A SURVEY OF ALGEBRAIC EXPONENTIAL SUMS AND SOME APPLICATIONS

E. KOWALSKI

1. Introduction

This survey is a written and slightly expanded version of the talk given at the ICMS workshop on motivic integration and its interactions with model theory and non-archimedean geometry. Its presence may seem to require a few preliminary words of explanation: not only does the title apparently fail to reflect any of the three components of that of the conference itself, but also the author is far from being an expert in any of these. However, one must remember that there is but a small step from summation to integration. Moreover, as I will argue in the last section, there are some basic problems in the theory of exponential sums (and their applications) for which it seems not impossible that logical ideas could be useful, and hence presenting the context to model-theorists in particular could well be useful. In fact, the most direct connection between exponential sums and the topics of the workshop will be a survey of the extension to exponential sums of the beautiful counting results of Chatzidakis, van den Dries and McIntyre's [CDM]. These may also have some further applications.

Acknowledgements. I wish to thank warmly the organizers of the workshop for preparing a particularly rich and inspiring program, and for inviting me to participate. I also thank B. Conrad for pointing out a potentially misleading statement related to Poincaré duality on page 10.

2. Where exponential sums come from

Exponential sums, in the most general sense, are any type of finite sums of complex numbers

$$S = \sum_{1\leqslant n\leqslant N} e(\theta_n)$$

where we write $e(z) = \exp(2i\pi z)$, as is customary in analytic number theory, and where the phases θ_n are real numbers. Such as sum is trivially bounded by the number of terms

$$|S| \leqslant N,$$

and of course if nothing more is known about θ_n, this can not be improved. However, in applications (whether arithmetic or otherwise), one knows something more, and the goal is very often to go from this to substantial improvements of the trivial bound: one typically wishes to prove

$$|S| \leqslant N\Sigma(N)^{-1}$$

with $\Sigma(N)$ increasing as $N \to +\infty$, as fast as possible. This is interpreted as being the result of substantial *oscillations* of the phases (in $\mathbf{R}/\mathbf{Z}$) which result in the sum being somewhat comparable with a random walk in the plane $\mathbf{C} = \mathbf{R}^2$.

In analytic number theory, exponential sums arise from many different sources. For the purpose of this survey, we will emphasize questions of *equidistribution* as leading naturally to exponential sums, as this will give a motivating framework for all the examples we want to consider here, and leads to problems and results which have obvious interest for all arithmeticians. So we will not speak about, e.g., the direct occurrence of exponential sums in the circle method, or in the distribution of primes (see, e.g., [IK, Ch. 20] for the former, and [IK, Ch. 5, 17–19] for the latter).

We therefore recall the definition of equidistribution, as well as the important Weyl criterion. Let (X, μ) be a compact topological space with a (Borel) probability measure μ. Given finite sequences $(x_n)_{n\leqslant N}$, where $x_n = x_n^{(N)}$ may depend on N, one says that they become μ-equidistributed in X if, for any open set U with $\mu(\partial U) = 0$, the sample counts

$$\frac{1}{N}|\{n \leqslant N \mid x_n \in U\}|$$

converge to the measure $\mu(U)$ of U as N gets large. Equivalently, for any continuous function f on X, the sample average

$$\frac{1}{N}\sum_{n\leqslant N} f(n)$$

is close to the integral of f over X. In the important special case where X is the torus $\mathbf{R}/\mathbf{Z}$ and μ the (unit) Lebesgue measure, this means that

$$\frac{1}{N}|\{n \leqslant N \mid a < x_n < b\}| \rightarrow (b - a)$$

for any a, b with $0 \leqslant a < b \leqslant 1$. The basic criterion of H. Weyl states that there is μ-equidistribution if and only if, for some orthonormal basis (f_h) of $L^2(X, \mu)$, elements of which are continuous functions with $f_0 = 1$, we have

$$\lim_{N\rightarrow+\infty} \frac{1}{N}\sum_{n\leqslant N} f_h(x_n) = 0, \qquad \text{for any fixed } h \neq 0.$$

If we consider again $X = \mathbf{R}/\mathbf{Z}$, we can take $f_h(x) = e(hx)$, and the criterion becomes

$$\lim_{N\rightarrow+\infty} \frac{1}{N} S_h(N) = 0, \qquad \text{where} \qquad S_h(N) = \sum_{n\leqslant N} e(hx_n),$$

which are clearly some sort of exponential sum, and the goal is to exhibit *some* cancellation, for every non-zero *frequency* h.

There are quite a few techniques available to deal with sums of the type

$$\sum_{1\leqslant n\leqslant N} e(f(n))$$

if f is some smooth real-valued function defined on $\mathbf{R}$, leading to many equidistribution statements. However, currently, the sums which are best understood are those of "algebraic nature", where extremely deep techniques of algebraic geometry are available to analyze the sums ("dissect" wouldn't be too strong a word!).

As an example of such algebraic exponential sum, which we will also carry along this survey, we take one of the most important one in analytic number theory, the *Kloosterman sums*. They are also among the first examples to have been considered historically in a non-trivial way.[1]

[1] Though the very first case is probably to be found in Gauss sums.

In keeping with our emphasis, we introduce those sums by means of an equidistribution statement which depends on estimates for Kloosterman sums.[2]

Theorem 1. *For $n \geqslant 1$, let $R_4(n)$ denote the set of 4-tuples $(n_1, \ldots, n_4) \in \mathbf{Z}^4$ such that*

$$n = n_1^2 + \cdots + n_4^2.$$

(1) *We have*

$$|R_4(n)| = 8n \prod_{\substack{p \mid n \\ p \geqslant 3}} \Bigl(1 + \frac{1}{p}\Bigr)$$

for all $n \geqslant 1$, in particular $r_4(n) = |R_4(n)|$ tends to infinity as n does.

(2) *Consider the set of points*

$$\tilde{R}_4(n) = \{x \in \mathbf{R}^4 \mid \|x\| = 1, \text{ and } n^{1/2}x \in R_4(n)\}$$

in the unit 3-sphere $\mathbf{S}^3 \subset \mathbf{R}^4$. Then, as $n \to +\infty$, $\tilde{R}_4(n)$ becomes equidistributed with respect to the Lebesgue measure on $\mathbf{S}^3$, i.e., for any continuous function f on the sphere, we have

$$\lim_{n \to +\infty} \frac{1}{r_4(n)} \sum_{x \in R_4(n)} f\Bigl(\frac{x}{\|x\|}\Bigr) = \int_{\mathbf{S}^3} f(x)dx.$$

It is the proof of Part (2) which requires exponential sums, while (1) is a result going back at least to Jacobi, which can be proved in many different ways (arithmetic of quaternions, theta functions, etc; see, e.g., [HW, §20.12] for an elementary approach).

