

# EXPONENTIAL SUMS OVER FINITE FIELDS

E. KOWALSKI

This text is an introduction to exponential sums over finite fields, and to the methods that have been developed in order to understand their properties. The main motivation is provided by the numerous applications of such sums to problems of analytic number theory, and we explain briefly how some of those arise. There exist many beautiful, fairly elementary techniques, in estimating exponential sums over finite fields, and they are still highly relevant in a number of difficult problems (e.g., when working with character sums involving polynomials of high degree compared with the cardinality of the underlying field). However, we will mostly try to explain and describe, as concretely as we can, the *cohomological methods*, which arose from the work and conjectures of A. Weil in the 1940's, and were then developed as one aspect of the tremendous foundational *tour de force* of Grothendieck and his school of algebraic geometers. In particular, we highlight the powerful formalism of weights of exponential sums which comes out of Deligne's (second) proof of the Weil Conjectures (i.e., the Riemann Hypothesis over finite fields), and part of the work of Katz who applied and refined this formalism. There is a very extensive theory of algebraic geometry in the background, but we will see that one can achieve a good basic understanding of the phenomena involved without knowing all the details. Indeed, we will conclude with a discussion of one of the most recent techniques found by Katz to compute the all-important "geometric monodromy group" of a family of exponential sums or  $L$ -functions, which he calls the *Larsen alternative*, and we will see that this carries a remarkably simple diophantine interpretation that brings the spotlight back to one of the very basic elementary method still used to study exponential sums (and other analytic quantities): the computation of moments.

It is a great pleasure to thank everyone involved with the organization of the ICTP Spring School and Conference on Analytic Number Theory, April 23 to May 11, 2007.

**Notation and prerequisites.** We try to keep the prerequisites minimal, however some familiarity with basic algebraic structures are required, though the first section at least does not require knowing much besides the fact that  $\mathbf{Z}/p\mathbf{Z}$  is a field if  $p$  is a prime number. In an Appendix, we list in particular the facts and definitions concerning finite fields and (very basic) algebraic geometry that we will use.

Throughout, we denote by  $\mathbf{F}_q$ , for  $q$  a power of a prime number  $p$ , a finite field with  $q$  elements. In particular,  $\mathbf{F}_p = \mathbf{Z}/p\mathbf{Z}$  if  $p$  is prime. Then  $\mathbf{F}_q^\times$  denotes the multiplicative group of invertible elements in  $\mathbf{F}_q$ .

The most fundamental truths about finite fields of characteristic  $p$  are the following: for  $\mathbf{F}_q$  such a field, and  $\mathbf{F}_{q^f}$  any extension of it, we have

$$(x + y)^q = x^q + y^q, \quad (xy)^q = x^q y^q, \quad \text{for all } x, y \in \mathbf{F}_{q^f},$$

and moreover

$$z \in \mathbf{F}_{q^f} \text{ is in } \mathbf{F}_q \text{ if and only if } z^q = z.$$

For any complex number  $z \in \mathbf{C}$ , we let  $e(z) = \exp(2i\pi z)$ . Note the following simple properties:  $e(z + w) = e(z)e(w)$ ,  $e(z) = 1$  if and only if  $z \in \mathbf{Z}$ ,  $|e(z)| = \exp(-2\pi \operatorname{Im}(z))$ .

This is a first draft where only the first part is complete. The second part will consider families of exponential sums.

## 1. WHAT AND WHY

1.1. **Introduction.** In following subsections, we will give three examples of very concrete and interesting problems of number theory, with an analytic flavor, that lead naturally, and ineluctably, to the question of finding non-trivial estimates for sums of the type

$$S = \sum_{x \in X} e(f(x))$$

where the finite summation set  $X$  is either  $\mathbf{Z}/p\mathbf{Z}$  or  $\mathbf{Z}/p\mathbf{Z} - \{0\}$  (or the complement of another finite set in  $\mathbf{Z}/p\mathbf{Z}$ ) and the phase function  $f$  is some real valued function defined on  $X$  (so that, in particular, we have  $|e(f(x))| = 1$  for all  $x$ ). The trivial bound is

$$|S| \leq |X|$$

and the goal of the investigations of exponential sums is to improve on this trivial bound, under some assumptions on this phase function. In the roughest approximation, the assumption must be some statement carrying the meaning that  $f(x)$  “varies a lot”, when seen a real number in  $[0, 1[\simeq \mathbf{R}/\mathbf{Z}$  (recall that  $e(y + 1) = e(y)$ ). Indeed, if  $f$  is constant (modulo  $\mathbf{Z}$ ), there can be no cancellation in the sum.

Before discussing further what are the phase functions which are relevant for our purpose here, let us mention a very important heuristic principle, sometimes called the “square-rooting principle”: unless some bias exists, the sum  $S$  should be of size roughly  $\sqrt{|X|}$  (of course, not exactly). This is justified for instance on probabilistic grounds as follows: assume for simplicity that we look at the average of  $|S|^2$ , performing the average over all  $2^p$  functions  $f(x)$  on  $\mathbf{Z}/p\mathbf{Z}$  taking value in  $\frac{1}{2}$  (so  $e(f(x)) = \pm 1$ ). This average is simply

$$\frac{1}{2^p} \sum_{\varepsilon_1 = \pm 1} \cdots \sum_{\varepsilon_p = \pm 1} \left| \sum_{i=1}^p \varepsilon_i \right|^2 = \frac{1}{2^p} \sum_{i,j=1}^p \sum_{\varepsilon_1 = \pm 1} \cdots \sum_{\varepsilon_p = \pm 1} \varepsilon_i \varepsilon_j$$

and the outer sum collapses to the diagonal terms with  $i = j$ , with  $\varepsilon_i \varepsilon_j = 1$ , since if  $i \neq j$  the multiple sum splits as a product containing a factor  $1 + (-1) = 0$  arising from the sum over the two values of  $\varepsilon_i$ . For each  $i$ , we obtain from  $i = j$  the contribution  $2^{p-2}$  and therefore the average is  $\frac{p}{4}$  which shows that the average (mean-square) value of  $|S|$  is  $\sqrt{p}/2$  in this case.<sup>1</sup>

In the cases which interest us, the phase function  $f(x)$  has a particular algebraic structure. We will give quite general examples later on, but the basic ones to keep in mind are the following:

– We may have

$$f(x) = \frac{g(x)}{p}$$

for some polynomial  $g \in \mathbf{Z}[X]$  (this ensures that  $e(f(x))$  can be computed from any integer in the congruence class of  $x$  modulo  $p$ ), giving an *additive character sum*. The simplest (non-trivial) such sum is

$$S_1 = \sum_{x \in \mathbf{F}_p} e\left(\frac{x^2}{p}\right),$$

---

<sup>1</sup> Probabilistically, this is the computation of the variance of a sum of independent Bernoulli variables.

which is called a *quadratic Gauss sum*. Another example, extending slightly the case of polynomials, is the *Kloosterman sum* defined by

$$S(1, 1; p) = \sum_{x \in \mathbf{F}_p^\times} e\left(\frac{x + \bar{x}}{p}\right)$$

where  $\bar{x}$ , according to a long-standing convention, is the inverse of  $x$  modulo  $p$  (so that in fact we have here  $g(x) = x + x^{-1}$ , which is a rational function).

– We may have instead

$$e(f(x)) = \chi(g(x))$$

for some polynomial  $g \in \mathbf{Z}[X]$ , or rational function, and some *multiplicative character*  $\chi$  of  $\mathbf{F}_p^\times$ , getting a *multiplicative character sum*. Among these, the simplest non-trivial case arises when  $\chi$  is the quadratic character modulo an odd prime  $p$ , i.e., the Legendre character  $\left(\frac{\cdot}{p}\right)$  defined equivalently by

$$\chi(x) = \left(\frac{x}{p}\right) = \begin{cases} -1 & \text{if } x \text{ is not a square modulo } p \\ 0 & \text{if } x = 0 \text{ modulo } p \\ 1 & \text{if } x \text{ is a square modulo } p. \end{cases}$$

This leads to the sums

$$S_3 = \sum_{x \in \mathbf{F}_p} \left(\frac{f(x)}{p}\right)$$

where  $f$  is a fixed polynomial in  $\mathbf{Z}[X]$ . Those sums are by no means completely understood. For instance, if  $f$  is of degree 3 (and has no repeated root), understanding the behavior of the sums  $S_3$  as a function of  $p$  is intrinsically related with the problem of the modularity of elliptic curves, which is one of the deepest parts of Wiles's proof of the Fermat Great Theorem.

– As a third example, we can mix the two types by multiplying one phase factor of each type:

$$e(f(x)) = \chi(f_1(x))e\left(\frac{f_2(x)}{p}\right)$$

for some character of  $\mathbf{F}_p^\times$ . The simplest case is when  $\chi$  is the quadratic character modulo an odd prime,  $f_1 = f_2 = X$ , i.e., when we look at

$$S_4 = \sum_{x \in \mathbf{F}_p} \left(\frac{x}{p}\right)e\left(\frac{x}{p}\right).$$

This is also called a *Gauss sum*, and in fact we have

$$S_1 = S_4$$

(which was known to Gauss) if  $p$  is an odd prime. This property is obtained quite simply from the simple remark that

$$1 + \left(\frac{x}{p}\right) = |\{y \in \mathbf{F}_p \mid y^2 = x\}|,$$

together with a rearranging of  $S_2$  according to the value of  $x^2$ :

$$S_4 = \sum_{y \in \mathbf{F}_p} e\left(\frac{y}{p}\right) |\{y \in \mathbf{F}_p \mid y^2 = x\}| = \sum_{y \in \mathbf{F}_p} e\left(\frac{y}{p}\right) + \sum_{y \in \mathbf{F}_p} \left(\frac{x}{p}\right)e\left(\frac{y}{p}\right)$$

and finally the property (called *orthogonality* of characters) that

$$\sum_{x \in \mathbb{F}_p} e\left(\frac{y}{p}\right) = 0$$

which in this case is simply a question of summing a finite geometric series.

