THE PRINCIPLE OF THE LARGE SIEVE

E. KOWALSKI

Pour les soixantes ans de J-M. Deshouillers

Abstract. We describe a very general abstract form of sieve based on a large sieve inequality which generalizes both the classical sieve inequality of Montgomery (and its higher-dimensional variants), and our recent sieve for Frobenius over function fields. The general framework suggests new applications. We give some first results on the number of prime divisors of "most" elements of an elliptic divisibility sequence, and we develop in some detail "probabilistic" sieves for random walks on arithmetic groups, e.g., estimating the probability of finding a reducible characteristic polynomial at some step of a random walk on $SL(n, \mathbf{Z})$. In addition to the sieve principle, the applications depend on bounds for a large sieve constant. To prove such bounds involves a variety of deep results, including Property (τ) or expanding properties of Cayley graphs, and the Riemann Hypothesis over finite fields.

Contents

1. Introduction

Classical sieve theory is concerned with the problem of the asymptotic evaluation of averages of arithmetic functions over integers constrained by congruence restrictions modulo a set of primes. Often the function in question is the characteristic function of some interesting sequence and the congruence restrictions are chosen so that those integers remaining after the sieving process are, for instance, primes or "almost" primes.

If the congruence conditions are phrased as stating that the only integers n which are allowed are those with reduction modulo a prime p not in a certain set Ω_p, then a familiar dichotomy arises: if Ω_p contains few residue classes (typically, a bounded number as p increases), the setting

Date: February 2, 2008.
2000 *Mathematics Subject Classification.* 11N35, 11N36, 11C99, 60G50, 22D10, 14G15, 11B37, 20C33.
Key words and phrases. Sieve methods, large sieve inequalities, distribution of Frobenius elements, elliptic divisibility sequences, random walks on groups, Property (τ), Deligne-Lusztig characters, inclusion-exclusion principle, pseudo-Anosov diffeomorphisms.

is that of a "small" sieve. The simplest such case is the detection of primes with $\Omega_p = \{0\}$. If, on the other hand, the size of Ω_p increases, the situation is that of a "large" sieve. The first such sieve was devised by Linnik to investigate the question of Vinogradov of the size of the smallest quadratic non-residue modulo a prime.

There have already been a number of works extending "small" sieves to more general situations, where the objects being sifted are not necessarily integers. Among these, one might quote the vector sieve of Brüdern and Fouvry [BF], with applications to Lagrange's theorem with almost prime variables, the "crible étrange" of Fouvry and Michel [FM], with applications to sign changes of Kloosterman sums, and Poonen's striking sieve procedure for finding smooth hypersurfaces of large degree over finite fields [Po].

Similarly, the large sieve has been extended in some ways, in particular (quite early on) to deal with sieves in $\mathbf{Z}^d$, $d \geqslant 1$, or in number fields (see e.g. [G]). Interesting applications have been found, e.g. Duke's theorem on elliptic curves over $\mathbf{Q}$ with "maximal" p-torsion fields for all p [D]. All these were much of the same flavor however, and in particular depended only on the character theory of finite abelian groups as far as the underlying harmonic analysis was concerned[1].

In [Ko1], we have introduced a new large sieve type inequality for the average distribution of Frobenius conjugacy classes in the monodromy groups of a family $(\mathcal{F}_\ell)$ of $\mathbf{F}_\ell$-adic sheaves on a variety over a finite field. Although the spirit of the large sieve is clearly recognizable, the setting is very different, and the harmonic analysis involves both non-abelian finite groups, and the deep results of Deligne on the Riemann Hypothesis over finite fields. Our first application of this new sieve was related to the "generic" arithmetic behavior of the numerator of the zeta function of a smooth projective curve in a family with large monodromy, improving significantly a result of Chavdarov [Ch].

Motivated by this first paper, the present one is interested with foundational issues related to the large sieve. We are able to describe a very general abstract framework which we call "the principle of the large sieve", with a pun on [Mo]. This leads to a sieve statement that may in particular be specialized to either the classical forms of the large sieve, or to a strengthening of [Ko1]. Roughly speaking, we deal with a set X that can be mapped to finite sets X_ℓ (for instance, integers can be reduced modulo primes) and we show how an estimate for the number of those $x \in X$ which have "reductions" outside $\Omega_\ell \subset X_\ell$ for all or some ℓ may be reduced to a bilinear form estimate of a certain kind. The form of the sieve statement we obtain is similar to Montgomery's formulation of the large sieve (see e.g. [Mo], [B], [IK, 7.4]). It should be mentioned that our "axioms" for the sieve may admit other variations. In fact, Zywina [Z] has developed a somewhat similar framework, and some of the flexibility we allow was first suggested by his presentation.

There remains the problem of estimating the bilinear form. The classical idea of duality and exponential sums is one tool in this direction, and we describe it also somewhat abstractly. We then find a convincing relation with the classical sieve axioms, related to equidistribution in the finite sets X_ℓ.

The bilinear form inequality also seemingly depends on the choice of an orthonormal basis of certain finite-dimensional Hilbert spaces. It turns out that in many applications, the sieve setting is related to the existence of a group G such that X_ℓ is the set of conjugacy classes in a finite quotient G_ℓ of G and the reduction $X \to X_\ell$ factors through G. In that case, the bilinear form inequality can be stated with a distinguished basis arising from the representation theory (or harmonic analysis) of the finite groups G_ℓ.

This abstract sieving framework has many incarnations. As we already stated, we can recover the classical large sieve and the "sieve for Frobenius" of [Ko1], but furthermore, we are led to

[1] There is, of course, an enormously important body of work concerning inequalities traditionally called "large sieve inequalities" for coefficients of automorphic forms of various types which have been developed by Iwaniec, Deshouillers-Iwaniec, Duke, Duke-Kowalski, Venkatesh and others (a short survey is in [IK, §7.7]). However, those generalize the large sieve inequality for Dirichlet characters, and have usually no relation (except terminological) with the traditional sieve principle.

a number of situations which are either new (to the author's knowledge), or have received attention only recently, although not in the same form in general. One of these concerns (small) sieves in arithmetic groups and is the subject of ongoing work of Bourgain, Gamburd and Sarnak [BGS], and some of the problems it is suited for have been raised and partly solved by Rivin [R], who also emphasized possible applications to some groups which are "close in spirit" to arithmetic groups, such as mapping class groups of surfaces or automorphism groups of free groups. Indeed, the large sieve strengthens significantly the results of Rivin (see Corollary 9.7).