What is the link with exponential sums? Kloosterman's approach was based on a refinement of the circle method of Hardy-Littlewood-Ramanujan (which remains of great importance today), but this particular problem is probably better understood by an appeal to modular forms. More precisely, Weyl's criterion shows that it is enough to prove that the limit exists and is equal to zero for non-constant spherical (harmonic) polynomials P on the three-sphere, which are the eigenvalues of the (Riemannian) Laplace operator on $\mathbf{S}^3$ (those functions, including the constant 1, are well-known to form an orthonormal basis of $L^2(\mathbf{S}^3)$). Fix one P; then one can form the generating function

$$\theta(z; f) = \sum_{n \geqslant 1} n^{d/2} \Bigl(\sum_{x \in R_4(n)} P(x/\|x\|) \Bigr) e(nz),$$

which exists for $z \in \mathbf{H} = \{z \mid \mathrm{Im}(z) > 0\}$, where d is the degree of f. Using modularity properties, it is possible to express the coefficient of $e(nz)$ as a finite linear combination of those of other functions called Poincaré series. These last coefficients, as shown already by Poincaré himself, are roughly of the form

$$\sum_{c \geqslant 1} \frac{1}{c} S(m, n; c) J(4\pi\sqrt{mn}/c)$$

where m is a non-zero integer, the function J is a Bessel function, and $S(m, n; c)$ is a *Kloosterman sum*:

$$S(m, n; c) = \sum_{\substack{x \,(\mathrm{mod}\, c) \\ (x,c)=1}} e\Bigl(\frac{mx + nx^{-1}}{c}\Bigr),$$

[2] This is not quite what Kloosterman himself did, which involved counting solutions to diagonal quadratic equations in four variables $a_1n_1^2 + \cdots + a_4n_4^2 = N$; however, the analysis which is required is very similar.

where the inverse of x is of course the inverse in the unit group $(\mathbf{Z}/c\mathbf{Z})^\times$. One then uses facts about Bessel functions to verify that the theorem follows from any non-trivial estimate for Kloosterman sums of the type

$$S(m,n;c) \ll c^{1-\delta}$$

for a fixed $\delta > 0$, and for all m, n coprime with c. This is what Kloosterman already proved:

Proposition 2. *Let $c \geqslant 1$ be an integer, let m, $n \in \mathbf{Z}$ be such that the gcd $(m,n,c) = 1$. Then*

$$|S(m,n;c)| < 2c^{3/4}.$$

The main difference between sums like Kloosterman's and general exponential sums is that the range of summation is the set of points of an algebraic variety over a finite ring, and the phases are obtained by evaluating rational functions defined on this variety (in his case, the variety is the multiplicative group, and the rational function is $x \mapsto mx + n/x$). This crucial algebraicity suggests that the theory of more general sums like

$$S(f,V,c) = \sum_{x \in V(\mathbf{Z}/c\mathbf{Z})} e(f(x)/c)$$

should be accessible for general algebraic varieties V (defined over $\mathbf{Z}/c\mathbf{Z}$ or over $\mathbf{Z}$ more simply), and for f an algebraic function on V.

There is indeed such a theory, which in fact splits into two fairly distinct questions. This can be seen from the original example of Kloosterman: his argument started by using the Chinese Remainder Theorem to relate $S(m,n;c_1c_2)$ to $S(m_1,n_1;c_1)$ and $S(m_2,n_2;c_2)$ when c_1 and c_2 are coprime (with m_1, n_1, m_2, n_2 simple rational functions of m, n, c_1, c_2). So Proposition 2 was reduced to the case where $c = p^k$, with $k \geqslant 1$, is the power of a prime number. Now it turns out (and this is a general feature) that the case $c = p$ is very different from $c = p^k$ with $k \geqslant 2$.

Precisely, in the case of Kloosterman sums, one gets *exact formulas* for $k \geqslant 2$ (due to Salié, see, e.g., [I, 4.3]); as an example, if $p \geqslant 3$ is $\equiv 1 \pmod 4$, we have

$$S(1,1;p^k) = 2p^{k/2}\cos\Bigl(\frac{4\pi n}{p^k}\Bigr)$$

if $k \geqslant 2$, and those formulas lead immediately to the desired estimate (with the better exponent $1/2$) without any more work.

Such a drastic simplification is not rare in concrete examples arising in analytic number theory, but it is not to be expected in all cases. Indeed, there is a whole theory of trying to understand the behavior of exponential sums of algebraic origin over $\mathbf{Z}/p^k\mathbf{Z}$ as $k \to +\infty$, to which the name of Igusa is most commonly attached; for recent results along these lines, see for instance the paper [C] of R. Cluckers.

We will now concentrate (almost) exclusively on the case $k = 1$, i.e., on sums over finite fields. For many concrete applications, this turns out to be the most important, and this justifies somewhat this special attention, but we repeat that one should not immediately consider a problem solved when the relevant sums are understood in that case: sometimes, it is higher powers of primes which are most difficult. An example of this is given in the work of Belabas and Fouvry [BF, 3.c].

3. Weil's interpretation of exponential sums over finite fields

Kloosterman's argument to prove Proposition 2 for $c = p$ a prime number was elegant, but not immediately generalizable to more situations (it can be seen in [I, 4.4]). Around

1940, A. Weil, following hints of Davenport and Hasse in particular, realized that the *Riemann Hypothesis* for curves over finite fields led to a stronger result and a better understanding of the underlying problem: namely, using results on the number of points over finite fields of the curves with equations

$$y^p - y = mx + \frac{n}{x}, \tag{1}$$

Weil showed that

$$|S(m,n;p)| \leqslant 2\sqrt{p}$$

for all primes p and m, n coprime with p. In fact, this was based on combining almost trivially two remarkable properties. The first one, already showing the importance of viewing the problem algebraically, is the *rationality* of the generating function

$$Z_{m,n}(T) = \exp\Bigl(\sum_{\nu\geqslant 1} S_\nu(m,n)\frac{T^\nu}{\nu}\Bigr) \in \mathbf{C}[[T]]$$

formed with the analogues of the Kloosterman sum *over extension fields* $\mathbf{F}_{p^\nu}/\mathbf{F}_p$ – not to be mistaken with the finite rings $\mathbf{Z}/p^k\mathbf{Z}$ – defined by

$$S_\nu(m,n) = \sum_{x\in\mathbf{F}_{p^\nu}^\times} e\Bigl(\frac{\mathrm{Tr}_\nu(xm+n/x)}{p}\Bigr)$$

in terms of the trace map $\mathbf{F}_{p^\nu} \xrightarrow{\mathrm{Tr}_\nu} \mathbf{F}_p$. These sums do not (as far as the author knows!) occur naturally in problems of analytic number theory,[3] but the whole family $(S_\nu(m,n))$ really contains the key to the general theory. Indeed, the rationality property already mentioned takes the form of the formula

$$Z_{m,n}(T) = \frac{1}{1 - S_1(m,n)T + pT^2}$$

with (of course) $S_1(m,n) = S(m,n;p)$, the original Kloosterman sum. What this means, in particular, is that the size of $S_1(m,n)$ is dictated precisely by the poles of the generating function, which is not surprisingly called the zeta function associated with the Kloosterman sum. Precisely, factor the quadratic term as

$$1 - S_1(m,n)T + pT^2 = (1-\alpha_{m,n}T)(1-\beta_{m,n}T),$$

so that

$$|S_1(m,n)| = |\alpha_{m,n}+\beta_{m,n}| \leqslant |\alpha_{m,n}| + |\beta_{m,n}|. \tag{2}$$

The next result, due to Weil, then explains his bound for Kloosterman sums: the two (inverse) roots $\alpha_{m,n}$, $\beta_{m,n}$, for m, n coprime with p, satisfy

$$|\alpha_{m,n}| = |\beta_{m,n}| = \sqrt{p}.$$

This is called the Riemann Hypothesis (for Kloosterman sums), because of the following interpretation: if we introduce a complex variable s and take $T = p^{-s}$, the resulting complex function $\zeta_{m,n}(s) = Z_{m,n}(p^{-s})$ is meromorphic and its only poles are situated on the line $\mathrm{Re}(s) = 1/2$.