Gauss also proved the remarkable result that

$$|S_4| = |S_1| = \sqrt{p},$$

which can be seen as a first (spectacular) confirmation of the square-rooting philosophy...

**1.2. Example 1: representations of integers by quadratic forms.** Few arithmetic problems are as classical as that of asking which integers are sums of 2, 3, or 4 square, or more generally, given  $k \geq 1$ , how many representations of  $n \geq 1$  as a sum

$$n = x_1^2 + \cdots + x_k^2$$

with  $x_i \geq 0$  integers there are, as a function of  $n$ . Let's call this number  $r_k(n)$ .

In particular cases, for instance for sums of 2, 3, or 4 squares, there are celebrated classical results leading to explicit formulas for the number of solutions. However, in general, one can not expect a very simple answer, and the reasons for this are well-understood (non-triviality of coefficients of some modular forms). A general analytic tool to attack such problems (and even more general ones) was developed by Ramanujan, Hardy and Littlewood: the *circle method*, which may be summarized very briefly by writing the expression

$$r(n) = \int_0^1 \left( \sum_{j \leq \sqrt{n}} e(j^2 t) \right)^k e(-nt) dt$$

for the number of desired solutions, which is a simple consequence of orthogonality of characters (and the positivity of squares, which is a non-trivial fact!).

Lagrange had proved (by more or less algebraic means) that  $r_4(n) \geq 1$  for all  $n \geq 1$ , hence trivially  $r_k(n) \geq 1$  for all  $k \geq 4$  and  $n \geq 1$ . The circle method was originally capable of proving  $r_k(n) \geq 1$  for fixed  $k \geq 5$  and  $n$  large enough. (In fact, with an asymptotic formula for  $r_k(n)$  as  $n$  gets large). It turns out that there are very intrinsic reasons why the simplest application of the circle method fails when  $k = 4$ , but Kloosterman (just after his thesis, in 1922) found a remarkable refinement of the original circle method, and in doing so showed that  $r_4(n) \geq 1$  could be proved analytically<sup>2</sup> (again, for  $n$  large enough; being more precise in this respect is another question entirely) *provided* one knew a non-trivial bound for the general sums of the type

$$(1) \quad S(a, b; p) = \sum_{(x,p)=1} e\left(\frac{ax + b\bar{x}}{p}\right)$$

if  $p \nmid ab$  (which generalize the Kloosterman sum written above). Moreover, he proved such a result, namely

$$|S(a, b; p)| < 2p^{3/4}$$

if  $p \nmid 2ab$ , by evaluating the fourth moment

$$\sum_{(a,p)=1} S(a, 1; p)^4 = 2p^3 - 3p^2 - p - 1.$$

We will come back to this in the last section...

---

<sup>2</sup> More importantly, he could deal with any positive definite quadratic form in four variables, where special methods in the Lagrange manner can not work.

1.3. **Example 2: short character sums.** Our next problem is well-known, though maybe not as much as the previous one, and it may look less motivated to beginners. It should be seen as a prototypical example of a large class of problems which are very difficult and challenging and (for the moment) resist complete resolution.

This question is as follows: let  $p$  be an odd prime number, and denote by  $q(p)$  the smallest prime number  $q \geq 2$  such that  $q$  is not a square modulo  $p$ , i.e., the equation  $x^2 = q$  has no solution in  $\mathbf{Z}/p\mathbf{Z}$ .<sup>3</sup> In algebraic number theory terms, this can be rephrased as saying that  $q(p)$  is the smallest prime which is *split* in the quadratic field  $\mathbf{Q}(\sqrt{p})$ . Then we wish to find an upper bound for  $q(p)$  as a function of  $p$ .

Clearly,  $q(p)$  is  $< (p - 1)/2$  since (non-zero) squares and non-squares are as numerous and are all found between 1 and  $p - 1$ . The link with character sums looks almost ridiculously simplistic: we can say that  $q(p)$  is the smallest integer  $q$  such that

$$\sum_{j=1}^q \left(\frac{j}{p}\right) < q.$$

Indeed, this sum (up to  $q(p)$ ) is equal to  $q - 2$ , and conversely, if for some upper bound  $q$  it is  $< q$ , this means there is *at least* one term  $j$  for which the summand is  $\neq 1$ , and this shows that  $q(p) < q$ . Now, what this means is that if, for  $X$  quite small compared with  $p$ , we can show that there is cancellation, we obtain automatically an upper bound for  $q(p)$ . For instance, if

$$\left| \sum_{x \leq X} \left(\frac{x}{p}\right) \right| \leq CX^{1-\delta}$$

for some constant  $C \geq 1$  and some  $\delta > 0$ , then it follows that  $q(p)$  is of order of magnitude at most  $X$ . Typically, one wishes to take  $X = p^\theta$  with  $\theta < 1$ .

Now notice that in this case, the summation set here is *not* the full set of residues modulo  $p$ , so the sum does not look like one of those we defined previously. However, it turns out that there are techniques to relate such *incomplete sums* to “full” character sums. The simplest way to do this is to write

$$\sum_{x \leq X} \left(\frac{x}{p}\right) = \sum_{x \pmod{p}} \left(\frac{x}{p}\right) \varphi(x)$$

where  $\varphi$  is the characteristic function of the desired *short interval*  $n \leq X$ . Then one expands the latter in (discrete) Fourier series

$$\varphi(x) = \sum_{a \pmod{p}} \alpha(a, X) e\left(\frac{ax}{p}\right)$$

for some (explicit) Fourier coefficients  $\alpha(a, X) \in \mathbf{C}$ . Then after exchanging the two sums resulting from inserting this in the short sum, we obtain

$$\sum_{a \pmod{p}} \alpha(a, X) T(a)$$

where

$$T(a) = \sum_{x \pmod{p}} \left(\frac{x}{p}\right) e\left(\frac{ax}{p}\right).$$

---

<sup>3</sup> It is easy to see that asking for the smallest integer  $q \geq 1$  with this property is equivalent; the first quadratic non-residue must be a prime number.

This looks like a Gauss sum (except for  $a = 0$ , where  $T(0) = 0$ , as a sum of values of a character over the full group); indeed, it satisfies again  $|T(a)| = \sqrt{p}$ , and after some checking of the modulus of  $|\alpha(a, X)|$  for  $a \neq 0$ , one finds the *Polya-Vinogradov inequality*

$$\left| \sum_{x \leq X} \left( \frac{x}{p} \right) \right| \leq C \sqrt{p} \log p$$

for some (explicit) constant  $C > 0$ . Thus the smallest quadratic non-residue can not be bigger than (roughly)  $\sqrt{p}$ . However, it is suspected that this first bound is very far from the truth. Indeed, using the Generalized Riemann Hypothesis and standard analytic methods of  $L$ -functions, one is led to  $q(p) \ll (\log p)^2$ .

It turns out that one can do better than the Polya-Vinogradov inequality for this problem using more sophisticated exponential sums over finite fields, precisely the full strength (almost) of Weil's estimates for one-variable character sums (described below). This was discovered by Burgess, and remains a remarkable result, and a striking example of the type of ingenious arguments required so exploit the full power of algebraic geometry estimates in analytic number theory.

**1.4. Example 3: an equidistribution problem.** Our last example is again a very special case of a general type of problems and results, this time related to the general notion of *equidistribution*. To start with a general definition (not in the fullest generality!), a sequence  $(x_n)$  of real numbers is *equidistributed modulo 1* if for any interval  $I = [\alpha, \beta] \subset [0, 1]$  with  $\alpha < \beta$ , we have

$$\lim_{N \rightarrow +\infty} \frac{1}{N} |\{n \leq N \mid \{x_n\} \in I\}| \rightarrow \beta - \alpha,$$

which means that the “right” proportion of elements of the sequence have fractional part  $\{x_n\}$  in the (fixed) interval  $I$ . In particular, a sequence which is equidistributed modulo 1 has the property that the set of values of the fractional parts of  $x_n$  is dense in  $[0, 1]$ .

Various criteria exist to prove equidistribution, most of which are based on the remark that for fixed  $N$ , the left-hand side in the limit above is

$$\frac{1}{N} \sum_{n \leq N} f_I(\{x_n\})$$

where  $f_I$  is the characteristic function of  $I$ , whereas the right-hand side's putative limit is

$$\int_0^1 f_I(x) dx$$

Now the point is that characteristic functions are typically difficult to investigate using harmonic analysis. So one tries instead to prove

$$\frac{1}{N} \sum_{n \leq N} \varphi(\{x_n\}) \rightarrow \int_0^1 \varphi(x) dx$$

for *better* functions  $\varphi : [0, 1] \rightarrow \mathbf{C}$ , and it turns out that if this holds for a set of functions which is (or rather, its linear combinations are) big enough, there is equidistribution. Note that, tautologically, this limit is valid if  $\varphi$  is the constant function 1.

We mention two possible choices, the first one of which is known usually as *Weyl's criterion*: the sequence  $(x_n)$  is equidistributed modulo 1 if and only if for all  $h \in \mathbf{Z}$ ,  $h \neq 0$ , we have

$$\frac{1}{N} \sum_{n \leq N} e(hx_n) \rightarrow 0,$$

which is all the more convenient because no fractional-part function is required, due to the 1-periodicity of  $z \mapsto e(z)$ . Note that the left-hand side is an exponential sum (not necessarily over a finite field)!

A second option is to use  $\varphi(x) = x^k$  for  $k \in \mathbf{Z}$ ; this is known as the *moment method* because the left-hand side

$$\frac{1}{N} \sum_{n \leq N} \{x_n\}^k$$

is the  $k$ -th moment of the fractional parts of  $x_n$ ,  $n \leq N$ .