Our main interest in writing this paper is the exploration of the general setting. Consequently, the paper is fairly open-ended and has a distinctly chatty style. We hope to come back to some of the new examples with more applications in the future. Still, to give a feeling for the type of results that become available, we finish this introduction with a few sample statements (the last one could in fact have been derived in [Ko1], with a slightly worse bound).

Theorem 1.1. *Let (S_n) be a simple random walk on $\mathbf{Z}$, i.e.,*

$$S_n = X_1 + \cdots + X_n$$

where (X_k) is a sequence of independent random variables with $\mathbf{P}(X_k = \pm 1) = \frac{1}{2}$ for all k.

Let $\varepsilon > 0$ be given, $\varepsilon \leqslant 1/4$. For any odd $q \geqslant 1$, any a coprime with q, we have

$$\mathbf{P}(S_n \textit{ is prime and } \equiv a \,(\mathrm{mod}\, q)) \ll \frac{1}{\varphi(q)} \frac{1}{\log n}$$

if $n \geqslant 1$, $q \leqslant n^{1/4-\varepsilon}$, the implied constant depending only on ε.

Theorem 1.2. *Let $n \geqslant 2$ be an integer, let $G = SL(n, \mathbf{Z})$ and let $S = S^{-1} \subset G$ be a finite generating set of G, e.g., the finite set of elementary matrices with ± 1 entries off the diagonal. Let (X_k) be the simple left-invariant random walk on Γ, i.e., a sequence of Γ-valued random variables such that $X_0 = 1$ and*

$$X_{k+1} = X_k \xi_{k+1} \textit{ for } k \geqslant 0,$$

where (ξ_k) is a sequence of S-valued independent random variables with $\mathbf{P}(\xi_k = s) = \frac{1}{|S|}$ for all $s \in S$. Then, almost surely, there are only finitely many k for which the characteristic polynomial $\det(X_k - T) \in \mathbf{Z}[T]$ is reducible, or in other words, the set of matrices with reducible characteristic polynomials in $SL(n, \mathbf{Z})$ is transient for the random walk.

In fact (see Theorem 9.4), we will derive this by showing that the probability that $\det(X_k - T)$ be reducible decays exponentially fast with k (in the case $n \geqslant 3$ at least). An analogue of this result (with some extra conditions) has the geometric/topological consequence that the set of non-pseudo-Anosov elements is transient for random walks on mapping class groups of closed orientable surfaces, answering a question of Maher [Ma, Question 1.3] (see Corollary 9.7 for details; this application was suggested by Rivin's paper [R]).

Theorem 1.3. *Let $n \geqslant 3$ be an integer, let $G = SL(n, \mathbf{Z})$, and let $S = S^{-1} \subset G$ be a finite symmetric generating set. Then there exists $\beta > 0$ such that for any $N \geqslant 1$, we have*

$$|\{w \in S^N \mid \textit{one entry of the matrix } g_w \textit{ is a square}\}| \ll |S|^{N(1-\beta)},$$

where $g_w = s_1 \cdots s_N$ for $w = (s_1, \ldots, s_N) \in S^N$, and β and the implied constant depend only on n and S.

Equivalently, for the random walk (X_k) on G defined as in the statement of the previous theorem, we have

$$\mathbf{P}(\textit{one entry of the matrix } X_k \textit{ is a square}) \ll \exp(-\delta k)$$

for $k \geqslant 1$ and some constant $\delta > 0$, where δ and the implied constant depend only on n and S.

Theorem 1.4. *Let $E/\mathbf{Q}$ be an elliptic curve with rank $r \geqslant 1$ given by a Weierstrass equation*

$$y^2 + a_1 xy + a_3 y = x^3 + a_2 x^2 + a_4 x + a_6, \qquad \textit{where } a_i \in \mathbf{Z}.$$

For $x \in E(\mathbf{Q})$, let $\omega_E(x)$ be the number of primes, without multiplicity, dividing the denominator of the coordinates of x, with $\omega_E(0) = +\infty$. Let $h(x)$ denote the canonical height on E.

Then for any fixed real number κ with $0 < \kappa < 1$, we have

$$|\{x \in E(\mathbf{Q}) \mid h(x) \leqslant T \text{ and } \omega_E(x) < \kappa \log\log T\}| \ll T^{r/2}(\log\log T)^{-1},$$

for $T \geqslant 3$, where the implied constant depends only on E and κ.

Theorem 1.5. *Let q be a power of a prime number $p \geqslant 5$, $g \geqslant 1$ an integer and let $f \in \mathbf{F}_q[T]$ be a squarefree polynomial of degree $2g$. For t not a zero of f, let C_t denote the smooth projective model of the hyperelliptic curve*

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

and let J_t denote its Jacobian variety. Then we have

$$|\{t \in \mathbf{F}_q \mid f(t) \neq 0 \text{ and } |C_t(\mathbf{F}_q)| \text{ is a square}\}| \ll q^{1-\gamma}(\log q),$$
$$|\{t \in \mathbf{F}_q \mid f(t) \neq 0 \text{ and } |J_t(\mathbf{F}_q)| \text{ is a square}\}| \ll q^{1-\gamma}(\log q)$$

where $\gamma = (4g^2 + 2g + 4)^{-1}$, and the implied constants are absolute.

It is well-known that the strong form of the large sieve is as efficient (qualitatively) as the best small sieves, as far as upper bound sieves are concerned. To put this in context, we will briefly recall the principles of small sieves (in the same abstract context) in an Appendix, and we will give a sample application (Theorem A.3) related to Theorem 1.5.

The plan of this paper is as follows. In the first sections, the abstract sieve setting is described, and the abstract large sieve inequality is derived; this is a pleasant and rather straightforward algebraic exercise. In Sections 4 and 6, we specialize the general setting to two cases ("group sieve" and "coset sieve") related to group theory, using the representation theory of finite groups. This leads to the natural problem of finding precise estimates for the degree and the sum of degrees of irreducible representations of some finite groups of Lie type, which we consider in some cases in Section 7. For this we use Deligne-Lusztig characters, and arguments shown to the author by J. Michel; this section may be omitted in a first reading.

Turning to examples of sieves, already in Section 5 we show how many classically-known uses of the large sieve are special cases of the setting of Section 4. In the same section, we also indicate the relation with the inclusion-exclusion technique in probability and combinatorics, which shows in particular that the general sieve bound is sharp (see Example 5.6).