Weil showed that this type of interpretation could be extended naturally (and beautifully) to all sums of algebraic origin in *one* variable, i.e., ranging over points of a curve $C/\mathbf{F}_p$, using suitable Artin-Shreier coverings of C and the corresponding zeta functions, and most importantly appealing to the Riemann Hypothesis for all curves over finite

[3] Of course, for a fixed ν, they may arise for problems over a number field with a finite residue field of order p^ν, but then extensions of the latter would not occur.

fields, which he proved during the 1940's. This leads, for instance, to the following estimate, which has been used in many contexts in analytic number theory:

Theorem 3 (Weil)**.** *Let $P \in \mathbf{Z}[X]$ be a monic polynomial of degree $d \geqslant 1$ which no repeated factor. Let p be a prime such that P has no repeated root modulo p, i.e., such that p does not divide the discriminant of P. Then we have*

$$\Bigl|\sum_{x \,(\mathrm{mod}\, p)} e\Bigl(\frac{P(x)}{p}\Bigr)\Bigr| \leqslant (d-1)\sqrt{p}.$$

4. The yoga of Grothendieck-Deligne

Weil's theory, beautiful and revolutionary as it was, did not extend well to situations involving sums in more than one variable, or in other words, beyond the case where the underlying set of summation is (the set of $\mathbf{F}_p$-points of) a curve. Or, rather, it doesn't provide a way to analyze those sums except by "fibering by curves", which is simply to say writing

$$\Bigl|\sum_{x,y} e(f(x,y)/p)\Bigr| = \Bigl|\sum_{x}\Bigl(\sum_{y} e(f(x,y)/p)\Bigr)\Bigr| \leqslant \sum_{x}\Bigl|\sum_{y} e(f(x,y)/p)\Bigr|$$

and trying to understand individually the inner sums before putting them all together. This can often be done using Theorem 3, but note that if x, y range over $\mathbf{F}_p$, this means that – in terms of p – the best possible estimate one can then expect is of order $p^{3/2}$, as no cancellation can be obtained from the sum over x.

Such estimates, although non-trivial, are often not enough for the purposes of applications. Here is an example, with an equidistribution problem. Recall that, for a prime p and a multiplicative character $\chi : \mathbf{F}_p^\times \to \mathbf{C}^\times$ of the field $\mathbf{F}_p$, the associated Gauss sum (itself a type of exponential sum over finite field) is given by

$$\tau(\chi) = \sum_{x \in \mathbf{F}_p^\times} \chi(x) e\Bigl(\frac{x}{p}\Bigr).$$

For a non-trivial character $\chi \neq 1$, it is well known that $|\tau(\chi)| = \sqrt{p}$, which means one can write

$$\tau(\chi) = p^{1/2} e(\theta(\chi)), \qquad \text{with} \qquad \theta(\chi) \in [0,1].$$

Is is then an interesting question to understand how the angles of the Gauss sums $\theta(\chi)$ vary. Indeed, Gauss considered the case where $\chi = (\cdot/p)$ is a real-valued character (the Legendre symbol modulo p), and succeeded in proving after much effort that, for $p \geqslant 3$, we have

$$\tau((\cdot/p)) = 1, \text{ i.e., } \theta((\cdot/p)) = 0, \text{ if } p \equiv 1 \,(\mathrm{mod}\, 4),$$
$$\tau((\cdot/p)) = i, \text{ i.e., } \theta((\cdot/p)) = \frac{1}{4}, \text{ if } p \equiv 3 \,(\mathrm{mod}\, 4).$$

However, the angles seem intractable for most other characters,[4] and one may suspect that they are spread all over the unit circle. To test this, equidistribution theory suggests strongly to try to estimate the Weyl-type sums, which are

$$W_n = \frac{1}{p-2} \sum_{\substack{\chi \,(\mathrm{mod}\, p)\\ \chi \neq 1}} e(n\theta(\chi))$$

[4] A famous conjecture of Kummer considered what happens when χ is of order 3; for this story, see, e.g., [HP].

for $n \in \mathbf{Z} - \{0\}$, where the sum is over non-trivial multiplicative characters modulo p (the number of which is $p-2$). For $n \geqslant 1$, this can be expressed as

$$\begin{aligned}
W_n &= \frac{1}{p-2} \sum_{\substack{\chi \,(\mathrm{mod}\, p) \\ \chi \neq 1}} \Bigl(\frac{\tau(\chi)}{\sqrt{p}}\Bigr)^n \\
&= \frac{1}{p^{n/2}(p-2)} \sum_{x_1,\ldots,x_n} e\Bigl(\frac{x_1 + \cdots + x_n}{p}\Bigr) \sum_{\chi \neq 1} \chi(x_1 \cdots x_n) \\
&= \frac{p-1}{p-2} \frac{K_n(p)}{p^{n/2}},
\end{aligned}$$

where

$$K_n(p) = \sum_{\substack{x_1,\ldots,x_n \,(\mathrm{mod}\, p) \\ x_1 \cdots x_n = 1}} e\Bigl(\frac{x_1 + \cdots + x_n}{p}\Bigr)$$

is an exponential sum with $n-1$ variables, called an *hyper-Kloosterman sum* (because, in the case $n = 2$, we recover $K_2(p) = S(1,1;p)$).

If we sum over one variable using Weil's bound, we can easily show that $K_n(p) \ll p^{n-3/2}$ for all p, but note that such a bound does not even allow us to recover the trivial fact that $|W_n| \leqslant 1$ if $n \geqslant 4$! On the other hand, standard probabilistic considerations ("square-root cancellation philosophy", taking root in the fact that a random walk of length N with random, uniformly distributed phases, has modulus of size about $\sqrt{N}$ with overwhelmingly large probability) suggest that one should have

$$K_n(p) \ll p^{(n-1)/2},$$

or in other words, each of the $n-1$ variables should, independently of the others, gain a factor $\sqrt{p}$. This would lead to

$$W_n \ll \frac{1}{\sqrt{p}}$$

for $p \geqslant 3$ (and n fixed), and therefore[5] we would conclude that the angles of the Gauss sums $\tau(\chi)$, where χ ranges over all non-trivial characters, become equidistributed, as $p \to +\infty$, on the unit circle.