Now consider an odd prime number  $p$ , and a multiplicative character  $\chi$  modulo  $p$ , which is non-trivial. One can define a Gauss sum similar to the one above as

$$\tau(\chi) = \sum_{x \pmod{p}} \chi(x) e\left(\frac{x}{p}\right).$$

Again, it turns out that  $|\tau(\chi)| = \sqrt{p}$ . Hence we can write

$$\tau(\chi) = \sqrt{p} e(\theta(\chi))$$

where  $\theta(\chi) \in [0, 1[$  is unique, and is called naturally enough the *argument of the Gauss sum*. Considerable ingenuity was expended in trying to express more concretely this argument; Gauss managed to prove that if  $\chi$  is quadratic, then

$$\tau(\chi) = \begin{cases} 1 & \text{if } p \equiv 1 \pmod{4} \\ i & \text{if } p \equiv 3 \pmod{4}. \end{cases}$$

However, no such explicit formula seems available for  $\theta(\chi)$  as  $\chi$  runs over all the  $p - 2$  non-trivial Dirichlet characters modulo  $p$ . One may wonder if this is for a good reason, or just because we are not imaginative enough. The following result shows that the answer can not be completely trivial (if it exists!): as  $p \rightarrow +\infty$ , the angles  $(\theta(\chi))_{\chi \neq 1}$  become equidistributed modulo 1. Now notice that this is a slightly different definition from the one above, however the meaning is fairly clear: instead of the averages over  $n \leq N$

$$\frac{1}{N} \sum_{n \leq N} \varphi(\{x_n\})$$

one considers limits of similar averages over  $\chi \neq 1$ , as  $p \rightarrow +\infty$ , i.e., one wishes to prove that

$$\lim_{p \rightarrow +\infty} \frac{1}{p-2} \sum_{\substack{\chi \pmod{p} \\ \chi \neq 1}} \varphi(\theta(\chi)) = \int_0^1 \varphi(x) dx.$$

In this situation the most convenient approach is the use of the Weyl Criterion. For  $h \geq 1$ , we have

$$\begin{aligned} \frac{1}{p-2} \sum_{\substack{\chi \pmod{p} \\ \chi \neq 1}} e(h\theta(\chi)) &= \frac{1}{p-2} \sum_{\substack{\chi \pmod{p} \\ \chi \neq 1}} e(\theta(\chi))^h \\ &= \frac{1}{p-2} \sum_{\substack{\chi \pmod{p} \\ \chi \neq 1}} \left(\frac{\tau(\chi)}{\sqrt{p}}\right)^h. \end{aligned}$$

We continue by replacing the Gauss sum using its expression as a character sum and expanding the  $h$ -fold product; this leads to

$$(2) \quad \sum_{x_1, \dots, x_h \pmod{p}} \cdots \sum_{\chi \neq 1} e\left(\frac{x_1 + \cdots + x_h}{p}\right) \chi(x_1 \cdots x_h),$$

after putting the summation over  $\chi$  in the inner sum. By orthogonality of characters, noting that all  $x_i$  are invertible, the sum over *all* characters is 0 except when  $x_1 \cdots x_h = 1$ , and then its value is  $p - 1$ . So without the trivial character we have either  $-1$ , or  $p - 2$ , which we write as  $p - 1 - 1$ . Taking the contribution of this  $p - 1$ , we obtain

$$\frac{p-1}{p^{h/2}} S_h(1; p)$$

where we have obtained a case of a *hyper-Kloosterman sum*, defined as

$$(3) \quad S_h(a; p) = \sum_{\substack{x_1, \dots, x_h \pmod{p} \\ x_1 \cdots x_h = a}} e\left(\frac{x_1 + \cdots + x_h}{p}\right),$$

generalizing the previous Kloosterman sum, which was the case  $h = 2$ . This expression is an exponential sum over finite fields, but over  $h - 1$  variables (if  $a \neq 0$ ,  $x_1, \dots, x_{h-1}$  are free to range over arbitrary elements modulo  $p$ , while they determine uniquely the last variable  $x_h$ ).

What remains is the sum

$$-\frac{1}{p-2} \frac{1}{p^{h/2}} \sum_{\substack{x_1, \dots, x_h \pmod{p} \\ x_1 \cdots x_h \neq 0}} e\left(\frac{x_1 + \cdots + x_h}{p}\right)$$

and the sum splits out as an  $h$ -th power again, and is equal to  $-1$  (orthogonality of additive characters this time).

We have therefore proved that

$$\frac{1}{p-2} \sum_{\substack{\chi \pmod{p} \\ \chi \neq 1}} e(h\theta(\chi)) = \frac{p-1}{p-2} \frac{S_h(1; p)}{p^{h/2}} - \frac{(-1)^h}{(p-2)p^{h/2}}$$

if  $h \geq 1$ , and by conjugation we also have

$$\frac{1}{p-2} \sum_{\substack{\chi \pmod{p} \\ \chi \neq 1}} e(-h\theta(\chi)) = \frac{p-1}{p-2} \frac{\overline{S_h(1; p)}}{p^{h/2}} - \frac{(-1)^h}{(p-2)p^{h/2}}$$

We therefore see that proving equidistribution amounts to showing that

$$\frac{S_h(1; p)}{p^{h/2}} \rightarrow 0$$

as  $p \rightarrow +\infty$ . Note that here the “square-rooting philosophy” predicts that  $S_h(1; p)$  should be of size roughly  $p^{(h-1)/2}$ , which would lead to the desired result. However, confirming this is not trivial at all. We will describe further how Deligne proved that

$$(4) \quad |S_h(1; p)| \leq hp^{(h-1)/2},$$

confirming the desired equidistribution of arguments of Gauss sums.

Note also that the trivial bound is  $|S_h(1; p)| \leq p^{h-1}$ , which gets further and further away from what is required as  $h$  grows, so that one cannot be content, in this application, to gain just a little on the trivial bound...

1.5. **The classical results of Weil.** We will now describe the way Weil interpreted exponential sums over finite fields and proved the first estimates concerning one-variable sums which are, in a sense, best possible. Those sums are of the type

$$S = \sum_{x \in \mathbf{F}_q}^* \chi(f(x))\psi(g(x))$$

where  $f$  and  $g$  rational functions, with  $f$  not identically zero, and  $\sum^*$  restricts the sum to elements in  $\mathbf{F}_q$  where  $f$  is defined.

There are obvious cases where  $S$  can not have modulus significantly smaller than  $q$ : if either  $f$  or  $g$  is constant, for instance. Weil's results showed for the first time that, except when the function  $\chi(f(x))\psi(g(x))$  is constant, a one variable sum exhibits optimal cancellation as far as  $p$  is concerned, i.e., there exists then a constant  $C(f, g)$  such that

$$|S| \leq C\sqrt{p}$$

for all primes  $p$ . This constant  $C$  is typically the product of two terms, one of which is  $(p, N)^{1/2}$ , where  $N$  is the product of finitely many primes where  $f$  or  $g$  becomes "degenerate", and a constant depending at most (and polynomially) on the degree of  $f$  and  $g$ .

The main thrust of Weil's work is very algebraic, in contrast with the type of elementary techniques used previously. There are three main steps in Weil's argument, two of which are rather formal, though by no means obvious, while the third is the Riemann Hypothesis for curves over finite fields, and was the crowning achievement of his work on the foundations of algebraic geometry. We will describe (briefly) those three steps in the case of the Kloosterman sums  $S(a, b; p)$  (see (1)).

– (Forming *companion* sums and an  $L$ -function) Recall from the elementary theory of finite fields that for any  $\nu \geq 1$ , there exists a unique (up to isomorphism) extension field  $\mathbf{F}_{p^\nu}$  of  $\mathbf{F}_p$  of degree  $\nu$ ; field theory then provides two interesting maps, the trace and the norm, which are homomorphisms

$$\text{Tr} : \begin{cases} \mathbf{F}_{p^\nu} \rightarrow \mathbf{F}_p \\ x \mapsto x + x^p + \dots + x^{p^{\nu-1}} \end{cases}$$

and

$$N : \begin{cases} \mathbf{F}_{p^\nu}^\times \rightarrow \mathbf{F}_p^\times \\ x \mapsto x \cdot x^p \dots x^{p^{\nu-1}} \end{cases}$$

(for the additive and multiplicative structures respectively).

For each such field, one can then extend the additive character  $x \mapsto e(x/p)$  defined on  $\mathbf{F}$  by the formula

$$\psi_\nu(x) = e(\text{Tr}(x)/p)$$

(and similarly, one can extend any multiplicative character  $\chi$  of  $\mathbf{F}_p^\times$  to  $\mathbf{F}_{p^\nu}^\times$  by

$$\chi_\nu(x) = \chi(N(x))$$

for  $x \in \mathbf{F}_{p^\nu}$ ,  $\nu \geq 1$ ).

Given  $a, b \in \mathbf{F}_p$ , define then the "companion" Kloosterman sums  $S(a, b; q)$  for  $q = p^\nu$  by the formula

$$S(a, b; q) = \sum_{x \in \mathbf{F}_q^\times} \psi_\nu(ax + bx^{-1}),$$

so that  $S(a, b; p)$  is indeed the original sum. This leads to a sequence  $(S(a, b; p^\nu))$  of exponential sums,  $\nu \geq 1$ . This sequence is then packaged as a generating series, which is (for reasons which will be clearer later)

$$L_{a,b} = \exp\left(\sum_{\nu \geq 1} \frac{S(a, b; p^\nu)}{\nu} T^\nu\right) \in \mathbf{Q}[[T]].$$

– (The  $L$ -function is a polynomial) The next step is the investigation of the nature of the  $L$ -function, and consists in proving that this formal power series is in fact a polynomial, more precisely it is

$$L_{a,b} = 1 + S(a, b; p^\nu)T + pT^2.$$

It is of course clear that  $L_{a,b}(0) = 1$ , and differentiating leads immediately to  $L'_{a,b}(0) = S(a, b; p^\nu)$ . However, it is not so clear that the terms with degree  $\geq 3$  vanish!

