ell \neq 2, p$. The symplectic nature of the polynomials shows this is as large as it can be. Although this is a finite-level analogue to the computation of geometric monodromy groups (25), this is in fact significantly more difficult: for instance, for every $m \geqslant 1$, the group

$$\{A \in Sp(2g, \mathbf{Z}_\ell) \mid A \equiv 1 \,(\mathrm{mod}\, \ell^m)\}$$

has Zariski closure $Sp(2g)$, but of course is trivial modulo ℓ. So the theorem of Yu must manage to eliminate this type of possibilities.

Example 36 (A problem of harmonic analysis). Here is a very recent example due to Bombieri and Bourgain [1], involving "classical" character sums. We select it (among many applications of the Riemann Hypothesis) to illustrate once more how it sometimes can be applied for problems which seem apparently very remote – maybe

this will help readers feel some of the same surprise which must have surrounded the discovery of the link between exponential sums and algebraic geometry...

In 1980, Kahane had proved the existence of trigometric polynomials P_n of degree n with coefficients of modulus 1, i.e.,

$$P_n(\theta) = \sum_{m=0}^{n} \hat{P}_n(m)e(m\theta), \qquad |\hat{P}_n(m)| = 1,$$

such that

$$|P_n(\theta)| = \sqrt{n} + O(n^{1/2-1/17}\sqrt{\log n}), \quad \text{for all } \theta \in \mathbf{R},$$

thereby confirming a conjecture of Littlewood (and disproving one of Erdős). His methods were probabilistic and did not allow the explicit construction of P_n.

In [1], Bombieri and Bourgain give an explicit construction of P_n having the required property (with the exponent $1/2 - 1/17$ replaced by $1/2 - 1/9 + \varepsilon$ for all $\varepsilon > 0$). One of their tools [1, §21] (by no means the only one!) is an estimate (with optimal cancellation) for the character sums [1, p. 689]

$$S(a_0, \dots, a_d; p) = \sum_{(y,x)\in \mathbf{F}_p \times \mathbf{F}_p^d} \chi(g(y))e\Big(\frac{a_0 y}{p}\Big) \prod_{j=1}^{d} \chi(f_j(x_j))e\Big(\frac{yx_j + a_j x_j}{p}\Big),$$

where $a_i \in \mathbf{Z}$, g and the f_j are integral polynomials with simple roots (and g is non-constant, $\deg(f_j) \leqslant 2$). In fact, Bombieri and Bourgain [1, Lemma 33] prove that

$$S(a; p) \ll p^{(1+d)/2},$$

the implied constant depending only on d and $\deg(g)$. Interestingly, their proof is an "elementary" argument based on the cohomological formalism and the Riemann Hypothesis,[25] which (in view of the fact that the number of variables is arbitrarily large) is a striking illustration of its power, and its versatility when combined with other tools (and certainly with clever ideas)...

7. Problems and speculations

To conclude this survey, we list – rather briefly – some problems and conjectures surrounding the Riemann Hypothesis over finite fields, emphasizing those closely connected to practical problems in analytic number theory.

– What happens when we consider character sums with "large degree", where the (known) degree of the L-function overwhelms the saving from the Riemann Hypothesis? For instance, consider a character sum

$$\sum_{1\leqslant x\leqslant p} e\Big(\frac{f(x)}{p}\Big)$$

[25] They also note that Katz has given a faster argument when $\deg(g) \geqslant 1 + d$, using more algebraic geometry.

where $f \in \mathbf{Z}[X]$ is such that $\deg(f) > p^{1/2}$. The Riemann Hypothesis gives only (via (20)) the bound

$$\Big| \sum_{1 \leqslant x \leqslant p} e\Big(\frac{f(x)}{p}\Big) \Big| \ll (\deg f)\sqrt{p}$$

which is worse than trivial! There have been quite a few investigations of such problems, in particular due to Bourgain, Konyagin, Heath-Brown, and these have shown that this type of questions is closely related to additive combinatorics and the sum-product phenomenon, for instance. However, no precise link between these results and the (still existing!) cohomological representation seems to be known. (See, e.g., [2]).