New (or emerging) situations are considered next, in four sections which are quite independent of one another (all of them involve either group or coset sieves). "Probabilistic" sieves are discussed briefly in Section 8, leading to Theorem 1.1. Sieving in arithmetic groups is described in Section 9, where Theorem 1.2 is proved. The crucial point (as in the work of Bourgain, Gamburd and Sarnak) is the expanding properties of Cayley graphs of $SL(n, \mathbf{Z}/d\mathbf{Z})$, phrased in terms of Property (τ). Then comes an amusing "elliptic sieve" which is related to the number of prime divisors of the denominators of rational points on an elliptic curve, leading to Theorem 1.4. In turn, this is linked to the analysis of the prime factorization of elements of so-called "elliptic divisibility sequences", and we find that "most" elements have many prime factors. This complements recent heuristics and results of Silverman, Everest, Ward and others concerning the paucity of primes and prime powers in such sequences. Finally, in Section 11, we extend the sieve result of [Ko1] concerning the distribution of geometric Frobenius conjugacy classes in finite monodromy groups over finite fields, and derive some new applications. To conclude, Appendix A briefly indicates the link with small sieve situations, for the purpose of comparison and reference, with a sample application, and Appendix B contains the proofs of some "local" density computations in matrix groups over finite fields. Those estimates have been used previously, but we defer the proof to not distract from the main thrust of the arguments underlying the principle of the sieve. Note that the techniques underlying those computations are in fact quite advanced and of independent interest, and involve work of Chavdarov [Ch] and non-trivial estimates for exponential sums over finite fields.

Notation. As usual, $|X|$ denotes the cardinality of a set; however if X is a measure space with measure μ, we sometimes write $|X|$ instead of $\mu(X)$.

For a group G, $G^\sharp$ denotes the set of its conjugacy classes, and for a conjugacy-invariant subset $X \subset G$, $X^\sharp \subset G^\sharp$ is the corresponding set of conjugacy classes. The conjugacy class of $g \in G$ is denoted $g^\sharp$.

By $f \ll g$ for $x \in X$, or $f = O(g)$ for $x \in X$, where X is an arbitrary set on which f is defined, we mean synonymously that there exists a constant $C \geqslant 0$ such that $|f(x)| \leqslant Cg(x)$ for all $x \in X$. The "implied constant" is any admissible value of C. It may depend on the set X which is always specified or clear in context. The notation $f \asymp g$ means $f \ll g$ and $g \ll f$. On the other hand $f(x) = o(g(x))$ as $x \to x_0$ is a topological statement meaning that $f(x)/g(x) \to 0$ as $x \to x_0$.

For $n \geqslant 1$ an integer, $\omega(n)$ is the number of primes dividing n, without counting multiplicity. For $z \in \mathbf{C}$, we denote $e(z) = \exp(2i\pi z)$.

In probabilistic contexts, $\mathbf{P}(A)$ is the probability of an event, $\mathbf{E}(X)$ is the expectation of a random variable X, $\mathbf{V}(X)$ its variance, and $\mathbf{1}_A$ is the characteristic function of an event A.

Acknowledgments. D. Zywina has developed [Z] an abstract setup of the large sieve similar to the conjugacy sieve described in Section 4. His remarks have been very helpful both for the purpose of straightening out the assumptions used, and as motivation for the search of new "unusual" applications. One of his nice tricks (the use of general sieve support) is also used here. The probabilistic setting was suggested in part by Rivin's preprint [R], who also mentioned to me the work of Bourgain, Sarnak and Gamburd. I also wish to thank P. Sarnak for sending me a copy of his email [Sa1] to his coauthors. Finally, I thank J. Michel for providing the ideas of the proof of Proposition 7.3 and explaining some basic properties of representations of finite groups of Lie type, and P. Duchon and M-L. Chabanol for help, advice and references concerning probability theory and graph theory.

2. The principle of the large sieve

We will start by describing a very general type of sieve. The goal is to reach an analogue of the large sieve inequality, in the sense of a reduction of a sieve bound to a bilinear form estimate.

We start by introducing the notation and terminology. The *sieve setting* is a triple $\Psi = (Y, \Lambda, (\rho_\ell))$ consisting of

- A set Y;
- An index set Λ;
- For all $\ell \in \Lambda$, a surjective map $\rho_\ell : Y \to Y_\ell$ where Y_ℓ is a finite set.

In combinatorial terms, this might be thought as a family of colorings of the set Y. In applications, Λ will often be a subset of primes (or prime ideals in some number field), but as first pointed out by Zywina, this is not necessary for the formal part of setting up the sieve, and although the generality is not really abstractly greater, it is convenient to allow arbitrary Λ.

Then, a *siftable set* associated to $\Psi = (Y, \Lambda, (\rho_\ell))$ is a triple $\Upsilon = (X, \mu, F)$ consisting of

- A measure space (X, μ) with $\mu(X) < +\infty$;
- A map $F : X \to Y$ such that the composites $X \to Y \to Y_\ell$ are measurable, i.e., the sets $\{x \in X \mid \rho_\ell(F_x) = y\}$ are measurable for all ℓ and all $y \in Y_\ell$.

The simplest case is when X is a finite set and μ is counting measure. We call this the *counting case*. Even when this is not the case, for notational convenience, we will usually write $|B| = \mu(B)$ for the measure of a measurable set $B \subset X$.

The last piece of data is a finite subset $\mathcal{L}^*$ of Λ, called the *prime sieve support*, and a family $\Omega = (\Omega_\ell)$ of *sieving sets*[2] of Y_ℓ, defined for $\ell \in \mathcal{L}^*$.

With this final data $(\Psi, \Upsilon, \mathcal{L}^*, \Omega)$, we can define the sieve problem.

[2] Sometimes, Ω will also denote a probability space, but no confusion should arise.

Definition 2.1. Let $\Psi = (Y, \Lambda, (\rho_\ell))$ be a sieve setting, $\Upsilon = (X, \mu, F)$ a siftable set, $\mathcal{L}^*$ a prime sieve support and Ω a family of sieving sets. Then the *sifted sets* are

$$S(Y, \Omega; \mathcal{L}^*) = \{y \in Y \mid \rho_\ell(y) \notin \Omega_\ell \text{ for all } \ell \in \mathcal{L}^*\},$$
$$S(X, \Omega; \mathcal{L}^*) = \{x \in X \mid \rho_\ell(F_x) \notin \Omega_\ell \text{ for all } \ell \in \mathcal{L}^*\}.$$

The latter is also $F^{-1}(S(Y, \Omega; \mathcal{L}^*))$ and is a measurable subset of X.