This fact could only be proved (by Deligne [D2]) after Grothendieck and his school of algebraic geometry had developed a new framework to study algebraic exponential sums, which goes much further than the one of Weil in many respects: not only does it encompass sums in arbitrarily many variables, but also it is formally much more flexible, and makes it possible to study and exploit variations of exponential sums in families, and indeed to analyze certain types of sums which do not look like the standard ones $S(f,V,p)$ we have mentioned earlier. However, because of considerations of space, we will only sketch how this formalism applies to exponential sums of this type, pointing (as an introduction) to the book of Katz [K2] for particularly striking illustrations of the more general sums that can naturally be considered, in the setting of the distribution of angles related to Kloosterman sums (see also another survey of the author [Ko2] for some motivation and discussion of Deligne's Equidistribution Theorem, which is an important part of this type of issues).

To present a fairly general case, we consider an affine scheme of finite type $V/\mathbf{Z}$, and two functions f, g on V, with g invertible. Then, for any prime number p, any additive

[5] For the application of the Weyl criterion to $n < 0$, one can use an obvious symmetry.

character $\psi : \mathbf{F}_p \to \mathbf{C}^\times$ and any multiplicative character $\chi : \mathbf{F}_p^\times \to \mathbf{C}^\times$, we look at the sum

$$S = S(V, f, g, \chi, \psi; p) = \sum_{x \in V(\mathbf{F}_p)} \psi(f(x))\chi(g(x)).$$

The case of hyper-Kloosterman sums corresponds to the affine subvariety $V \subset \mathbf{A}^n$ defined by the equation $x_1 \cdots x_n = 1$, to the additive character $x \mapsto e(x/p)$ (note that all additive characters are of the form $x \mapsto e(ax/p)$ for some a modulo p) and to $g = 1$, $\chi = 1$.

The analysis of such sums in the Grothendieck framework (see [D1] for a more detailed presentation, [IK, 11.11] for another survey tailored to analytic number theorists) starts by choosing, for a given prime p, another prime $\ell \neq p$. Then the formalism of the so-called *Lang torsor* (which is, concretely, a fairly systematic analysis of Artin-Schreier coverings (1) and of Kummer coverings for the multiplicative part) gives an object, called an ℓ-adic (lisse) sheaf of rank 1, $\mathcal{L} = \mathcal{L}_{\psi(f)} \otimes \mathcal{L}_{\chi(g)}$, depending on all the data, which can be seen as associating (in particular) to every rational point $x \in V(\mathbf{F}_p)$ a one-dimensional $\bar{\mathbf{Q}}_\ell$-vector space (the "stalk" $\mathcal{L}_x$ of $\mathcal{L}$ at x) together with an action of the Frobenius automorphism, which is the natural generator of the Galois group of $\mathbf{F}_p$, in such a way that the basic formula

$$\mathrm{Tr}(\mathrm{Fr}_x \mid \mathcal{L}_x) = \psi(f(x))\chi(g(x))$$

holds, for Fr_x the inverse of $x \mapsto x^p$ (the geometric Frobenius; up to changing f by $-f$ and g by g^{-1}, we could use the standard Frobenius as well). Of course, since the stalk is one-dimensional, speaking of the trace is somewhat pedantic, but the generalizations briefly mentioned earlier will involve similar sheaves such that $\mathcal{L}_x$ is a vector space of higher dimension.

Hence the exponential sums take the form

$$S = \sum_{x \in V(\mathbf{F}_p)} \mathrm{Tr}(\mathrm{Fr}_x \mid \mathcal{L}_x),$$

and the next basic steps will work for sums defined in this way for an arbitrary lisse ℓ-adic sheaf on $V/\mathbf{F}_p$.

The first transformation which is done is the analogue of the application of the rationality of the zeta function for Kloosterman sums (and it can be interpreted in this manner, although this would be anachronistic, since the rationality, in general, is proved exactly in this way): the *trace formula* of Grothendieck, a deep analogue of a formula of Lefschetz in classical algebraic topology, states that

$$\sum_{x \in V(\mathbf{F}_p)} \mathrm{Tr}(\mathrm{Fr}_x \mid \mathcal{L}_x) = \sum_{i=0}^{2d} (-1)^i \,\mathrm{Tr}(F \mid H^i_c(V_{\bar{\mathbf{F}}_p}, \mathcal{L})),$$

where the sum now runs only over integers up to $2d$, with d the dimension of $V/\mathbf{F}_p$, and F denotes the global action of the geometric Frobenius automorphism on the various ℓ-adic cohomology spaces with compact support of V, base-changed to an algebraic closure of $\mathbf{F}_p$.

Example 4. The trace formula is already interesting for the "trivial" sheaf $\bar{\mathbf{Q}}_\ell$ itself, for which all local traces are equal to 1, so that the associated sum is

$$\sum_{x \in V(\mathbf{F}_p)} \mathrm{Tr}(\mathrm{Fr}_x \mid \bar{\mathbf{Q}}_\ell) = |V(\mathbf{F}_p)|,$$

which is the number of points on V over $\mathbf{F}_p$. Indeed, the original Weil conjectures, which motivated the general theory (and in fact, much of the development of modern algebraic geometry, see the summary in [Ha, Appendix C]) concerned precisely this case. The trace formula states that

$$|V(\mathbf{F}_p)| = \sum_{i=0}^{2d} (-1)^i \operatorname{Tr}(F \mid H^i_c(V_{\bar{\mathbf{F}}_p}, \mathbf{Q}_\ell))$$

which is highly non-trivial in all but the simplest case.

One can see these formulas as black boxes, but of course this may become somewhat unsatisfactory. As a baby step towards enlightenment, let us consider one of the very few elementary situations where the formula becomes transparent. Consider the case where V is 0-dimensional, given by the equation $f(x) = 0$, for some monic polynomial $f \in \mathbf{Z}[X]$ of degree $\deg(f) \geqslant 1$. Then, for any prime p, $V/\mathbf{F}_p$ is zero-dimensional and $V(\mathbf{F}_p)$ is the number of zeros of f in the base field $\mathbf{F}_p$. Since $d = 0$, the trace formula gives

$$|\{x \,(\operatorname{mod} p) \mid f(x) = 0\}| = \operatorname{Tr}(F \mid H^0_c(V_{\bar{\mathbf{F}}_p}, \mathbf{Q}_\ell)).$$

But what is the 0-th cohomology space? Since $V_{\bar{\mathbf{F}}_p}$ is simply the finite collection of the zeros of f in $\bar{\mathbf{F}}_p$, the intuition from topology (which can be confirmed by the barest introduction to the definition of étale cohomology) states that $H^0_c(V_{\bar{\mathbf{F}}_p}, \bar{\mathbf{Q}}_\ell)$ should be isomorphic to $\bar{\mathbf{Q}}_\ell^\delta$, where δ is the number of distinct zeros of f (which was not assumed to be squarefree, so repeated roots are possible). And how should the global Frobenius act? It seems obvious (and again is confirmed easily) that its action on $\bar{\mathbf{Q}}_\ell^\delta$ is simply obtained from the permutation action of $x \mapsto x^p$ (or rather its inverse) on the zeros of f. In other words, the matrix representing this action is the permutation matrix associated with this permutation of the zeros of f. What is its trace? It is, as is well-known, the number of fixed points of the permutation, and this is precisely the number of zeros in $\mathbf{F}_p$.