– What are general, uniform bounds, for the sums of Betti numbers C (see (21)) occurring in the rough bound (20)? In particular, how does this vary with p for sheaves of rank > 1, and is there a good theory of algebraic sums over the integers that explains the various uniformity statements which are known empirically (such as the bounds for Deligne-type character sums)? See the survey of Katz [15] for some speculations on this problem.

–Related to the previous item is the general question of understanding families of L-functions over finite fields when the base field (i.e., q) is fixed, but one has a sequence of sheaves with growing rank. The basic example here is that of families of curves with increasing genus, and in particular one can ask about the limiting behavior of the central value of the L-function of hyperelliptic curves given by equations

$$C_f \, : \, y^2 = f(x)$$

where f runs over the set $H_g(\mathbf{F}_q)$ of monic squarefree polynomials in $\mathbf{F}_q[X]$ of degree $2g+1$, the limit considered being $g \to +\infty$ (see the introduction of [18] for some discussion). There are conjectures about this problem, which are related to conjectures concerning moments of the Riemann zeta function (due to Keating-Snaith), but despite the availability of the Riemann Hypothesis, not much more is known, e.g., about the asymptotic behavior as $g \to +\infty$ of

$$\frac{1}{|H_g(\mathbf{F}_q)|} \sum_{f \in H_g(\mathbf{F}_q)} |L(C_f, q^{-1/2})|^k,$$

for $k \geqslant 1$. (Compare with the vertical limit (26)).

However, there is one special case that is much easier: in the 0-dimensional case (Example 15), it is possible to understand for instance the order of the pole at $T = 1$ of the zeta function of a 0-dimensional variety defined by the equation $f(x) = 0$, $f \in \mathbf{F}_q[X]$, with q fixed and $\deg(f) \to +\infty$ (see (15)), which (for f squarefree) is just the number of irreducible factors of f. We refer to [21] for a study of this question, where random permutations and probabilistic models of divisibility by fixed irreducible polynomials combine; the structure of the asymptotic formulas obtained closely parallels the conjectures for L-functions.

– Many conjectures about the distribution of Frobenius conjugacy classes in an "horizontal" direction remain very mysterious, the prototypical example being the

horizontal Sato-Tate conjecture: in the notation of Example 4, are the angles

$$\{\theta_{p,1} \in [0,\pi] \mid p \leqslant x\}$$

equidistributed with respect to the Sato-Tate measure as $x \to +\infty$? (Here we fix the parameter a as p varies, instead of averaging over it). This is an example where we do not have, a priori, a fixed "source group" Π_1 with Frobenius classes Fr_p, and homomorphism

$$\rho : \Pi_1 \longrightarrow GL(2,k)$$

with $\mathrm{Tr}(\rho(\mathrm{Fr}_p)) = HK(2;1,p)$. So there is not even a good reason to expect equidistribution with respect to a measure with group-theoretic origin.[26]

The strongest result in this direction is a theorem of Duke, Friedlander and Iwaniec (see [11, Cor. 21.9]), who (using spectral theory of automorphic forms for $GL(2)$ and sophisticated sieve methods) have solved the analogue conjecture for the Salié sums defined by

$$\sum_{x\in\mathbf{F}_p^\times} \Big(\frac{x}{p}\Big) e\Big(\frac{x+\bar{x}}{p}\Big), \qquad \Big(\frac{x}{p}\Big) \text{ the Legendre symbol.}$$

However, Salié sums are much simpler than Kloosterman sums from the cohomological point of view – this translates into the equidistribution measure being the Lebesgue measure, instead of the Sato-Tate measure –, so this does not give much hint about the way to proceed for Kloosterman sums.

– Finally, a vexing philosophical question: can one make the theory "easier to apply"? This is not merely a reflection on the mathematical complexity (or sophistication) of the cohomological framework, which in some ways is probably, in fact, as simple as it can be; rather, it has more to do with the lack of any direct link between a result like the wonderful upper bound

$$L(1/2,\chi_d) \ll_\varepsilon |d|^{1/6+\varepsilon}, \qquad \text{for all } \varepsilon > 0,$$

where χ_d is a real primitive character modulo d (which is due to Conrey and Iwaniec [3]), and the – crucial – estimation of the character sums

$$\sum_{x,y\in\mathbf{F}_p} \chi(xy(x+1)(y+1)) e\Big(\frac{xy-1}{p}\Big)$$

(where $\chi \neq 1$ is a multiplicative character modulo p) which enters in the proof.