The problem we will consider is to find estimates for the measure $|S(X, \Omega; \mathcal{L}^*)|$ of the sifted set. Here we think that the sieve setting is fixed, while there usually will be an infinite sequence of siftable sets with size $|X|$ going to infinity; this size will be the main variable in the estimates.

Example 2.2. The classical sieve arises as follows: the sieve setting is

$$\Psi = (\mathbf{Z}, \{\text{primes}\}, \mathbf{Z} \to \mathbf{Z}/\ell\mathbf{Z})$$

and the siftable sets are $X = \{n \mid M < n \leqslant M + N\}$ with counting measure and $F_x = x$ for $x \in X$. Then the sifted sets become the classical sets of integers in an interval with reductions modulo primes in $\mathcal{L}^*$ lying outside a subset $\Omega_\ell \subset \mathbf{Z}/\ell\mathbf{Z}$ of residue classes.

In most cases, (X, μ) will be a finite set with counting measure, and often $X \subset Y$ with $F_x = x$ for $x \in X$. See Section 11 for a conspicuous example where F is not the identity, Section 8 for interesting situations where the measure space (X, μ) is a probability space, and F a random variable, and Section 9 for another example.

We will now indicate one type of inequality that reduces the sieve problem to the estimation of a *large sieve constant* Δ. The latter is a more analytic problem, and can be attacked in a number of ways. This large sieve constant depends on most of the data involved, but is independent of the sieving sets.

First we need some more notation. Given a sieve setting Ψ, we let $S(\Lambda)$ denote the set of finite subsets $m \subset \Lambda$. Since $S(\Lambda)$ may be identified with the set of squarefree integers $m \geqslant 1$ in the classical case where Λ is the set of primes, to simplify notation we write $\ell \mid m$ for $\ell \in m$ when $\ell \in \Lambda$ and $m \in S(\Lambda)$, and similarly for $n \mid m$ instead of $n \subset m$ if $n, m \in S(\Lambda)$.

A *sieve support* $\mathcal{L}$ associated to a prime sieve support $\mathcal{L}^*$ is any finite subset of $S(\Lambda)$ such that

$$\ell \in m,\ m \in \mathcal{L} \text{ implies } \ell \in \mathcal{L}^*, \text{ and } \{\ell\} \in \mathcal{L} \text{ if } \ell \in \mathcal{L}^*. \tag{2.1}$$

This implies that $\mathcal{L}$ determines $\mathcal{L}^*$ (as the set of elements of singletons in $\mathcal{L}$). If Λ is a set of primes, $\mathcal{L}$ "is" a set of squarefree integers only divisible by primes in $\mathcal{L}^*$ and containing $\mathcal{L}^*$ (including possibly $m = 1$, not divisible by any prime).

For $m \in S(\Lambda)$, let

$$Y_m = \prod_{\ell \mid m} Y_\ell$$

and let $\rho_m : Y \to Y_m$ be the obvious product map. (In other words, we look at all "refined" colorings of Y obtained by looking at all possible finite tuples of colorings). If $m = \emptyset$, Y_m is a set with a single element, and ρ_m is a constant map.

We will consider functions on the various sets Y_m, and it will be important to endow the space of complex-valued functions on Y_m with appropriate and consistent inner products. For this purpose, we assume given for $\ell \in \Lambda$ a density

$$\nu_\ell : Y_\ell \to [0, 1]$$

(often denoted simply ν when no ambiguity is possible) such that the inner product on functions $f : Y_\ell \to \mathbf{C}$ is given by

$$\langle f, g \rangle = \sum_{y \in Y_\ell} \nu_\ell(y) f(y) \overline{g(y)}.$$

We assume that $\nu(y) > 0$ for all $y \in Y_\ell$, in order that this hermitian form be positive definite (it will be clear that $\nu(y) \geqslant 0$ would suffice, but the stronger assumption is no problem for applications), and that ν is a probability density, i.e., we have

$$\sum_{y \in Y_\ell} \nu_\ell(y) = 1. \tag{2.2}$$

Using the product structure we define corresponding inner products and measures on the spaces of functions $Y_m \to \mathbf{C}$. Property (2.2) still holds. We will interpret ν as a measure on Y_ℓ or Y_m, so we will write for instance

$$\nu(\Omega_\ell) = \sum_{y \in \Omega_\ell} \nu(y), \qquad \text{for } \Omega_\ell \subset Y_\ell.$$

We denote by $L^2(Y_m)$ the space of complex-valued functions on Y_m with the inner product thus defined.

The simplest example is when $\nu(y) = 1/|Y_m|$, but see Sections 4 and 6 for important natural cases where ν is not uniform. It will be clear in the remarks and sections following the statement of the sieve inequality that, in general, the apparent choice of ν_ℓ is illusory (only one choice will lead to good results).

Note that ρ_m is not necessarily surjective, but it turns out to be true, and a crucial fact, in most applications of the sieve, so we make a definition (the terminology will be clearer in later applications).

Definition 2.3. A sieve setting $\Psi = (Y, \Lambda, (\rho_\ell))$ is *linearly disjoint* if the map $\rho_m : Y \to Y_m$ is onto for all $m \in S(\Lambda)$.

Here is now the first sieve inequality.

Proposition 2.4. *Let Ψ, Υ, $\mathcal{L}^*$ be as above. For any sieve support $\mathcal{L}$ associated to $\mathcal{L}^*$, i.e, any finite subset of $S(\Lambda)$ satisfying* (2.1)*, let $\Delta = \Delta(X, \mathcal{L})$ denote the* large sieve constant*, which is by definition the smallest non-negative real number such that*

$$\sum_{m \in \mathcal{L}} \sum_{\varphi \in \mathcal{B}_m^*} \Bigl| \int_X \alpha(x) \varphi(\rho_m(F_x)) d\mu(x) \Bigr|^2 \leqslant \Delta \int_X |\alpha(x)|^2 d\mu(x) \tag{2.3}$$

for any square integrable function $\alpha : X \to \mathbf{C}$, where φ in the outer sum ranges over $\mathcal{B}_m^*$, where $\mathcal{B}_\ell^* = \mathcal{B}_\ell - \{1\}$, $\mathcal{B}_\ell$ is an orthonormal basis, containing the constant function 1, of the space $L^2(Y_\ell)$, and for all m we let

$$\mathcal{B}_m = \prod_{\ell | m} \mathcal{B}_\ell, \qquad \mathcal{B}_m^* = \prod_{\ell | m} \mathcal{B}_\ell^*,$$

the function on Y_m corresponding to (φ_ℓ) being given by

$$(y_\ell) \mapsto \prod_{\ell | m} \varphi_\ell(y_\ell),$$

and for $m = \emptyset$, we have $\mathcal{B}_m = \mathcal{B}_m^* = \{1\}$.