This example illustrates also one property which is important and may seem doubtful at first: the trace formula works equally well for non-reduced schemes (e.g., if there are repeated roots) as for reduced ones. This is useful for applications, where checking that $V/\mathbf{F}_p$ is reduced might be quite bothersome.

As we leave this example, note that – even in this very simple case – the variation of $|V(\mathbf{F}_p)|$ with p is by no means an easy question!

Coming back to the general application of the trace formula, we note that it does not yet lead to any non-trivial estimate, because it might be that the traces on the various cohomology groups are enormous. In fact, it is not even clear (and it is open in general!) that the various terms on the right-hand side are independent of the choice of the auxiliary prime $\ell \neq p$. The eigenvalues may also conceivably be elements of $\bar{\mathbf{Q}}_\ell$ which are not algebraic over $\mathbf{Q}$.

However, the extraordinary general Riemann Hypothesis for varieties and sheaves over finite fields, proved by Deligne [D3], leads to quite precise information concerning those eigenvalues, from which non-trivial estimates may often be deduced.

In the context we consider, the result is the following: (1) any eigenvalue α of the Frobenius acting on a cohomology space $H^i_c(V_{\bar{\mathbf{F}}_p}, \mathcal{L})$, for $\mathcal{L}$ of the type described, is an algebraic integer; (2) any such α has the property that if $\beta \in \mathbf{C}$ is an arbitrary Galois conjugate of α (e.g., $\beta = \alpha$ if $\bar{\mathbf{Q}}$ is identified with a subfield of $\mathbf{C}$), we have

$$|\beta| = p^{j/2}$$

where the *weight* $j = j(\alpha)$ depends only on α and satisfies $0 \leqslant j \leqslant i$.

It is this last estimate of the weight which "is" the Riemann Hypothesis. To see why, consider the example of Kloosterman sums $S(1,1;p)$: since the variety V is the multiplicative group, of dimension 1, the yoga leads to

$$S(1,1;p) = \sum_{i=0}^{2} (-1)^i \operatorname{Tr}(F \mid H^i_c(\mathbf{G}_{m,\bar{\mathbf{F}}_p}, \mathcal{K})),$$

for some suitable sheaf, and the result of Deligne says that the eigenvalues of F on H^0_c are of modulus $\leqslant 1$, those on H^1_c are of modulus $\leqslant \sqrt{p}$, and those on H^2_c of modulus $\leqslant p$. Comparing with Weil's expression, one can guess (and it is true) that in fact $H^0_c = H^2_c = 0$ here, and H^1_c is 2-dimensional with eigenvalues of Frobenius given by the algebraic integers $\alpha_{1,1}$ and $\beta_{1,1}$ occurring in (2).

Here is a more general explanation of the relevance of Deligne's result. We expect that the number of points in the sum over $x \in V(\mathbf{F}_p)$ is roughly of size p^d (in simple cases, such as hyper-Kloosterman sums $K_n(p)$, V is so simple that this is obvious; there, $\dim V = n-1$ and $|V(\mathbf{F}_p)| = (p-1)^{n-1}$). Deligne's bound shows that only the topmost cohomology group H^{2d}_c may have eigenvalues as large as $p^d = p^{2d/2}$. If we let $\beta_i(\mathcal{L}) = \dim H^i_c$, we get

$$|S| = |S(V,f,g,\chi,\psi;p)| \leqslant \beta_{2d} p^d + R, \qquad |R| \leqslant q^{\kappa/2} B$$

(after choosing a fixed embedding of $\bar{\mathbf{Q}}$ in $\mathbf{C}$), where

$$\kappa = \max\{j \leqslant 2d-1 \mid \beta_j \neq 0\} \leqslant 2d-1, \qquad B = B(p) = \sum_{0 \leqslant i \leqslant \kappa} \dim H^i_c(V_{\bar{\mathbf{F}}_p}, \mathcal{L}).$$

Thus *getting some non-trivial estimate* depends on showing that the topmost group has dimension $\beta_{2d} = 0$. This, it turns out, is often the case, because Poincaré duality can be used to relate H^{2d}_c to some H^0, which can be analyzed fairly simply (there is a subtlety, which is avoided if one assumes that V is smooth over $\mathbf{F}_p$, which is most often not a problem in analytic number theory). Here is a special case: for a fixed $V/\mathbf{Z}$, and for any prime p large enough, if $V/\mathbf{F}_p$ is geometrically connected, we have $\beta_{2d} = 0$ *unless* the function f is constant on $U(\mathbf{F}_p)$ for some dense open subset $U/\mathbf{F}_p$ of V (equal to V if V is smooth over $\mathbf{F}_p$). In that case (if $\chi = 1$ at least), it is clear that we can not expect cancellation, since we are just counting the number of points of summation. In particular, this condition is trivially true for hyper-Kloosterman sums, and thus one gets "for free" that

$$|K_n(p)| \leqslant B p^{n-3/2}. \tag{3}$$

Strangely enough, in terms of exponential sums, this remains a trivial estimate, because B has not been estimated explicitly,[6] and it depends on p because the sheaves $\mathcal{L}$ that were introduced to represent the exponential sums themselves depend on p (through the additive and multiplicative characters). Thus the following theorem is crucial (see [K1] for a very general statement):

Theorem 5 (Dwork, Bombieri, Adolphson-Sperber)**.** *Let V, p, χ, ψ, $f \in \mathbf{Z}[V]$, $g \in \mathbf{Z}[V]^\times$ be as above. Then there exists a constant B_0, independent of p, such that*

$$\sum_{0 \leqslant i \leqslant 2d} \dim H^i_c(V_{\bar{\mathbf{F}}_p}, \mathcal{L}) \leqslant B_0$$

for all p.

[6] The corresponding estimate for the sums ranging over finite extensions $\mathbf{F}_{p^\nu}$ would be non-trivial.

In fact one can often write down a concrete value for B_0.

Let us come back, to conclude this section, to the special case of hyper-Kloosterman sums. The bound (3), reinforced by the bound $B(p) \leqslant B_0$, is no better than the one coming from Weil's bound for one-dimensional sums and "fibering". This is because it is based on assuming (a worst case scenario) that $\kappa = 2(n-1)-1$, i.e., that the cohomology space "topmost minus one" is non-zero, namely that

$$H_c^{2(n-1)-1} \neq 0.$$

This *would be the case* if the hyper-Kloosterman sums were replaced, indeed, by sums where only one variable is involved non-trivially, such as

$$\sum_{\substack{x_1,\ldots,x_n \\ x_1\cdots x_n=1}} e\Bigl(\frac{x_1^2}{p}\Bigr)$$

where nothing may be gained from the contribution of the $n-2$ variables $x_2, \ldots, x_n$. Of course, the actual hyper-Kloosterman sums do not look like that, but it is quite a bit more delicate to prove that, in fact, the only possible contribution to the cohomology comes from the "middle" dimension $i = n-1$. This is a result of Deligne:

Theorem 6 (Deligne)**.** *For the variety $V \subset \mathbf{A}^n$ with equation $X_1 \cdots X_n = 1$ and the sheaf $\mathcal{K}_n$ corresponding to hyper-Kloosterman sums, we have*

$$H_c^i(V_{\bar{\mathbf{F}}_p}, \mathcal{K}_n) = 0 \textit{ if } i \neq n-1,$$
$$\dim H_c^{n-1}(V_{\bar{\mathbf{F}}_p}, \mathcal{K}_n) = n.$$

As a corollary, the very precise bound

$$|K_n(p)| \leqslant np^{(n-1)/2}$$

follows from Deligne's analysis, hence also the equidistribution of angles of Gauss sums!