References

[1] E. Bombieri and J. Bourgain: *On Kahane's ultraflat polynomials*, J. Eur. Math. Soc. 11 (2009), 627–703.

[2] J. Bourgain: *Mordell's exponential sum estimate revisited*, Journal A.M.S 18 (2005), 477–499.

[26] Things are slightly better for the distribution as p grows of the coefficients of a fixed elliptic curve, since the Sato-Tate conjecture has been proved in this case by Clozel, Harris, Sheperd-Barron, Taylor (see Mazur's survey [22]). But there, the group-theoretic framework does exist.

[3] J.B. Conrey and H. Iwaniec: *The cubic moment of central values of automorphic L-functions*, Ann. of Math. 151 (2000), 1175–1216.

[4] P. Deligne: *Cohomologie étale*, S.G.A $4\frac{1}{2}$, L.N.M 569, Springer Verlag (1977).

[5] P. Deligne: *La conjecture de Weil : I*, Publ. Math. IHÉS 43 (1974), 273–307

[6] P. Deligne: *La conjecture de Weil, II*, Publ. Math. IHÉS 52 (1980), 137–252.

[7] M. Fried, D. Haran and M. Jarden: *Effective counting of the points of definable sets over finite fields*, Israel J. of Math. 85 (1994), 103–133.

[8] C. Hall: *Big orthogonal or symplectic monodromy mod ℓ*, Duke Math. J. 141 (2008), 179–203; see also `arXiv:math.NT/0608718`.

[9] R. Hartshorne: *Algebraic geometry*, Grad. Texts in Math. 52, Springer-Verlag (1977).

[10] H. Iwaniec: *Topics in classical automorphic forms*, Grad. Studies in Math. 17, A.M.S (1997).

[11] H. Iwaniec and E. Kowalski: *Analytic Number Theory*, A.M.S Colloq. Publ. 53, A.M.S (2004).

[12] N. Katz: *Moments, monodromy and perversity: a diophantine perspective*, Annals of Math. Studies 159, Princeton Univ. Press 2005.

[13] N. Katz: *Larsen's alternative, moments, and the monodromy of Lefschetz pencils*, Contributions to automorphic forms, geometry, and number theory, 521–560, Johns Hopkins Univ. Press, Baltimore, MD, 2004.

[14] N. Katz: *Gauss sums, Kloosterman sums and monodromy*, Annals of Math. Studies, 116, Princeton Univ. Press, 1988.

[15] N. Katz: *Exponential sums over finite fields and differential equations over the complex numbers: some interactions*, Bull. A.M.S 23 (1990), 269–309.

[16] N. Katz: *Sums of Betti numbers in arbitrary characteristic*, Finite Fields Appl. 7 (2001), no. 1, 29–44.

[17] N. Katz: *Twisted L-functions and monodromy*, Annals of Math. Studies 150, Princeton Univ. Press 2002.

[18] N. Katz and P. Sarnak: *Random matrices, Frobenius eigenvalues and monodromy*, A.M.S Colloquium Publ. 45, 1999.

[19] E. Kowalski: *The large sieve and its applications*, Cambridge Tract. in Math. 175, Cambridge Univ. Press, 2008.

[20] E. Kowalski: *On the rank of quadratic twists of elliptic curves over function fields*, International J. of Number Theory 2 (2006), 267–288.

[21] E. Kowalski and A. Nikeghbali: *Mod-Poisson convergence in probability and number theory*, International Math. Res. Notices, to appear. `arXiv:0905.0318`.

[22] B. Mazur: *Finding meaning in error terms*, Bull. A.M.S 45 (2008), 185–228.

[23] J. Silverman: *The arithmetic of elliptic curves*, Grad. Texts in Math. 106, Springer Verlag 1986.

[24] A. Weil: *Numbers of solutions of equations in finite fields*, Bull. A.M.S 55 (1949), 497–508.

[25] A. Weil: comments on [24], Collected Works, vol. I, 568–569, Springer 1979.

E. Kowalski
ETH Zürich – D-MATH
Rämistrasse 101
8092 Zürich
Switzerland
e-mail: kowalski@math.ethz.ch

Received: November 30, 2009.