Weil's proof is based on an interpretation of  $L_{a,b}$  as an analogue of a Dirichlet  $L$ -function. More precisely, consider the group  $G$  of rational functions  $f/g \in \mathbf{F}(X)$  where  $f$  and  $g$  are monic polynomials both defined and non-zero at  $X = 0$ , and let

$$\psi(f/g) = \psi(f)\psi(g)^{-1}, \quad \text{with} \quad \psi(X^d + a_1X^{d-1} + \cdots + a_{d-1}X + a_d) = e\left(\frac{aa_1 + ba_{d-1}\bar{a}_d}{p}\right).$$

Then one shows that  $\psi$  is multiplicative and that

$$L_{a,b}(T) = \sum_h \psi(h)T^{\deg(h)},$$

the sum running over  $h \in \mathbf{F}[X]$  monic and non-zero at  $X = 0$ .

Another way to see this result is to write  $T = p^{-s}$ , for  $s$  a complex variable, so that with  $N(h) = p^{\deg(h)}$ , the formula becomes

$$L_{a,b}(p^{-s}) = \sum_h \psi(h)N(h)^{-s} = 1 + S(a, b; p)p^{-s} + p^{1-2s},$$

and this formula shows that  $L_{a,b}(p^{-s})$ , a priori only defined for  $\text{Re}(s)$  large enough, is in fact an entire function. We will not pursue this here this analogy with Dirichlet  $L$ -functions.

– (The Riemann Hypothesis) Finally, Weil managed to prove the analogue of the Riemann Hypothesis; in terms of  $L_{a,b}(p^{-s})$ , this means that the zeros of  $L_{a,b}(p^{-s})$  satisfy  $\text{Re}(s) = \frac{1}{2}$ . Since  $T = p^{-s}$ , this means exactly that the zeros (two at most) of  $L_{a,b}(T)$  have modulus  $1/\sqrt{p}$ , which in turn implies that

$$1 + S(a, b; p)T + pT = (1 - \alpha T)(1 - \beta T)$$

with  $|\alpha| = |\beta| = \sqrt{p}$ , and hence  $S(a, b; p) = -(\alpha + \beta)$ , so that  $|S(a, b; p)| \leq 2\sqrt{p}$ .

*Remark 1.* In the late 1960's, Stepanov found a new beautiful elementary argument to prove bounds of the same quality as Weil's bound for one variable exponential sums, i.e., in the case of  $S(a, b; p)$ , he proved

$$|S(a, b; p)| \leq C\sqrt{p}$$

for  $p \nmid 2ab$  and  $C$  some (explicit) constant. In combination with the first two (simpler) steps of Weil's method, this leads to an alternative proof of

$$|S(a, b; p)| \leq 2\sqrt{p}$$

with the "right" constant. Moreover, it should be emphasized that Stepanov's method is not simply about recovering results coming from algebraic geometry without learning

this subject: in some situations, its explicitness may be very useful (e.g., for sums over  $\mathbf{F}_p$  involving polynomials of high degree, such as

$$S = \sum_{x \in \mathbf{F}_p} e(x^{2d} + x^d + 1)$$

where Weil's constant is so large that the estimate is trivial if  $d$  is larger than  $\sqrt{p}$ ).

## 2. THE COHOMOLOGICAL FORMALISM

**2.1. Introduction.** Weil's algebraic theory of exponential sums over finite fields was not well-suited to sums involving more than one variable. Although arguments based on induction on the dimension were possible, they typically could not lead to the square-root cancellation expected in many multi-variable exponential sums (such as the hyper-Kloosterman sum (3)).

However, Weil also formulated conjectures concerning point-counting over finite fields which led Grothendieck and his school to develop techniques of algebraic geometry which, especially after the ground-breaking work of Deligne and the deep studies of Katz, lead to a very general formalism of exponential sums, expanding Weil's insight not only in the direction of dealing with many variables, but also in that of being able to work with sums of new types, which have considerable importance in modern analytic number theory.

We will try to discuss this formalism, in as concrete a way as possible ; the apparent complexity of the whole framework involved should not hide its underlying simplicity from analytic number theorists.

**2.2. Sheaf formalism.** We saw in Section 1.5 that Weil had interpreted one-variable exponential sums by writing companion sums  $S_\nu$  and interpreting their generating series as a sum analogue to a Dirichlet series involving a polynomial ring over  $\mathbf{F}_p[X]$  and a finite quotient of it.

The underlying thrust of the sheaf-theoretic approach can then be said to be, first of all, the effect of the process of looking instead at characters of *Galois groups*. The correspondance between Weil's work and this new approach is then an instance of a *reciprocity law*. However, it is in fact rather simpler to find and define the Galois characters than it is to find Weil's (highly ingenious) characters, though some objects involved may look more abstract.

Here is the basic type of argument involved, still in the case of a Kloosterman sum. We start with a prime  $p$ ; for all powers  $q = p^\nu$  of  $p$ , we have the subset  $\mathbf{F}_q^\times$  of  $\mathbf{F}_q$ , and the union  $\bar{\mathbf{F}}_q^\times$  over all  $\nu$ . Consider the curve  $C$  with equation

$$(5) \quad y^p - y = x + 1/x,$$

seen as a subset of  $\bar{\mathbf{F}}_p^\times \times \bar{\mathbf{F}}_p$ , which we see as mapping to  $\bar{\mathbf{F}}_p$  by  $(x, y) \mapsto x$ . This curve has some nice symmetry : since  $(y + a)^p = y^p + a^p$  for all  $a \in \bar{\mathbf{F}}_p$ , and  $a^p = a$  for  $a \in \mathbf{F}_p$  itself, the additive group  $G = \mathbf{F}$  acts by

$$\sigma_a : (x, y) \mapsto (x, y + a).$$

Moreover, a basic result states that if  $z \in \bar{\mathbf{F}}_p$  is arbitrary, and if we fix a root  $z_0$  of  $y^p - y = z$ , we can obtain all other roots as  $z_0 + a$  for  $a \in \mathbf{F}_p$ , and this means that for any fixed  $x \in \bar{\mathbf{F}}_p^\times$ , the pre-image  $f^{-1}(x)$  consists of the orbit under the  $G$ -action of the  $\sigma_a$ 's of a fixed solution of  $y^p - y = x + 1/x$ .

Now consider any  $x \in \bar{\mathbf{F}}_q$  and  $y$  such that  $(x, y) \in C$ . We have  $f(x^q, y^q) = x^q = x$ , and therefore there exists some  $a \in G$  such that

$$(6) \quad y^q = y + a;$$

notice that  $a$  is independent of the choice of  $y$ : if we were to replace  $y$  by another point on the pre-image, it would be of the form  $\sigma_b(y) = b + y$ , with  $b \in G$ , and then  $(b + y)^q - (b + y) = y^q - y = a$ .

We denote  $\tilde{\text{Fr}}_{x,q} = a$ , and call this (somewhat temporarily) the *Frobenius at  $x$  relative to  $\mathbf{F}_q$* . Now we claim that the phase

$$e\left(\frac{\text{Tr}(x + x^{-1})}{p}\right)$$

which is the basis of the Kloosterman sum can be written as

$$e\left(\frac{\text{Tr}(x + x^{-1})}{p}\right) = e\left(\frac{\tilde{\text{Fr}}_{x,q}}{p}\right).$$

Notice that this is tautologous for  $q = p$  by comparing (6) and (5). For  $q = p^\nu$ , we simply write a telescoping sum

$$\begin{aligned} y^q - y &= y^{p^\nu} - y = (y^{p^\nu} - y^{p^{\nu-1}}) + \cdots + (y^p - y) \\ &= (y^p - y)^{p^{\nu-1}} + \cdots + (y^p - y)^p + (y^p - y) \\ &= \text{Tr}(y^p - y) \\ &= \text{Tr}(x + x^{-1}) \end{aligned}$$

by means of (5) again.

In other words, we have shown that the companion Kloosterman sums can be written as

$$S_\nu = \sum_{x \in \mathbf{F}_{p^\nu}^\times} e\left(\frac{\text{Tr}(x + x^{-1})}{p}\right) = \sum_{x \in \mathbf{F}_{p^\nu}^\times} e\left(\frac{\tilde{\text{Fr}}_{x,q}}{p}\right).$$

If we now define  $V$  to be the vector space  $\mathbf{C}$  and

$$\tilde{\psi} : G \rightarrow GL(V)$$

to be the multiplication by  $e(a/p)$ , then we can also write

$$S_\nu = \sum_{x \in \mathbf{F}_{p^\nu}^\times} \text{Tr}(\tilde{\psi}(\tilde{\text{Fr}}_{x,q}) | V),$$

where  $\text{Tr}(A | V)$  is the trace of an endomorphism of any vector space.

Now it turns out that this can be enormously generalized, and more to the point, that this enormous generality does buy *something!* So, to define what can be considered as the most general type of “algebraic” exponential sum, we consider first an algebraic “summation set”, which will be an algebraic variety  $U$  over some finite field  $\mathbf{F}_q$ , i.e., the set of solutions of some polynomial equations with coefficient in  $\mathbf{F}_q$ , with restrictions such as the non-vanishing of some other polynomials.<sup>4</sup> We do not need to assume any non-singularity condition at the beginning, and we do not try to speak of projective variety (most summation sets for exponential sums are not defined as projective varieties). We write  $U(\mathbf{F}_{q^\nu})$  for the set of points of  $U$  with coefficients in  $\mathbf{F}_{q^\nu}$  (in other words, solutions

---

<sup>4</sup> By replacing conditions like  $x \neq 0$  with  $xy = 1$  where  $y$  is an additional variable, allowing this type of conditions is seen to bring no more generality.

to the defining equations in this field), and either  $\bar{U}$  or  $U(\bar{\mathbf{F}}_q)$  for the set of all solutions with coefficient in an algebraic closure of  $\mathbf{F}_q$ .

Now if we were interested in an additive character sum of the type

$$\sum_{x \in U(\mathbf{F}_q)} e\left(\frac{\text{Tr}(f(x))}{p}\right),$$

we could argue as above with the algebraic variety  $V$  with equation

$$y^q - y = f(x),$$

involving as automorphism group (i.e., as analogue of  $G$ ) the additive group of  $\mathbf{F}_q$ .