The proof of this theorem is quite intricate, but note that, because of the base change, this is a *geometric* statement, not so much a theorem of arithmetic anymore. This, in fact, explains partly why the method is so successful: it isolates *geometric* reasons for the smallness of the exponential sums, and these reasons may be accessible to arguments (e.g., deformation, "continuity") which are not available or visible from the purely arithmetic viewpoint.

A more direct illustration of this is found in the next theorem, also due to Deligne:

Theorem 7 (Deligne)**.** *Let $\mathbf{F}_q$ be a finite field, $P \in \mathbf{F}_q[X_1, \ldots, X_n]$ a homogeneous polynomial of degree d. Assume that the homogeneous part of degree d of P has the property that its zero set in the projective space of dimension $d-1$ (over $\bar{\mathbf{F}}_q$) is a smooth hypersurface. Then we have*

$$\Bigl|\sum_{x \in \mathbf{F}_q^n} e\Bigl(\frac{P(x)}{p}\Bigr)\Bigr| \leqslant (d-1)^n q^{n/2}.$$

Notice that this generalizes in a powerful way Weil's Theorem 3 to arbitrarily many dimensions (the condition on P, though natural, is not necessary for the estimate to hold). Deligne's proof, which is quite instructive, can be sketched very roughly as follows: the goal is to prove that the relevant cohomology groups (say of $\mathcal{L}_P$) satisfy

$$H_c^i(V_{\bar{\mathbf{F}}_p}, \mathcal{L}_P) = 0 \text{ if } i \neq n, \ \dim H_c^n(V_{\bar{\mathbf{F}}_p}, \mathcal{L}_P) = (d-1)^n.$$

However, when P varies among all the "Deligne" polynomials of degree d (those satisfying the condition concerning the zero set of the part of degree d), it turns out that the dimensions of the cohomology groups remain constant (this is a type of "smoothness" having to do with the fact that the Deligne polynomials are themselves parameterized by a nice affine algebraic variety), and hence it is enough to check the desired property for a *single* well-chosen P. This can be done for instance for

$$P = X_1^d + \cdots + X_n^d,$$

for which the exponential sums factors

$$\sum_{x\in \mathbf{F}_q^n} e\Bigl(\frac{P(x)}{p}\Bigr) = \Bigl(\sum_{x\in \mathbf{F}_q} e\Bigl(\frac{x^d}{p}\Bigr)\Bigr)^n$$

and (reverting the link between estimates and dimensions, as can be done) one obtains the required statement, e.g. from Weil's estimate (although these special one-variable sums, which are variants of Gauss sums, can also be estimated more directly).

Remark 8. The richness of the formalism of Grothendieck-Deligne is such that, quite often, it is possible to recover comparably precise estimates without requiring geometric analysis as detailed as that of Theorem 6 – in other words, by combining some extra arithmetic information that may well be available, for a given problem, with the geometric structure. For hyper-Kloosterman sums, this was done very cleverly by Bombieri; see [IK, 11.11, Example 2] for his argument based on mean-square averages of a family of hyper-Kloosterman sums and Galois-conjugacy.

5. Exponential sums over definable sets

We come back to the motivations from analytic number theory. Quite frequently, a natural problem is to estimate sums which are obtained from an exponential sum of algebraic origin by *shortening the range of summation*: for instance, estimates of

$$\sum_{n\leqslant p^\delta} \chi(p), \tag{4}$$

where χ is a multiplicative character modulo p and $0 < \delta < 1$ are of considerable importance in the theory of Dirichlet L-functions. This set is not the set of points of an algebraic variety,[7] but one may wonder if it is possible to extend the language used to define algebraic sets to allow for a richer spectrum of possibilities that might contain these "short intervals". This, and the encounter with the paper [CDM], led the author (see [Ko1]) to try to look at exponential sums over *definable sets* over finite fields. These are probably the simplest generalizations of algebraic varieties: they amount to replacing the set of points $V(\mathbf{F}_p)$ with a definable set $\varphi(\mathbf{F}_p)$ associated with a formula in the *first order language of rings*, i.e., the set of $x \in \mathbf{F}_p$ for which the formula is satisfied.

The case of algebraic varieties corresponds to positive formulas without quantifiers: if $V/\mathbf{F}_p$ is affine, embedded in $\mathbf{A}^n$, we have

$$x = (x_1, \ldots, x_n) \in V(A) \Leftrightarrow f_1(x) = \ldots = f_m(x) = 0$$

for any $\mathbf{F}_p$-algebra A, where $(f_1, \ldots, f_m)$ are generators of the ideal in $\mathbf{F}_p[X_1, \ldots, X_n]$ defined by V (i.e., those are possible equations defining V).

But, over finite fields at least, definable sets are more general. As a very simple example, consider the formula φ given by

$$\varphi(x) \,:\, \exists\, y,\ x = y^2,$$

[7] In a reasonable way, at least: of course, any finite set is the set of zeros of suitable polynomials.

so that $\varphi(A)$, for any ring A, is the set of squares in A (recall that all quantifiers are implicitly running over the ring A for which the formula is "evaluated"). In particular, note that

$$|\varphi(\mathbf{F}_p)| = \frac{p+1}{2}, \quad \text{if } p \text{ is odd,}$$

and there is no subvariety $V/\mathbf{Z} \subset \mathbf{A}^1/\mathbf{Z}$ with this number of $\mathbf{F}_p$-points for infinitely many primes.

Example 9. There are other examples of definable sets which are quite a bit more refined. Here are two types of important examples.

(1) Let $f(X,Y) \in \mathbf{Z}[X,Y]$ be a non-constant irreducible polynomial in two variables. There is then a formula $\varphi_f(x)$ such that, for any A, we have

$$\varphi_f(A) = \{x \in A \mid \text{the polynomial } f(x,Y) \in A[Y] \text{ is irreducible}\}$$

(these sets occur in the context of the Hilbert Irreducibility Theorem, and are often called *Hilbert sets*). Even more is true here: if we let $(a_{i,j})_{i+j\leqslant d}$ denote variables representing the coefficients of a polynomial $f \in \mathbf{Z}[X,Y]$ of degree $\leqslant d$, there is a single formula $\varphi_d(x, a_{i,j})$ with parameters $(a_{i,j})$ such that

$$\varphi_d(A, a_{i,j}) = \{x \in A \mid \text{the polynomial } \sum_{i,j} a_{i,j}x^iY^j \in A[Y] \text{ is irreducible}\}$$

for any choice of parameters $(a_{i,j})$. In other words, even the variation of the Hilbert sets $\varphi_f(A)$ with f is "definable", and this is an important property which can be crucial in applications. In algebraic geometry, the analogue variation would be that of the fibers $V_y = \pi^{-1}(y)$ for a morphism $V \xrightarrow{\pi} W$, where W is seen as the space of parameters.