Then for arbitrary sieving sets $\Omega = (\Omega_\ell)$, we have

$$|S(X, \Omega; \mathcal{L}^*)| \leqslant \Delta H^{-1}$$

where

$$H = \sum_{m \in \mathcal{L}} \prod_{\ell | m} \frac{\nu(\Omega_\ell)}{\nu(Y_\ell - \Omega_\ell)} = \sum_{m \in \mathcal{L}} \prod_{\ell | m} \frac{\nu(\Omega_\ell)}{1 - \nu(\Omega_\ell)}. \tag{2.4}$$

Remark 2.5. The large sieve constant as defined above is independent of the choices of basis $\mathcal{B}_\ell$ (containing the constant function 1). Here is a more intrinsic definition which shows this, and provides a first hint of the link with classical (small) sieve axioms. It's not clear how much this intrinsic definition can be useful in practice, which explains why we kept a concrete version in the statement of Proposition 2.4.

By definition, the inequality (2.3) means that Δ is the square of the norm of the linear operator

$$T\begin{cases} L^2(X,\mu) & \longrightarrow & \displaystyle\bigoplus_{m\in\mathcal{L}} L(L_0^2(Y_m),\mathbf{C}) \\ \alpha & \mapsto & \displaystyle\Bigl(f\mapsto \int_X \alpha(x)f(\rho_m(F_x))d\mu(x)\Bigr)_m \end{cases}$$

where the direct sum over m is orthogonal and $L(L_0^2(Y_m),\mathbf{C})$ is the space of linear functionals on

$$L_0^2(Y_m)=\bigotimes_{\ell\mid m} L_0^2(Y_\ell),\quad \text{where}\quad L_0^2(Y_\ell)=\{f\in L^2(Y_\ell)\mid \langle f,1\rangle=\sum_y \nu(y)f(y)=0\}$$

(the space $L_0^2(Y_m)$ may be thought of as the "primitive" subspace of the functions on Y_m), with the norm

$$\|f^*\|=\max_{f\neq 0}\frac{|\langle f^*,f\rangle|}{\|f\|}.$$

Since we are dealing with Hilbert spaces, $L(L_0^2(Y_m),\mathbf{C})$ is canonically isometric to $L_0^2(Y_m)$, and Δ is the square of the norm of the operator

$$T_1\begin{cases} L^2(X,\mu) & \longrightarrow & \displaystyle\bigoplus_{m\in\mathcal{L}} L_0^2(Y_m) \\ \alpha & \mapsto & T_1(\alpha) \end{cases}$$

where $T_1(\alpha)$ is the vector such that $\langle f,T_1(\alpha)\rangle=T(\alpha)(f)$ for $f\in L_0^2(Y_m)$, $m\in\mathcal{L}$. This vector is easy to identify: we have

$$\int_X \alpha(x)f(\rho_m(F_x))d\mu(x)=\sum_{y\in Y_\ell} f(y)\Bigl(\int_{\{\rho_m(F_x)=y\}}\alpha(x)d\mu(x)\Bigr)$$

which means that $T_1(\alpha)$ is the complex-conjugate of the projection to $L_0^2(Y_m)$ of the function

$$y\mapsto \frac{1}{\nu_m(y)}\int_{\{\rho_m(F_x)=y\}}\alpha(x)d\mu(x)$$

on Y_m. For $m=\{\ell\}$, this projection is obtained by subtracting the contribution of the constant function, i.e., subtracting the average over y: it is

$$\begin{aligned} y\mapsto &\frac{1}{\nu(y)}\int_{\{\rho_m(F_x)=y\}}\alpha(x)d\mu(x)-\sum_y\int_{\{\rho_m(F_x)=y\}}\alpha(x)d\mu(x) \\ &=\frac{1}{\nu(y)}\int_{\{\rho_m(F_x)=y\}}\alpha(x)d\mu(x)-\int_X\alpha(x)d\mu(x). \end{aligned}$$

In the case of counting measure and a uniform density ν, this becomes the quantity

$$\sum_{\rho_m(F_x)=y}\alpha(x)-\frac{1}{|Y_\ell|}\sum_x\alpha(x)$$

after multiplying by $\nu(y)$, which is a typical "error term" appearing in sieve axioms.

To prove Proposition 2.4, we start with two lemmas. For $m\in S(\Lambda)$, $y\in Y_m$, an element φ of the basis $\mathcal{B}_m$, and a square-integrable function $\alpha\in L^2(X,\mu)$, we denote

$$S(m,y)=\int_{\{\rho_m(F_x)=y\}}\alpha(x)d\mu(x),\quad\text{and}\quad S(\varphi)=\int_X\alpha(x)\varphi(\rho_m(F_x))d\mu(x), \tag{2.5}$$

where the integral is defined because $\mu(X)<+\infty$ by assumption. The first lemma is the following:

Lemma 2.6. *We have for all $\ell \in \Lambda$ the relation*

$$\sum_{\varphi \in \mathcal{B}_\ell^*} |S(\varphi)|^2 = \sum_{y \in Y_\ell} \frac{|S(\ell, y)|^2}{\nu(y)} - \Bigl| \int_X \alpha(x) d\mu(x) \Bigr|^2.$$

Proof. Expanding the square by Fubini's Theorem, the left-hand side is

$$\int_X \int_X \alpha(x)\overline{\alpha(y)} \sum_{\varphi \in \mathcal{B}_\ell^*} \varphi(\rho_\ell(F_x))\overline{\varphi(\rho_\ell(F_y))} d\mu(x) d\mu(y).$$