However, we are going to consider more complicated settings, and for this we introduce a black-box group of variegated shape, denoted  $\Pi_1$ , which is technically known as the *arithmetic fundamental group* of  $U$ .<sup>5</sup> The link with the previous construction is that, corresponding to the curve  $C$ , there is a surjective group homomorphism  $\Pi_1 \rightarrow G$ ,<sup>6</sup> so we can see the map  $\tilde{\psi}$  above as defined on  $\Pi_1$  instead of  $G$  by composing, say  $\tilde{\psi} : \Pi_1 \rightarrow G \rightarrow GL(V)$ . More importantly, the elements  $\tilde{\text{Fr}}_{x,q}$  are the image using  $\Pi_1 \rightarrow G$  of elements  $\text{Fr}_{x,q}$  which lie in  $\Pi_1$ , and the sum becomes of the form

$$\sum_{x \in U(\mathbf{F}_q)} \text{Tr}(\psi(\text{Fr}_{x,q}) | V).$$

Now comes the generalization: for any homomorphism  $\Pi_1 \xrightarrow{\rho} GL(W)$ , where  $W$  is a finite-dimensional vector space over some field  $K$ , we can define

$$S_\nu(\rho) = \sum_{x \in U(\mathbf{F}_{q^\nu})} \text{Tr}(\rho(\text{Fr}_{x,q^\nu}) | W),$$

which lies in  $K$  and generalizes the exponential sums. Such a  $\rho$  is called a *representations of  $\Pi_1$  on  $W$* .

This is indeed what the sums described below will look like, but before going further, there are technical issues needed which arise in the actual construction:

— The Frobenius elements  $\text{Fr}_{x,q^\nu}$  are, in fact, not well-defined in  $\Pi_1$ , but are defined up to conjugation. This is not a problem since we take a trace afterwards – and explains why one can not take consider more general expressions which are not conjugacy-invariant (e.g., some coefficient of the matrix  $\rho(\text{Fr}_{x,q^\nu})$  in some fixed basis).

— For geometric reasons, it is more convenient to replace  $\text{Fr}_{x,q}$  by its inverse, the so-called (local) *geometric Frobenius* (conjugacy class) at  $x$  relative to  $\mathbf{F}_q$ ; this is not a difficulty, as it amounts to replacing  $e(a/p)$  by  $e(-a/p)$  to recover the original sum after summing the traces of geometric Frobenius.

— More importantly, the group  $\Pi_1$  is an infinite topological group, and one must only consider continuous homomorphisms to  $GL(K)$ , with respect to some topology; but moreover, the topology is “profinite” (meaning that there are many finite quotients and kernels of surjections to finite groups are a basis of neighborhoods of the identity) which implies that, e.g., all maps to  $GL(n, \mathbf{C})$  have finite image. This is not such a problem here (since  $\psi$  takes values in the group of  $p$ -th roots of unity), but becomes one when working with more complicated situations, as we will do below. This is bypassed by using for  $K$  the field  $\mathbf{Q}_\ell$  of  $\ell$ -adic numbers, or an/its algebraic closure: there are plenty of interesting representations of  $\Pi_1$  on  $\mathbf{Q}_\ell$ -vector fields, with infinite image, and we will

<sup>5</sup> And in precise technical terms depends on the choice of an algebraic closure of  $\mathbf{F}_q$ .

<sup>6</sup> In the same way that given a tower  $K \subset L_1 \subset L_2$  of fields, where  $L_2/K$  is a Galois extension, the restriction to  $L_1$  leads to a surjective homomorphism  $\text{Gal}(L_2/K) \rightarrow \text{Gal}(L_1/K)$ .

indicate (without full details!) some of them below. One intuitive idea to keep in mind about  $\mathbf{Q}_\ell$  is the following: it is of characteristic 0, contains a ring  $\mathbf{Z}_\ell$ , itself containing  $\mathbf{Z}$ , and is generated by  $\ell^{-1}$  together with  $\mathbf{Z}_\ell$ , and moreover for any  $k \geq 1$ , we can “reduce modulo  $\ell^k$ ” as on  $\mathbf{Z}$ , i.e., we have surjective maps

$$\mathbf{Z}_\ell \rightarrow \mathbf{Z}_\ell / \ell^k \mathbf{Z}_\ell \simeq \mathbf{Z} / \ell^k \mathbf{Z}.$$

On the other hand, for any prime  $s \neq \ell$ , we have  $\mathbf{Z}_\ell / s \mathbf{Z}_\ell = 0$ .

Now, to start anew: a *lisse  $\ell$ -adic sheaf of rank  $r$*  on  $U$  (simply called a sheaf in the following, and denoted usually with curly letters) boils down exactly to a continuous group homomorphism

$$\Pi_1 \rightarrow GL(r, \bar{\mathbf{Q}}_\ell)$$

for some  $\ell \neq p$ . Note that, e.g., the  $p$ -th roots of unity lie in  $\bar{\mathbf{Q}}_\ell$ , so the map  $\psi$  above could equivalently be seen as defined as

$$\Pi_1 \rightarrow GL(1, \bar{\mathbf{Q}}_\ell)$$

for any  $\ell \neq p$ . An *algebraic exponential sum* associated to a sheaf  $\mathcal{F}$  over  $U$  is the data of the sums

$$S_\nu(\mathcal{F}) = \sum_{x \in U(\mathbf{F}_{q^\nu})} \text{Tr}(\text{Fr}_{x, q^\nu} | \mathcal{F})$$

where we abuse notation by denoting  $\text{Tr}(\text{Fr}_{x, q^\nu} | \mathcal{F})$  the trace of the image of the geometric Frobenius  $\text{Fr}_{x, q^\nu}$  acting on the vector space which “is”  $\mathcal{F}$ .

Those sums are packaged in the generating series

$$L(\mathcal{F}, T) = \exp\left(\sum_{\nu \geq 1} \frac{S_\nu(\mathcal{F})}{\nu} T^\nu\right)$$

about which one must be careful to notice that it belongs to  $\bar{\mathbf{Q}}_\ell[T]$  *a priori*.

**Example 2.** Besides the additive example discussed previously, consider the most trivial instance of this: take  $\rho$  to be the  $\ell$ -adic sheaf of degree 1 mapping all of  $\Pi_1$  to  $1 \in GL(1, \mathbf{Q}_\ell)$ . This is called the *trivial sheaf* and is denoted (confusingly)  $\mathcal{F} = \mathbf{Q}_\ell$ . In this case, the  $L$ -function is the zeta function of the variety  $U$ , counting points in extensions fields.

*Remark 3.* There is one important potential problem in working with sums in this generality: since the target of  $\rho$  is not a vector space over the complex numbers, the trace of local Frobenius live in the field  $\bar{\mathbf{Q}}_\ell$ , and hence *their modulus does not make sense*. We will see how this is finessed in the next section for the most important applications.<sup>7</sup>

One of the important points of the formalism is that representations are very flexible objects: one can combine representations in various ways to obtain new ones, define homomorphisms, etc, in a way hardly distinguishable from ordinary linear and multilinear algebra. For instance, the *direct sum* of two sheaves<sup>8</sup> simply acts in the obvious way on the direct sum of the associated vector spaces. Similarly, there is a tensor product of representations and a dual (also called the *contragredient*). A *homomorphism*  $\mathcal{F}_1 \rightarrow \mathcal{F}_2$  is simply a linear map  $\varphi : W_1 \rightarrow W_2$  such that

$$\varphi(\rho_1(g) \cdot w_1) = \rho_2(g) \cdot \varphi(w_1)$$

<sup>7</sup> This point led Deligne to some amusing play with the choice of an isomorphism  $\iota$  between  $\bar{\mathbf{Q}}_\ell$  and  $\mathbf{C}$ , which depends on the axiom of choice...

<sup>8</sup> Associated to the same prime  $\ell \neq p$ .

for  $g \in \Pi_1$ ,  $w_1 \in W_1$ . We then have kernels, surjective and injective homomorphisms, and hence isomorphisms. All this is *essential* to proving the wonderful results which now follow...

**2.3. Cohomology, weights, and the Riemann Hypothesis.** The previous section, by itself, looks remarkably close to abstract nonsense. However, some remarkable theorems prove that something truly amazing is going on; those are undoubtedly among the greatest achievements of 20th century mathematics. Specifically, the first main theorem shows that the rationality of zeta functions of curves over finite fields (or of  $L$ -functions for Weil-type exponential sums) extends to this setting.

**Theorem 4.** *For any algebraic exponential sum (associated to some lisse  $\ell$ -adic sheaf  $\mathcal{F}$ ), the  $L$ -function  $L(\mathcal{F}, T)$  is a rational function, in  $\bar{\mathbf{Q}}_\ell[T]$ , of the form*

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

for some integer  $d \geq 0$ , which is the dimension of  $\bar{U}$  as an algebraic variety (roughly speaking, this is the number of variables in the exponential sum), and for some polynomial  $P_i(\mathcal{F}, T) \in \bar{\mathbf{Q}}_\ell[T]$  such that  $P_i(T, 0) = 1$ .

When dealing with the trivial sheaf (i.e., counting points) on a smooth projective variety, this was first proved by Dwork, using different interpretations and methods. From the point of view of analytic number theory, what this result states is that, for any  $\nu \geq 1$ , we can write the algebraic exponential sum  $S_\nu(\mathcal{F})$  as

$$(7) \quad S_\nu(\mathcal{F}) = \sum_{i=0}^{2d} (-1)^i \sum_{j=1}^{\deg(P_i)} \omega_{i,j}^\nu$$

where the  $\omega_{i,j}$  are the inverse of the roots of  $P_i$ , i.e., we have

$$P_i(\mathcal{F}, T) = \prod_{1 \leq j \leq \deg(P_i)} (1 - \omega_{i,j} T).$$

This is already quite remarkable, but useless for analytic applications if we do not know more... (In particular, at the moment the polynomials do not seem to be uniquely determined...)