Concretely, assume $d = 2$ (the simplest case). Then the parameters are $(a,b,c,d,e,f) = (a_{0,0}, a_{1,0}, a_{0,1}, a_{1,1}, a_{2,0}, a_{0,2})$ for the polynomials

$$a + bX + cY + dXY + eX^2 + fY^2 \in A[X,Y],$$

and since a polynomial of degree 2 is irreducible if and only it has no zero, we have

$$\varphi_2(x,a,b,c,d,e,f) \; : \; \forall\; y,\; a + bx + ex^2 + (c + dx)y + fy^2 \neq 0.$$

The nature of the sets $\varphi_2(\mathbf{F}_p, a, b, c, d, e, f)$ is not clear a priori, and even less so when the number of variables grows.

(2) Let $\mathbf{G}/\mathbf{F}_p$ be a linear algebraic group defined over $\mathbf{F}_p$, for instance $\mathbf{G} = GL(n)$ or $SL(n)$, or a symplectic group. After embedding $\mathbf{G}$ in a suitable affine space $\mathbf{A}^m$, we have coordinates $x = (x_1, \ldots, x_m)$ for $\mathbf{G}$, and $\mathbf{G}$ is defined by finitely many equations in terms of these, while the product and inverse maps of $\mathbf{G}$ are also given by polynomials with coefficients in $\mathbf{F}_p$. Then, we can see for instance that the conjugacy classes C_g are definable: if $\psi_{\mathbf{G}}(x_1, \ldots, x_m)$ is a formula defining $\mathbf{G}$ in the affine space, then

$$\varphi_{conj}(x,g) \; : \; \psi_{\mathbf{G}}(g) \wedge \psi_{\mathbf{G}}(x) \wedge (\exists\; z,\; \psi_{\mathbf{G}}(z) \wedge x = zgz^{-1})$$

with variables $x = (x_1, \ldots, x_m)$ and parameters $g = (g_1, \ldots, g_m)$ is such that $\varphi_{conj}(A, g)$ is naturally identified with the conjugacy class in $\mathbf{G}(A)$ of $g \in \mathbf{G}(A)$. In general, of course, such conjugacy classes are not the set of points of an algebraic variety.

Chatzidakis, van den Dries and Macintyre studied such definable sets (with parameters) in [CDM], in the case of finite fields, and established a beautiful result concerning the possible number of points. (See also [FHJ]).

Theorem 10 (Chatzidakis, van den Dries, Macintyre). *Let $\varphi(x,y)$ be a formula in the language of rings with n variables $x=(x_1,\ldots,x_n)$ and m parameters $y=(y_1,\ldots,y_m)$. There exist a set D of finitely many pairs $(\delta,\mu)\in\mathbf{Q}^+\times\mathbf{N}$ and a constant C depending only on $\varphi(x,y)$, with the following properties:*

(1) *For any finite field $\mathbf{F}_q$ with q elements, and $y\in\mathbf{F}_q^m$ such that $\varphi(\mathbf{F}_q,y)\neq\emptyset$, there exist $(\delta,\mu)\in D$ for which*

$$\Big||\varphi(\mathbf{F}_q,y)|-\delta q^{\mu}\Big|\leqslant Cq^{\mu-1/2}.$$

(2) *For any $(\delta,\mu)\in D$, there exists a formula $\mathcal{C}_{d,\mu}(y)$ with m parameters in the language of rings such that* (1) *holds for $\mathbf{F}_q$ and $y\in\mathbf{F}_q^m$ if and only if $y\in\mathcal{C}_{d,\mu}(\mathbf{F}_q)$.*

So, intuitively, the number of points $|\varphi(\mathbf{F}_p,y)|$ always looks like δq^{μ} for some rational "density" δ and some integral "dimension" μ, and although those may vary with y (and thus with p), there are only finitely many possibilities for a given φ – a type of "tameness" property of the variation of definable sets –, and moreover, the sets of parameters for which the density and dimension are fixed are themselves definable.

Some remarkable applications of this result in group theory where found by Hrushovski and Pillay in [HP], including a new proof of a difficult theorem of Mathews-Vaserstein-Weisfeiler on Strong Approximation for algebraic groups over $\mathbf{Z}$, which has many important applications in analytic number theory, in particular in applications of sieve methods.

It was fairly natural to try to extend this to results concerning exponential sums over definable sets, with the hope that these could have number-theoretic applications. More precisely (restricting to additive sums only for simplicity), given a formula $\varphi(x,y)$, a polynomial $f\in\mathbf{Z}[X]$, we can define a family (indexed by y) of exponential sums

$$S_{\varphi}(f,y)=S_{\varphi}(\mathbf{F}_q,f,\psi,y)=\sum_{x\in\varphi(\mathbf{F}_q,y)}\psi(f(x))$$

where ψ is an additive character of $\mathbf{F}_q$. These sums, to an analytic number theorist, are even worth investigating independently of possible applications, and this study was begun in [Ko1].

It turns out that the proof of Theorem 10 can be combined with the procedure of Section 4, with some care. This leads to the following statement (see [Ko1, Th. 13], which is a bit more general and more precise).

Theorem 11. *Let $\varphi(x,y)$ be a formula in the language of rings with n variables $x=(x_1,\ldots,x_n)$ and m parameters $y=(y_1,\ldots,y_m)$, and let $f\in\mathbf{Z}[X]$. Let D be the set given by Theorem 10 for $\varphi(x,y)$. There exists $p_0\geqslant 1$, a constant $\eta>0$, and $B\geqslant 0$, depending only on φ and the degree of f, such that for $p\geqslant p_0$ and $y\in\mathbf{F}_q^m$, we have*

$$|S_{\varphi}(f,y)|\leqslant\frac{BS_{\varphi}(1,y)}{\sqrt{p}}=\frac{B|\varphi(\mathbf{F}_p,y)|}{\sqrt{p}},$$

unless there exists $c\in\mathbf{F}_p$ with

$$|\{x\in\varphi(\mathbf{F}_p,y)\ \mid\ f(x)=c\}|\geqslant\eta S_{\varphi}(1,y)=\eta|\varphi(\mathbf{F}_p,y)|.$$

This statement is not entirely satisfactory, in that by itself it does not provide a way to understand when (or if) there is more cancellation than the $p^{-1/2}$ gain stated, which is comparable with the outcome of the Weil and fibering method. The condition which is excluded to obtain cancellation is easily understood: it means that f is constant for a positive proportion of the points of summation. If that is the case, any compensation in the sum (there may not be any, e.g., if $f = 1$) will have entirely different origins than what we expect generically. (This happens already for sums over varieties V if $V_{\bar{\mathbf{F}}_p}$ is not geometrically connected, where f might be constant on the geometric components, and the sum of these values might lead to cancellation).