Since $(\varphi)_{\varphi \in \mathcal{B}_\ell}$ is an orthonormal basis of the space of functions on Y_ℓ, expanding the delta function $z \mapsto \delta(y, z)$ in the basis gives

$$\sum_{\varphi \in \mathcal{B}_\ell} \varphi(y)\overline{\varphi(z)} = \frac{1}{\nu(y)} \delta(y, z).$$

Taking on the right-hand side the contribution of the constant function 1, we get in particular

$$\sum_{\varphi \in \mathcal{B}_\ell^*} \varphi(\rho_\ell(F_x))\overline{\varphi(\rho_\ell(F_y))} = \frac{1}{\nu(F_x)} \delta(F_x, F_y) - 1.$$

Inserting this in the first relation, we obtain

$$\begin{aligned}
\sum_{\varphi \in \mathcal{B}_\ell^*} |S(\varphi)|^2 &= \int\int_{\{F_x = F_y\}} \frac{\alpha(x)\overline{\alpha(y)}}{\nu(F_x)} d\mu(x) d\mu(y) - \int_X \int_X \alpha(x)\overline{\alpha(y)} d\mu(x) d\mu(y) \\
&= \sum_{z \in Y_\ell} \frac{1}{\nu(z)} \int\int_{\{F_x = z = F_y\}} \alpha(x)\overline{\alpha(y)} d\mu(x) d\mu(y) - \Bigl| \int_X \alpha(x) d\mu(x) \Bigr|^2 \\
&= \sum_{y \in Y_\ell} \frac{|S(\ell, y)|^2}{\nu(y)} - \Bigl| \int_X \alpha(x) d\mu(x) \Bigr|^2,
\end{aligned}$$

as desired. □

Here is the next lemma.

Lemma 2.7. *Let $(\Psi, \Upsilon, \Omega, \mathcal{L}^*)$ be as above, and let $\mathcal{L}$ be any sieve support associated to $\mathcal{L}^*$. For any square-integrable function $x \mapsto \alpha(x)$ on X supported on the sifted set $S(X, \Omega; \mathcal{L}^*) \subset X$, and for any $m \in \mathcal{L}$, we have*

$$\sum_{\varphi \in \mathcal{B}_m^*} |S(\varphi)|^2 \geqslant \Bigl| \int_X \alpha(x) d\mu(x) \Bigr|^2 \prod_{\ell | m} \frac{\nu(\Omega_\ell)}{\nu(Y_\ell - \Omega_\ell)},$$

where $S(\varphi)$ is given by (2.5).

Proof. Since this does not change the sifted set, we may replace $\mathcal{L}$ if necessary by the full power set of $\mathcal{L}^*$. Then, as in the classical case (see e.g. [IK, Lemma 7.15]), the proof proceeds by induction on the number of elements in m. If $m = \emptyset$, the inequality is trivial (there is equality, in fact). If $m = \{\ell\}$ with $\ell \in \Lambda$ (in the arithmetic case, m is a prime), then $\ell \in \mathcal{L}^*$ by (2.1). Using Cauchy's inequality and the definition of the sifted set with the assumption on $\alpha(x)$ to restrict the support of integration to elements where $\rho_\ell(F_x) \notin \Omega_\ell$, we obtain:

$$\begin{aligned}
\Bigl| \int_X \alpha(x) d\mu(x) \Bigr|^2 = \Bigl| \sum_{\substack{y \in Y_\ell \\ y \notin \Omega_\ell}} S(\ell, y) \Bigr|^2 &\leqslant \Bigl(\sum_{y \notin \Omega_\ell} \nu(y) \Bigr) \Bigl(\sum_{y \in Y_\ell} \frac{|S(\ell, y)|^2}{\nu(y)} \Bigr) \\
&= \nu(Y_\ell - \Omega_\ell) \sum_{y \in Y_\ell} \frac{|S(\ell, y)|^2}{\nu(y)} \\
&= \nu(Y_\ell - \Omega_\ell) \Bigl\{ \sum_{\varphi \in \mathcal{B}_\ell^*} |S(\varphi)|^2 + \Bigl| \int_X \alpha(x) d\mu(x) \Bigr|^2 \Bigr\}
\end{aligned}$$

(by Lemma 2.6), hence the result by moving $|\int \alpha(x)d\mu|^2$ on the left-hand side, since $\nu(Y_\ell)=1$.

The induction step is now immediate, relying on the fact that the function α is arbitrary and the sets $\mathcal{B}_m^*$ are "multiplicative": for $m\in\mathcal{L}$, not a singleton, write $m=m_1m_2=m_1\cup m_2$ with m_1 and m_2 non-empty. Then we have[3]

$$\sum_{\varphi\in\mathcal{B}^*_{m_1m_2}}|S(\varphi)|^2=\sum_{\varphi_1\in\mathcal{B}^*_{m_1}}\sum_{\varphi_2\in\mathcal{B}^*_{m_2}}|S(\varphi_1\otimes\varphi_2)|^2$$

where $\varphi_1\otimes\varphi_2$ is the function $(y,z)\mapsto\varphi_1(y)\varphi_2(z)$. For fixed φ_1, we can express the inner sum as

$$S(\varphi_1\otimes\varphi_2)=\int_X\beta(x)\varphi_2(\rho_{m_2}(F_x))d\mu(x)$$

with $\beta(x)=\alpha(x)\varphi_1(\rho_{m_1}(F_x)))$, which is also supported on $S(X,\Omega;\mathcal{L}^*)$. By the induction hypothesis applied first to m_2, then to m_1, we obtain

$$\begin{aligned}\sum_{\varphi\in\mathcal{B}^*_{m_1m_2}}|S(\varphi)|^2&\geqslant\prod_{\ell|m_2}\frac{\nu(\Omega_\ell)}{\nu(Y_\ell-\Omega_\ell)}\sum_{\varphi_1\in\mathcal{B}^*_{m_1}}\Bigl|\int_X\beta(x)d\mu(x)\Bigr|^2\\&=\prod_{\ell|m_2}\frac{\nu(\Omega_\ell)}{\nu(Y_\ell-\Omega_\ell)}\sum_{\varphi_1\in\mathcal{B}^*_{m_1}}|S(\varphi_1)|^2\geqslant\prod_{\ell|m_1m_2}\frac{\nu(\Omega_\ell)}{\nu(Y_\ell-\Omega_\ell)}\Bigl|\int_X\alpha(x)d\mu(x)\Bigr|^2.\end{aligned}$$

□

Now the proof of Proposition 2.4 is easy.