– More information requires some assumption on the exponential sum, which are natural from the point of view of wishing to estimate the size of  $|S_\nu(\mathcal{F})|$ , since we explained that this does not make sense a priori. Following Deligne, say that  $\mathcal{F}$  is *pointwise pure of weight  $k$* , for some integer  $k \in \mathbf{Z}$ , if for every  $\nu \geq 1$ , and every  $x \in U(\mathbf{F}_{q^\nu})$ , every eigenvalue  $\alpha$  of  $\rho(\text{Fr}_{x,q^\nu})$  acting on  $\mathcal{F}$  is an *algebraic number*, and every conjugate  $\sigma(\alpha)$  of  $\alpha$  in  $\mathbf{C}$  satisfies (since this makes sense!)

$$|\sigma(\alpha)| = q^{\nu k/2}.$$

As a basic example, note that if the eigenvalues of  $\rho(\text{Fr}_{x,q^\nu})$  are all roots of unity, then it is pointwise pure of weight 0; this of course applies to the type of (rank 1) sheaves constructed for exponential sums with additive and/or multiplicative characters. Later on, we will see more sophisticated examples where the weight can be different from 0, but for the moment let us note that if this assumption holds for a sheaf  $\mathcal{F}$ , the sums  $S_\nu$  are themselves algebraic numbers, hence it makes sense to speak of “the” modulus  $|S_\nu(\mathcal{F})|$ , if we define it to be

$$|\alpha| = \max_{\sigma} |\sigma(\alpha)|$$

for any algebraic number  $\alpha$ ,  $\sigma$  running over the embeddings of  $\mathbf{Q}(\alpha)$  in  $\mathbf{C}$ .

Then we have Deligne's extraordinary generalization of the Riemann Hypothesis for curves:

**Theorem 5.** *For any algebraic exponential sum over  $\mathbf{F}_q$  (associated to some lisse  $\ell$ -adic sheaf  $\mathcal{F}$ ) which is pointwise pure of weight  $k \in \mathbf{Z}$ , and for any  $i$ ,  $0 \leq i \leq 2d$ , the  $\omega_{i,j}$  which are the inverse of the roots of  $P_i(\mathcal{F}, T)$  are algebraic integers mixed of weight  $\leq k + i$ , in the sense that for any  $j$ ,  $\omega_{i,j}$  has the property that there exists  $k_{i,j} \leq k + i$ , which may depend on  $j$ , such that*

$$|\sigma(\omega_{i,j})| = q^{k_{i,j}/2}.$$

The corollary of this is that for an algebraic exponential sum associated to a pointwise pure sheaf of weight  $k$ , we have

$$|S_\nu(\mathcal{F})| \leq Bq^{\delta\nu/2}$$

for all  $\nu \geq 1$ , where

$$(8) \quad B = \sum_{0 \leq i \leq 2d} \deg(P_i(\mathcal{F}, T)),$$

$$(9) \quad \delta = \max_{0 \leq i \leq 2d} \max_{1 \leq j \leq \deg(P_i)} k_{i,j}.$$

Remarkably enough, this is *still* useless, as is, for most practical applications! In other words, no source of cancellation in exponential sums is yet revealed. Indeed, consider a sheaf which is pointwise of weight 0; note that  $d$  being the number of variables in the sum, we can expect about  $q^{\nu d}$  points of summation in  $S_\nu(\mathcal{F})$ , which coincides with the trivial bound for  $\delta$ , using  $k_{i,j} \leq i \leq 2d$  in this situation. This is not surprising: recall that we can take the trivial sheaf, which means exactly counting points on  $U$ :

$$S_\nu(\bar{\mathbf{Q}}_\ell) = |U(\mathbf{F}_{q^\nu})|$$

and we obtain from the theorem above

$$U(\mathbf{F}_{q^\nu}) \ll q^{d\nu} \quad \text{for } \nu \geq 1.$$

Another point is that in most applications to analytic number theory, the sums genuinely depend on  $p$ , and it is the variation with  $p$  (called the "horizontal" direction, by opposition with the "vertical" direction  $\nu \rightarrow +\infty$  for fixed  $p$ ) which is most crucial, e.g., to prove Weil's bound for Kloosterman sums! But then it might be the case that, even if  $\delta$  is small, giving cancellation in the vertical direction, the constant  $B$  depends on  $p$  in such a way as to make the bound above useless when  $p$  varies.

We will come back to applications and explanations of the source of possible cancellation in the next section. However, we will first describe those two results somewhat more precisely, explaining where the polynomials  $P_i$  come from. This is where the dreaded cohomology enters...

Yet, formally speaking this is quite easy to describe. The work of the Grothendieck school of algebraic geometry produced (among other things) a way to assign, to any lisse  $\ell$ -adic sheaf  $\mathcal{F}$  on  $U$  defined over  $\mathbf{F}_q$ , a sequence of finite-dimensional vector spaces, denoted

$$H_c^i(\bar{U}, \mathcal{F})$$

together with an action of the *global* (geometric) Frobenius automorphism  $\text{Fr}$  of  $\bar{U}$  such that  $H_c^i = 0$  for all  $i > 2d$  and

$$(10) \quad P_i(\mathcal{F}, T) = \det(1 - TF \mid H_c^i(\bar{U}, \mathcal{F})).$$

This is a remarkable fact. Note in particular the following: the zeros (and poles) of the  $L$ -function  $L(\mathcal{F}, T)$  are revealed to be the (inverse of the) eigenvalues of some operator on some vector space, and from this springs much of the speculation and current philosophizing concerning Random Matrices and zeros of  $L$ -functions.

The vector space  $H_c^i(\bar{U}, \mathcal{F})$  is called the  $i$ -th compactly supported cohomology group of  $\bar{U}$  with coefficients in  $\mathcal{F}$ . There are again some subtleties to note. First of all, again, those are vector spaces over the  $\ell$ -adic field  $\mathbf{Q}_\ell$ , not over  $\mathbf{C}$ . Also, it is sometimes important to be aware of the fact that the cohomology groups depend on  $\bar{U}$  only; practically speaking, this means that if  $U$  and  $U'$  are algebraic varieties which are in bijection using polynomial maps with coefficients in  $\bar{\mathbf{F}}_q$ , then the cohomology groups are identical as vector spaces; however, the Frobenius  $\text{Fr}$  is not the same. For instance, consider the two curves

$$y^2 = x^3 - x, \quad -y^2 = x^3 - x$$

for  $p$  a prime congruent to 3 mod 4; they are “geometrically” isomorphic using the map  $(x, y) \mapsto (x, iy)$  where  $i \in \mathbf{F}_{p^2}$  is a square root of  $-1$ , but of course they usually do not have the same number of points over  $\mathbf{F}_q$  itself.

The formula (10) together with the rationality and the definition of  $\omega_{i,j}$  means that we have

$$S_\nu(\mathbf{F}_{q^\nu}) = \sum_{i=0}^{2d} (-1)^i \text{Tr}(F^\nu \mid H_c^i(\bar{U}, \mathcal{F})).$$

This beautiful formula is called the *Grothendieck-Lefschetz Trace Formula*. In fact, it is easy to check that it is equivalent with Theorem 4, and this is the way the latter was proved by Grothendieck.

The proof of Theorem 5 defies a simple explanation. My own understanding of the first proof by Deligne (which only covered the counting problem for points over smooth projective curves) is described in the old notes [Ko1]. Some comments are however in order. First of all, everyone may have his own opinions, but to my mind this is the deepest result of arithmetic in the twentieth-century. Secondly, it certainly deserves much more attention from the analytic number theory community. Young analytic number theorists must put themselves in the position of continuing the use and understanding of Deligne’s work in the future. This looks difficult, but I believe the intellectual investment would be well rewarded – not only in gaining insight in such beautiful mathematics, but also in proving new results yet unsuspected. And the earlier one starts trying to gain this understanding, the more likely it is that patience and hard work will lead to it. Crucially, I believe that despite the intrinsic difficulties involved, one can penetrate layer by layer into the heart of the subject, starting with a big black box and opening it to find a smaller one, and then on... And finally, the link with analytic number theory is *not* a one-way street: analytic number theory can bring new results and new intuition in the algebraic theory. Indeed, it already has! Deligne’s first proof relied crucially on the use of an analogue of the Rankin-Selberg convolution of modular forms (precisely, Deligne acknowledges the influence of the paper of Rankin where the first non-trivial estimate for the Ramanujan  $\tau$ -function was achieved; repaying the debt, it is from his work that the best possible bound for the latter was deduced...) Deligne’s second proof (the one which led to Theorem 5 for arbitrary  $\mathcal{F}$ ) uses instead the classical method of Hadamard and de la Vallée Poussin for proving that some  $L$ -functions do not vanish on the line at the edge of the critical line.

This leads naturally to a discussion of why this is called the Riemann Hypothesis over finite fields. Assume that the polynomials  $P_i$  have coefficients in  $\mathbf{Z}$  (or in a number field),

which is the case in many situations. Then we can define

$$\mathcal{L}(\mathcal{F}, s) = L(\mathcal{F}, q^{-s})$$

for  $s \in \mathbf{C}$ ; this is a meromorphic function of  $s$ . Its zeros (resp. poles) satisfy  $q^{-s} = \omega_{i,j}^{-1}$  for  $i$  even (resp.  $i$  odd), so they are of the form

$$s = \frac{k_{i,j}}{2} + i \frac{t_{i,j} + 2k\pi}{\log q}$$

where  $k \in \mathbf{Z}$  and  $t_{i,j}$  is a complex number such that

$$e^{it_{i,j}} = \omega_{i,j} q^{-k_{i,j}/2}.$$

So the zeros and poles are on vertical lines! However, those lines can have different real parts. And indeed, such situations (i.e., “mixed” cohomology) do arise quite frequently, and only in particular cases is it possible to combine Deligne’s theorem with duality results that imply, e.g., to the eigenvalues of  $F$  on  $H_c^i$  are “pure” of weight  $k + i$ .