From the proof of this theorem, one can extract some geometric information to help investigate further cancellation, which could, in principle, sometimes lead to stronger bounds. It would be highly interesting to find cases where this can be done, beyond cases involving only algebraic varieties, but the geometry is not as easily understood as that case (which may already be quite involved!). In particular, note that there may be little relation between $\varphi(\bar{\mathbf{F}}_p)$ and $\varphi(\mathbf{F}_p)$, i.e., no analogue of the fact that $\mathbf{F}_p$-rational points on V are fixed point of the Frobenius among the $\bar{\mathbf{F}}_p$-points. This is already clear for the set of squares in $\mathbf{F}_p$, which is non-trivial, but which of course is the whole line geometrically.

On the positive side, at least in the case of one-variable sums, the result is essentially optimal. So although we have not yet been able to find convincing purely arithmetic applications, we can prove the following statement (see [Ko1, Remark 19]), where equidistribution reappears:

Corollary 12. *Let $\varphi(x)$ be a formula in one variable in the first order language of rings. Let p run over any increasing sequence of primes, if it exists, such that $|\varphi(\mathbf{F}_p)| \to +\infty$ as p grows. Then the finite families $(\{(x/p\})_{x\in\varphi(\mathbf{F}_p)}$, where $\{x\}$ denotes the fractional part, become equidistributed in $\mathbf{R}/\mathbf{Z}$, with respect to Lebesgue measure, as $p \to +\infty$.*

Indeed, we simply apply the Weyl Criterion: for $h \neq 0$, we need to show that

$$\frac{1}{|\varphi(\mathbf{F}_p)|} \sum_{x\in\varphi(\mathbf{F}_p)} e\Bigl(\frac{hx}{p}\Bigr)$$

converges to 0 as p grows. This is an exponential sum over the definable set $\varphi(\mathbf{F}_p)$, hence by Theorem 11, we can find constants $B \geqslant 0$ and $\eta > 0$ for which

$$\sum_{x\in\varphi(\mathbf{F}_p)} e\Bigl(\frac{hx}{p}\Bigr) \leqslant B|\varphi(\mathbf{F}_p)|p^{-1/2},$$

unless the function $x \mapsto hx$ is constant for at least $\eta|\varphi(\mathbf{F}_p)|$ values of $x \in \varphi(\mathbf{F}_p)$. However, the latter is clearly impossible if $|\varphi(\mathbf{F}_p)| \to +\infty$.

In particular, this corollary means that it is not possible to find a formula $\varphi(x)$ such that, for infinitely many primes p, we have

$$\varphi(\mathbf{F}_p) = \{x \,(\mathrm{mod}\, p) \mid p/2 \leqslant x \leqslant 3p/2\}.$$

Note that this could not be derived directly from the point counting of Theorem 10, since the right-hand side forms a set which always has a density (namely, 1/2) and a dimension (namely, 1) in the sense of that theorem.

6. Questions

We conclude with a few questions which, possibly, could be the occasion of useful meetings of the minds between analytic number theory, algebraic geometry and model-theoretic ideas.

– We have seen that for applications to number theory, when p varies, Theorem 5 is crucial to the efficiency of the Grothendieck-Deligne approach to algebraic exponential sums. Except in the case where the additive character is trivial, its proof is not entirely satisfactory: one would hope to derive it as a type of “uniformity in parameters”, flowing naturally from the fact that the sums over finite fields come from an object “defined over **Z**". However, algebraic geometry does not provide such a global object for additive character sums, because the Artin-Schreier coverings (1) depend on p. Can one use model theory, possibly using a richer language than that of rings, to develop such a theory of "exponential sums over **Z**"? (See the paper [K3] of Katz for some speculations on this hypothetical theory).

– In some cases, "elementary" methods give better results than the algebraic methods (either Weil's, or those of Grothendieck-Deligne): consider for instance sums like

$$S_{m,p} = \sum_{x \,(\mathrm{mod}\, p)} e\Big(\frac{x^m + x}{p}\Big)$$

where now m is not bounded, but increases with p. By Theorem 3, we get

$$|S_{m,p}| \leqslant m\sqrt{p}$$

(this is a one-variable sum, and in the cohomological framework, we have $H^0_c = H^2_c = 0$ and $\dim H^1_c = d - 1$, so the method of Section 4 give the same result). Thus, if $m > \sqrt{p}$, this result is *worse* than the trivial bound $|S_{m,p}| \leqslant p$. Analytic heuristics, however, still suggest that $S_{m,p}$ should be "small" for larger values of m, provided certain conditions hold. Such results are indeed known (see, e.g., [B]), but the methods are completely different (based on additive combinatorics). Can progress be made on understanding this type of examples using algebraic geometry or model theory?

– Are there applications of Theorem 11 to model theory or algebraic geometry, maybe in the spirit of the work of Hrushovski and Pillay [HP]? This is also suggested by the work of Tomašić [T], but the author lacks competence to suggest any plausible track to follow...

– And the most important, maybe, for analytic number theory: is there an approach to more general exponential sums (such as those over short intervals (4)) involving a formalism as efficient (or comparable!) to the Grothendieck-Deligne approach?

References

[BF] K. Belabas and É. Fouvry: *Sur le 3-rang des corps quadratiques de discriminant premier ou presque premier*, Duke Math. J. 98 (1999), no. 2, 217–268.

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

[CDM] Z. Chatzidakis, L. van den Dries and A. Macintyre: *Definable sets over finite fields*, J. reine angew. Math. 427 (1992), 107–135

[C] R. Cluckers: *Igusa and Denef-Sperber conjectures on nondegenerate p-adic exponential sums*, Duke Math. J. 141 (2008), no. 1, 205–216.

[D1] P. Deligne: *Cohomologie étale*, S.G.A $4\frac{1}{2}$, L.N.M 569, Springer Verlag (1977).

[D2] P. Deligne: *La conjecture de Weil : I*, Publ. Math. IHÉS 43 (1974), 273–307

[D3] P. Deligne: *La conjecture de Weil, II*, Publ. Math. IHÉS 52 (1980), 137–252.

[FHJ] 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.

[HW] G.H. Hardy and E.M. Wright: *An introduction to the theory of numbers*, 5th ed., Oxford Univ. Press, 1979.

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

[HP] D.R. Heath-Brown and S.J. Patterson: *The distribution of Kummer sums at prime arguments*, J. reine angew. Math. 310 (1979), 111–130.

[HP] E. Hrushovski and A. Pillay: *Definable subgroups of algebraic groups over finite fields*, J. reine angew. Math 462 (1995), 69–91.

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

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

[K1] N. Katz: *Sums of Betti numbers in arbitrary characteristic*, Finite Fields Appl. 7 (2001), no. 1, 29–44.

[K2] N. Katz: *Gauss sums, Kloosterman sums and monodromy*, Annals of Math. Studies, 116, Princeton Univ. Press, 1988.

[K3] N. Katz: *Exponential sums over finite fields and differential equations over the complex numbers: some interactions*, Bull. A.M.S 23 (1990), 269–309.

[Ko1] E. Kowalski: *Exponential sums over definable subsets of finite fields*, Israel J. Math. 160 (2007), 219–251.

[Ko2] E. Kowalski: *Some aspects and applications of the Riemann Hypothesis over finite fields*, Milan J. of Mathematics 78 (2010), 179–220.

[T] I. Tomašić: *Exponential sums in pseudofinite fields and applications*, Illinois J. Math. 48 (2004), no. 4, 1235–1257.

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