Proof of Proposition 2.4. Take $\alpha(x)$ to be the characteristic function of $S(X,\Omega;\mathcal{L}^*)$ and sum over $m\in\mathcal{L}$ the inequality of Lemma 2.7; since

$$\int_X\alpha(x)d\mu(x)=\int_X\alpha(x)^2d\mu(x)=|S(X,\Omega;\mathcal{L}^*)|,$$

it follows that

$$|S(X,\Omega;\mathcal{L}^*)|^2\sum_{m\in\mathcal{L}}\prod_{\ell|m}\frac{\nu(\Omega_\ell)}{\nu(Y_\ell-\Omega_\ell)}\leqslant\sum_{m\in\mathcal{L}}\sum_{\varphi\in\mathcal{B}^*_m}|S(\varphi)|^2\leqslant\Delta|S(X,\Omega;\mathcal{L}^*)|,$$

hence the result. □

Example 2.8. In the classical case, with $Y=\mathbf{Z}$ and $Y_\ell=\mathbf{Z}/\ell\mathbf{Z}$, we can identity Y_m with $\mathbf{Z}/m\mathbf{Z}$ by the Chinese Remainder Theorem. With $\nu(y)=1/\ell$ for all ℓ and all y, the usual basis of functions on Y_m is that of additive characters

$$x\mapsto e\Bigl(\frac{ax}{m}\Bigr)$$

for $a\in\mathbf{Z}/m\mathbf{Z}$. It is easy to check that such a character belongs to $\mathcal{B}^*_m$ if and only if a and m are coprime.

At this point a "large sieve inequality" will be an estimate for the quantity Δ. There are various techniques available for this purpose; see [IK, Ch. VII] for a survey of some of them.

The simplest technique is to use the familiar duality principle for bilinear forms or linear operators. Since Δ is the square of the norm of a linear operator, it is the square of the norm of its adjoint. Hence we have:

Lemma 2.9. *Let $\Psi=(Y,\Lambda,(\rho_\ell))$ be a sieve setting, (X,μ,F) a siftable set, $\mathcal{L}$ a sieve support associated to $\mathcal{L}^*$. Fix orthonormal basis $\mathcal{B}_\ell$ and define $\mathcal{B}_m$ as above. Then the large sieve constant $\Delta(X,\mathcal{L})$ is the smallest number Δ such that*

$$\int_X\Bigl|\sum_{m\in\mathcal{L}}\sum_{\varphi\in\mathcal{B}^*_m}\beta(m,\varphi)\varphi(\rho_m(F_x))\Bigr|^2d\mu(x)\leqslant\Delta\sum_m\sum_\varphi|\beta(m,\varphi)|^2\tag{2.6}$$

for all vectors of complex numbers $(\beta(m,\varphi))$.

[3] Here we use the enlargement of $\mathcal{L}$ at the beginning to ensure that $m_i\in\mathcal{L}$.

The point is that this leads to another bound for Δ in terms of bounds for the "dual" sums $W(\varphi, \varphi')$ obtained by expanding the square in this inequality, i.e.

$$W(\varphi, \varphi') = \int_X \varphi(\rho_m(F_x))\overline{\varphi'(\rho_n(F_x))}d\mu(x),$$

where $\varphi \in \mathcal{B}_m$ and $\varphi' \in \mathcal{B}_n$ for some m and n in $S(\Lambda)$. Precisely, we have:

Proposition 2.10. *Let $\Psi = (Y, \Lambda, (\rho_\ell))$ be a sieve setting, $\Upsilon = (X, \mu, F)$ a siftable set, $\mathcal{L}^*$ a prime sieve support and $\mathcal{L}$ an associated sieve support. Then the large sieve constant satisfies*

$$\Delta \leqslant \max_{n\in\mathcal{L}} \max_{\varphi\in\mathcal{B}_n^*} \sum_{m\in\mathcal{L}} \sum_{\varphi'\in\mathcal{B}_m^*} |W(\varphi, \varphi')|. \tag{2.7}$$

Proof. Expanding the left-hand side of (2.6), we have

$$\int_X \Bigl|\sum_{m\in\mathcal{L}} \sum_{\varphi\in\mathcal{B}_m^*} \beta(m, \varphi)\varphi(\rho_m(F_x))\Bigr|^2 d\mu(x) = \sum\sum_{m,n} \sum\sum_{\varphi,\varphi'} \beta(m, \varphi)\overline{\beta(n, \varphi')}W(\varphi, \varphi')$$

and applying $|uv| \leqslant \frac{1}{2}(|u|^2 + |v^2|)$ the result follows as usual. □

The point is that sieve results are now reduced to individual *uniform* estimates for the "sums" $W(\varphi, \varphi')$. Note that, here, the choice of the orthonormal basis may well be very important in estimating $W(\varphi, \varphi')$ and therefore Δ.

However, at least formally, we can proceed in full generality as follows, where the idea is that in applications $\rho_m(F_x)$ should range fairly equitably (with respect to the density ν_m) over the elements of Y_m, so the sum $W(\varphi, \varphi')$ should be estimated by exploiting the "periodicity" of $\varphi(\rho_m(F_x))\overline{\varphi'(\rho_n(F_x))}$. To do this, we introduce further notation.

Let m, n be two elements of $S(\Lambda)$, $\varphi \in \mathcal{B}_m$, $\varphi' \in \mathcal{B}_n$. Let $d = m \cap n$ be the intersection (g.c.d. in the case of integers) of m and n, and write $m = m'd = m' \cup d$, $n = n'd = n' \cup d$ (disjoint unions). According to the multiplicative definition of $\mathcal{B}_m$ and $\mathcal{B}_n$, we can write

$$\varphi = \varphi_{m'} \otimes \varphi_d, \quad \varphi' = \varphi'_{n'} \otimes \varphi'_d$$

for some unique basis elements $\varphi_{m'} \in \mathcal{B}_{m'}$, φ_d, $\varphi'_d \in \mathcal{B}_d$ and $\varphi'_{n'} \in \mathcal{B}_{n'}$.

Let $[m, n] = mn = m \cup n$ be the "l.c.m" of m and n. We have the decomposition

$$Y_{[m,n]} = Y_{m'} \times Y_d \times Y_{n'},$$

the (not necessarily surjective) map $\rho_{[m,n]} : Y \to Y_{[m,n]}$ and the function

$$[\varphi, \overline{\varphi'}] = \varphi_{m'} \otimes (\varphi_d\overline{\varphi'_d}) \otimes \overline{\varphi'_{n'}} : (y_1, y_d, y_2) \mapsto \varphi_{m'}(y_1)\varphi_d(y_d)\overline{\varphi'_d(y_d)\varphi'_{n'}(y_2)}, \tag{2.8}$$

(which is not usually a basis element in $\mathcal{B}_{[m,n]}$).