**2.4. Basic results on cohomology groups.** Before continuing with examples, we need to say a few words on the simplest cases of computations of cohomology groups and on what is known about the crucial constant  $B$  in (8); the matter of (9) will be discussed in the next section. Those correspond to  $i = 0$  and  $i = 2d$  (the extremities of the range  $0 \leq i \leq 2d$ ; note that for  $d = 1$ , this leaves only  $H_c^1$  to be understood). In both cases, we need to introduce the *geometric fundamental group*  $\bar{\Pi}_1$  of  $\bar{U}$ , which is a normal subgroup of  $\Pi_1$  with the property that the quotient  $\Pi_1/\bar{\Pi}_1$  is (algebraically) the Galois group of  $\bar{\mathbf{F}}_q/\mathbf{F}_q$ , and therefore it is (topologically) generated a single element, which is none other than the Frobenius automorphism of  $\mathbf{F}_q$ .

— For  $i = 0$ , the space  $H_c^0(\bar{U}, \mathcal{F})$  is 0 except when  $U$  is projective. In the latter case, if we denote by  $\bar{\Pi}_1$  the normal subgroup of  $\Pi_1$  which is the *geometric fundamental group* of  $U$ , i.e., the fundamental group of  $\bar{\Pi}_1$ , then  $H_c^0(\bar{U}, \mathcal{F})$  is canonically isomorphic with the space of vectors in the vector space defining  $\mathcal{F}$  invariant under  $\bar{\Pi}_1$ . This last case does not occur for typical exponential sums over varieties since  $U$  is then not projective.

— For  $i = 2d$ , there is a duality result; assuming that  $\bar{U}$  is irreducible and smooth (e.g.,  $U$  is defined as the locus of points where a single polynomial in  $d$  variables is non-zero), then  $H_c^{2d}(\bar{U}, \mathcal{F})$  is the co-invariant space of the vector space  $V$  defining  $\mathcal{F}$ , but the action of the global Frobenius on this space is  $q^d$  times the natural action of  $F$  on the co-invariant space. To see how this action works, we need only remark the following general fact: if  $H$  is a normal subgroup of a group  $G$ , then for any vector space  $V$  on which  $G$  acts, the invariant space and the co-invariant space, defined by

$$V^H = \{v \in V \mid h \cdot v = v \text{ for all } h \in H\}$$

$$V_H = V/W, \text{ where } W \text{ is generated by } h \cdot v - v, \text{ for } v \in V, h \in H$$

are stable under the action of the whole group  $G$ ; indeed, taking the case of  $V_H$  for instance, we check that  $g \cdot v$  is well-defined in  $V_H$  for  $g \in G$ , since

$$g \cdot (h \cdot w - w) = (ghg^{-1})(g \cdot v - g \cdot v),$$

and  $ghg^{-1} \in H$ .

In many cases of interest, the global Frobenius actually acts trivially on  $\mathcal{F}^G$  and  $\mathcal{F}_G$ . In particular if  $\mathcal{F}$  is trival, this is so, and then we see that if  $U$  is irreducible, and not projective, we have

$$\begin{aligned} \text{Tr}(F^\nu \mid H_c^0(\bar{U}, \bar{\mathbf{Q}}_\ell)) &= 0, \\ \text{Tr}(F^\nu \mid H_c^{2d}(\bar{U}, \bar{\mathbf{Q}}_\ell)) &= q^d. \end{aligned}$$

This, by the Riemann Hypothesis, leads to

$$|U(\mathbf{F}_{q^\nu})| \sim q^{d\nu}$$

as  $\nu \rightarrow +\infty$  for any irreducible algebraic variety of dimension  $d$  over  $\mathbf{F}_q$ .

As another particular case, consider an exponential sum with additive character over  $U/\mathbf{F}_q$ . Then the space on which the corresponding  $\rho$  acts is one-dimensional. We find therefore that  $H_c^{2d}$  is zero in this case, except when the subgroup  $\bar{\Pi}_1$  acts trivially. What does the latter condition mean? Algebraically, this means that the action of some element  $g \in \Pi_1$  is determined by its class modulo  $\bar{\Pi}_1$ , which we have mentioned is generated by the Frobenius  $F$  over  $\mathbf{F}_q$ . This class, when  $g$  is of the form  $\text{Fr}_{x,q^\nu}$ , is  $F^{-\nu}$  (the minus sign comes from the fact that  $F^{-1}$  is the geometric Frobenius). If one thinks about this means, it is simply the expression of the fact that the phase  $e(\text{Tr}(f(x))/p)$  depends only on the degree  $\nu$ ; in particular for  $\nu = 1$ , this means the phase is constant. In this case, there is trivially no cancellation. So we have the following preliminary result:

**Proposition 6.** *Let  $U/\mathbf{F}_q$  be a geometrically irreducible smooth algebraic variety of dimension  $d$  over a finite field, let*

$$S_\nu = \sum_{x \in U(\mathbf{F}_{q^\nu})} e\left(\frac{\text{Tr}(f(x))}{p}\right)$$

*be a system of exponential sums with additive characters where  $f$  is a polynomial function on  $U/\mathbf{F}_q$ . Then there exists a constant  $B$  such that*

$$|S_\nu| \leq Bq^{d\nu-1/2}$$

*for  $\nu \geq 1$ , except when for all  $\nu \geq 1$ , the phase  $e(\text{Tr}(f(x))/p)$  is independent of  $x \in U(\mathbf{F}_{q^\nu})$ .*

This proposition is again only of interest in the direction where  $\nu$  grows. Let us look at this once more from the point of view of  $U$  defined over  $\mathbf{Z}$  (by vanishing and non-vanishing of polynomial with integral coefficients), with  $f$  defined as polynomial with integer coefficients. Assume  $U$  is smooth and irreducible over  $\mathbf{C}$  (e.g.,  $U$  defined by non-vanishing of a single non-zero polynomial), so that for all but finitely many primes  $p$ , we are in the situation of the proposition when  $U$  is looked at as defined over  $\mathbf{F}_p$ . The condition that  $e(f(x)/p)$  is independent of  $x \in U(\mathbf{F}_p)$ , i.e.,  $f(x)$  is constant modulo  $p$ . However, another special case of Deligne's result implies that for any non-constant polynomial  $f$ , we have

$$|\{x \in U(\mathbf{F}_p) \mid f(x) = \alpha\}| = p^{d-1} + O(p^{d-1/2})$$

and this is uniform with respect to  $\alpha$ . Hence, taking  $p$  large enough, we see that the non-cancellation condition can only hold if  $f$  was constant as a polynomial.

To deduce cancellation, and indeed to prove the estimate above uniformly with respect to  $p$ , we need to ensure that uniformity of the constant  $B$  in the proposition with respect to  $p$ . This crucial fact for analytic number theorists is (for additive character sums) a theorem of Bombieri, which was much generalized by Adolphson and Sperber, and then by Katz. We quote a version of what is proved by Katz, which is very general.

**Proposition 7 (Katz).** *For any algebraic exponential sum associated to sums of the type*

$$\sum_{x \in U(\mathbf{F}_q)} e\left(\frac{\text{Tr}(f(x))}{p}\right) \prod_{j=1}^s \chi_j(Ng_j(x))$$

where  $U$  is defined by the vanishing of  $\leq B$  polynomials in  $N$  variables of degree  $\leq D$ , and the non-vanishing of  $g_1, \dots, g_s$ , which are polynomials of degree  $\leq D'$ , and where  $f$  is a polynomial of degree  $D''$ , we have

$$\sum_i \dim H_c^i \leq 3(s + 2 + D + D' + sD'')^{N+B}.$$

See [K2] for the details.

**2.5. Sample applications.** In this section we will describe a few applications of the formalism of sheaves and cohomology to fairly classical exponential sums of interest in analytic number theory (i.e., corresponding to sheaves of pointwise pure of weight 0). The next section will discuss applications where the weight is  $> 0$ .

— Deligne himself proved the following general estimate for exponential sums involving additive characters, in many variables, of a special type: let  $f \in \mathbf{Z}[x_1, \dots, x_m]$  be a non-zero polynomial of degree  $d$ , and assume that the hypersurface

$$H_f : f_d(x_1, \dots, x_m) = 0$$

defined by the homogeneous component of highest degree terms is smooth (in other words, there is no common zero of  $f_d$  and all its partial derivatives, except for  $(0, \dots, 0)$ ). Then we have

$$\left| \sum_{x_1, \dots, x_m \in \mathbf{F}_{p^\nu}} \dots \sum e\left(\frac{\text{Tr}(f(x_1, \dots, x_m))}{p}\right) \right| \leq (d-1)^m q^{m\nu/2}$$

for all prime  $p$  such that the reduction of  $H_f$  modulo  $p$  remains smooth, and for all  $\nu \geq 1$ .

As a sample: we have for any  $m \geq 2$ , any  $d \geq 1$ , any prime  $p \nmid d$ , the estimate

$$\left| \sum_{x_1, \dots, x_m \pmod{p}} \dots \sum e\left(\frac{x_1^d + \dots + x_m^d + x_1 \dots x_m}{p}\right) \right| \leq (d-1)^m p^{m/2}.$$

Indeed the partial derivatives of the term of degree  $d$  are  $dx_i^{d-1}$ , hence vanish modulo  $p$  only for  $x_i = 0$  if  $p \nmid d$ . Notice that here we obtain squareroot cancellation in the sum, and this does not depend at all on the lower order terms (the latter property is reminiscent of features in the methods of Weyl, van der Corput and Vinogradov).

Deligne proved this twice; the first proof was geometrically delicate, based on the theory as it existed for smooth projective varieties and the trivial sheaf only, and the second exploited the whole formalism with lisse sheaves. Indeed, Deligne proved the following results for the associated sheaf (say  $\mathcal{F}_f$ , on the affine  $m$ -space  $\mathbf{A}^m/\mathbf{F}_q$ , such that  $\mathbf{A}^m(\mathbf{F}_{q^\nu}) = \mathbf{F}_{q^\nu}^m$  for all  $\nu$ ): (1)  $H_c^i = 0$  except for  $i = m$ ; (2)  $H_c^m$  is pure of weight  $m$ , i.e., all eigenvalues of  $F$  are algebraic integers with conjugates of modulus  $q^{\mu/2}$ ; (3) the dimension of  $H_c^m$  is  $(d-1)^m$ . The last result explains the independence of the constant from lower order terms, and ultimately can be seen as an expression of the fact that this dimension is a “topological” invariant, and can be computed after deformation and specialization to any fixed Deligne polynomial. Now, notice that if  $f = x_1^d + \dots + x_m^d$ , the sum (over extension fields) factors as

$$\left( \sum_{x \in \mathbf{F}_{p^\nu}} e\left(\frac{\text{Tr}(x^d)}{p}\right) \right)^m$$

and it is a consequence of works on 1-dimensional sums (e.g., Weil’s earlier results!) that this is bounded by  $(d-1)^m q^{\nu/2}$ , which implies the result of Deligne.

Even better, this sum can be rewritten in terms of Gauss sums using characters of order  $d$  to detect the condition  $x^d = 1$ ; we can indeed assume that  $d \mid p - 1$  since  $x \mapsto x^d$  is bijective if  $(d, p - 1) = 1$ , and then we find that

$$\sum_{x \in \mathbf{F}_{p^\nu}} e\left(\frac{\mathrm{Tr}(x^d)}{p}\right) = \sum_{\chi^d=1} \sum_{y \in \mathbf{F}_{p^\nu}} \chi(y) e\left(\frac{\mathrm{Tr}(y)}{p}\right)$$

(remembering that  $x \mapsto x^d$  is  $d$ -to-one if  $d \mid p - 1$ ). Since all Gauss sums with nontrivial character are of modulus  $p^{\nu/2}$  and the Gauss sum with trivial character vanishes, we obtain

$$\left| \sum_{x \in \mathbf{F}_{p^\nu}} e\left(\frac{\mathrm{Tr}(x^d)}{p}\right) \right| \leq (d - 1)p^{\nu/2}.$$

Hence we see here a case where having a bound for a single (very easy!) sum in a “family” with many parameters suffices to derive bounds in all cases. This is a typical feature of the general theory, which will be (somewhat) clearer in the next section.

— In the case of hyper-Kloosterman sums, in addition to the proof by Deligne of the estimate (4) (in the form

$$\left| \sum \cdots \sum_{\substack{(x_1, \dots, x_h) \in \mathbf{F}_q^\times \\ x_1 \cdots x_h = a}} e\left(\frac{\mathrm{Tr}(x_1 + \cdots + x_h)}{p}\right) \right| \leq hq^{(h-1)/2}$$

for extension fields), a clever trick of Bombieri exploits the general formalism with some analytic ideas to prove in much simpler way the bound

$$S_h(a; p) \leq Bp^{(h-1)/2}$$

for  $p$  prime, the constant  $B$  being absolute.

Instead of presenting this case (the details are explained in [IK, p. 308–309]), we will treat in a similar way a two-variable exponential sum worked out by L. Rozenzweig (after discussions with P. Kurlberg and myself) during the ICTP conference.

The sum in question is  $L(\chi_1, \chi_2, 1; p)$ , where we denote

$$L(\chi_1, \chi_2, a; q) = \sum_{(x, y, z) \in U(\mathbf{F}_q)} \chi_1(Nx) \chi_2(Ny) e\left(\frac{\mathrm{Tr}(a\psi(x, y, z))}{p}\right)$$

for any prime power  $q = p^f$ ,  $a \in \mathbf{F}_q$  and multiplicative characters  $\chi_1, \chi_2$  modulo  $p$ , where the summation set is

$$U(\mathbf{F}_q) = \{(x, y, z) \in \mathbf{F}_q^\times \mid xyz = 1, (x - 1)(y - 1)(z - 1) \neq 0\}$$

while the rational function  $\psi$  defining the phase is

$$\psi(x, y, z) = \frac{1 + x}{1 - x} - \frac{1 + y}{1 - y} + \frac{1 + z}{1 - z}.$$

Note this rational function is well-defined on  $\bar{U}$ . The introduction of the parameter  $a$  is crucial, and it anticipates in a simple way the topics of the next section.

It is clear that  $U$  is of dimension 2; the general formalism and Deligne’s theorem therefore allow us to assert that there exist  $\omega_{i,j}(\chi_1, \chi_2, a)$  which satisfy  $|\omega_{i,j}(\chi_1, \chi_2, a)| \leq p^{i/2}$  such that for  $q = p^\nu$  we have

$$L(\chi_1, \chi_2, a; q) = \sum_{0 \leq i \leq 3} \sum_j \omega_{i,j}(\chi_1, \chi_2, a)^\nu$$

and moreover total number of  $\omega_{i,j}(\chi_1, \chi_2, a)$  is bounded uniformly by Proposition 7.

Clearly  $\psi$  is not constant on  $U(\mathbf{F}_p)$  so we let  $i$  run only up to  $3 = 2d - 1$  in the above. We target square-root cancellation, which amounts to proving that  $|\omega_{i,j}(\chi_1, \chi_2, a)| \leq p$  for all  $i, j$ . (This will be true in particular if  $H_c^3 = 0$ , but we will not actually confirm this; recall that  $H_c^3$  might contain eigenvalues of smaller weight).

The trick is now to compute a mean-square average  $M_2$  over  $a \in \mathbf{F}_q$ . Namely, we define and compute as follows:

$$\begin{aligned} M_2 &= \sum_{a \in \mathbf{F}_q} |L(\chi_1, \chi_2, a; q)|^2 \\ &= \sum_a \sum_{x,y,z} \sum_{\alpha,\beta,\gamma} \chi_1(Nx) \bar{\chi}_1(N\alpha) \chi_2(Ny) \bar{\chi}_2(N\beta) e\left(\frac{1}{p} \operatorname{Tr}\{a(\psi(x,y,z) - \psi(\alpha,\beta,\gamma))\}\right) \end{aligned}$$

obtaining by orthogonality (after summing first over  $a$ ) the formula

$$\begin{aligned} M_2 &= q \sum_{\substack{x,y,z,\alpha,\beta,\gamma \\ \psi(x,y,z)=\psi(\alpha,\beta,\gamma)}} \chi_1(Nx) \bar{\chi}_1(N\alpha) \chi_2(Ny) \bar{\chi}_2(N\beta) \\ &= q \sum_{c \in \mathbf{F}_q} \left| \sum_{\substack{(x,y,z) \in U(\mathbf{F}_q) \\ \psi(x,y,z)=c}} \chi_1(x) \chi_2(y) \right|^2. \end{aligned}$$

Now the inner sums are one-variable sums because there are two equations ( $xyz = 1$  and  $\psi(x,y,z) = c$ ), and it is not hard to deduce that each is bounded by  $Cq^{1/2}$  for some constant  $C \geq 0$  which is independent of  $c$  and  $q$ . Hence it follows that

$$M_2 \leq Cq^3.$$

On the other hand, assume some  $\omega_{i,j}(\chi_1, \chi_2, 1)$ , for some prime  $p$ , has modulus  $p^{3/2}$ . Then applying an automorphism of  $\mathbf{C}$  that maps  $e(1/p)$  to  $e(a/p)$  for any  $a \in \mathbf{F}_p^\times$ , and leaves the values of  $\chi_1$  and  $\chi_2$  invariant (and hence maps  $L(\chi_1, \chi_2, 1; p^\nu)$  to  $L(\chi_1, \chi_2, a; p^\nu)$  for all  $\nu \geq 1$ ) we see that

$$\limsup_{\nu \rightarrow +\infty} \frac{L(\chi_1, \chi_2, a; p^\nu)}{p^{3\nu/2}} > 0.$$

It is then not hard to show that each of these  $a \in \mathbf{F}_p^\times$  will contribute simultaneously to  $M_2$  for  $q = p^\nu$  in the limit  $\nu \rightarrow +\infty$ , so that

$$M_2 \gg (p-1)p^{3\nu/2}$$

for infinitely many  $\nu$ . The upper and lower bounds on  $M_2$  being contradictory if  $p$  is big enough, we conclude that the “bad”  $\omega_{i,j}(\chi_1, \chi_2, 1)$  could not exist (if  $p$  is large enough), and hence

$$L(\chi_1, \chi_2, 1; p) \ll p^2$$

for  $p \geq 2$  (with absolute implied constant).

## REFERENCES

- [IK] H. Iwaniec and E. Kowalski: *Analytic Number Theory*, A.M.S Colloquium Publications, vol 53 (2004).
- [K1] N. Katz: *Gauss sums, Kloosterman sums and monodromy groups*, Princeton Univ. Press (1988).
- [K2] N. Katz: *Sums of Betti numbers in arbitrary characteristic*, Finite Fields Appl. 7 (2001), no. 1, 29–44.
- [KS1] N. Katz and P. Sarnak: *Random matrices, Frobenius eigenvalues, and monodromy*, A.M.S Colloquium Publications, vol. 45 (1999).

- [KS2] N. Katz and P. Sarnak: *Zeroes of zeta functions and symmetry*, Bull. Amer. Math. Soc. 36 (1999), 1–26.
- [K01] E. Kowalski: *Deligne’s proof of the Weil conjectures for varieties over finite fields*, available online at [www.math.u-bordeaux1.fr/~kowalski/deligne.pdf](http://www.math.u-bordeaux1.fr/~kowalski/deligne.pdf)

UNIVERSITÉ BORDEAUX I – IMB, UMR 5251, 351, COURS DE LA LIBÉRATION, 33405 TALENCE  
CEDEX, FRANCE

*E-mail address:* [emmanuel.kowalski@math.u-bordeaux1.fr](mailto:emmanuel.kowalski@math.u-bordeaux1.fr)