The motivation for all this is the following tautology:

Lemma 2.11. *Let m, n, φ and φ' be as before. We have*

$$[\varphi, \overline{\varphi'}](\rho_{[m,n]}(y)) = \varphi(\rho_m(y))\overline{\varphi'(\rho_n(y))}$$

for all $y \in Y$, hence

$$W(\varphi, \varphi') = \int_X [\varphi, \overline{\varphi'}](\rho_{[m,n]}(F_x))d\mu(x).$$

Now we can hope to split the integral according to the value of $y = \rho_{[m,n]}(F_x)$ in $Y_{[m,n]}$, and evaluate it by summing the main term in an equidistribution statement.

More precisely, for $d \in S(\Lambda)$ and $y \in Y_d$, we *define* $r_d(X; y)$ as the "error term" in the expected equidistribution statement:

$$|\{\rho_d(F_x) = y\}| = \int_{\{\rho_d(F_x)=y\}} d\mu(x) = \nu_d(y)|X| + r_d(X; y). \tag{2.9}$$

Then we can write $W(\varphi,\varphi')$ as described before:

$$
\begin{aligned}
W(\varphi,\varphi') &= \int_X [\varphi,\overline{\varphi'}](\rho_{[m,n]}(F_x))d\mu(x)\\
&= \sum_{y\in Y_{[m,n]}} [\varphi,\overline{\varphi'}](y)\int_{\{\rho_{[m,n]}(F_x)=y\}} d\mu(x)\\
&= m([\varphi,\overline{\varphi'}])|X| + O\Bigl(\sum_{y\in Y_{[m,n]}} \|[\varphi,\overline{\varphi'}]\|_\infty |r_{[m,n]}(X;y)|\Bigr) \qquad (2.10)
\end{aligned}
$$

after inserting (2.9), where the implied constant is of modulus $\leqslant 1$ and

$$m([\varphi,\overline{\varphi'}]) = \sum_{y\in Y_{[m,n]}} \nu_{[m,n]}(y)[\varphi,\overline{\varphi'}](y) = \langle[\varphi,\overline{\varphi'}],1\rangle,$$

the inner product in $L^2(Y_{[m,n]})$. One would then hope that $m([\varphi,\varphi'])$ is the delta-symbol $\delta((m,\varphi),(n,\varphi'))$ which would select the diagonal in the main term of the sums $W(\varphi,\varphi')$. In Sections 4 and 6, we will see how to evaluate this quantity for the special case of group and coset sieves. But first, a short digression...

3. The "dual" sieve

The equivalent definition of the large sieve constant by means of the duality principle (i.e, Lemma 2.9) is quite useful in itself. For instance, it yields the following type of sieve inequality, which in the classical case goes back to Rényi.

Proposition 3.1. *Let $(Y,\Lambda,(\rho_\ell))$ be a sieve setting, (X,μ,F) a siftable set and $\mathcal{L}^*$ a prime sieve support. Let Δ be the large sieve constant for $\mathcal{L}=\mathcal{L}^*$.*[4] *Then for any sifting sets (Ω_ℓ), we have*

$$\int_X \Bigl(P(x,\mathcal{L})-P(\mathcal{L})\Bigr)^2 d\mu(x) \leqslant \Delta Q(\mathcal{L}) \qquad (3.1)$$

where

$$P(x,\mathcal{L}) = \sum_{\substack{\ell\in\mathcal{L}\\ \rho_\ell(F_x)\in\Omega_\ell}} 1,\qquad P(\mathcal{L})=\sum_{\ell\in\mathcal{L}}\nu(\Omega_\ell),\qquad Q(\mathcal{L})=\sum_{\ell\in\mathcal{L}}\nu(\Omega_\ell)(1-\nu(\Omega_\ell)). \qquad (3.2)$$

Proof. By expanding the characteristic function $\chi(\Omega_\ell)$ of $\Omega_\ell\subset Y_\ell$ in the orthonormal basis $\mathcal{B}_\ell$, we obtain

$$P(x,\mathcal{L}) = P(\mathcal{L}) + \sum_{\ell\in\mathcal{L}}\sum_{\varphi\in\mathcal{B}_\ell^*} \beta(\ell,\varphi)\varphi(\rho_\ell(F_x)),$$

where

$$\beta(\ell,\varphi) = \sum_{y\in\Omega_\ell} \nu_\ell(y)\overline{\varphi}(y),$$

and we used the fact that $\mathcal{B}_\ell^* = \mathcal{B}_\ell - \{1\}$ for $\ell\in\Lambda$. Thus we get

$$
\begin{aligned}
\int_X \Bigl(P(x,\mathcal{L})-P(\mathcal{L})\Bigr)^2 d\mu(x) &= \int_X \Bigl|\sum_{\ell\in\mathcal{L}}\sum_{\varphi\in\mathcal{B}_\ell^*}\beta(\ell,\varphi)\varphi(\rho_\ell(F_x))\Bigr|^2 d\mu(x)\\
&\leqslant \Delta\sum_{\ell\in\mathcal{L}}\sum_{\varphi\in\mathcal{B}_\ell^*}|\beta(\ell,\varphi)|^2
\end{aligned}
$$

by applying (2.6). Since we have

$$\sum_{\varphi\in\mathcal{B}_\ell^*}|\beta(\ell,\varphi)|^2 = \sum_{\varphi\in\mathcal{B}_\ell}|\beta(\ell,\varphi)|^2 - |\beta(\ell,1)|^2 = \|\chi(\Omega_\ell)\|^2 - \nu(\Omega_\ell)^2 = \nu(\Omega_\ell)(1-\nu(\Omega_\ell)),$$

this implies the result. ∎

[4] Precisely, $\mathcal{L}$ is the set of singletons $\{\ell\}$ for $\ell\in\mathcal{L}^*$.

In particular, since $P(x,\mathcal{L})=0$ for $x\in S(X;\Omega,\mathcal{L}^*)$ and $Q(\mathcal{L})\leqslant P(\mathcal{L})$, we get (by positivity again) the estimate

$$|S(X;\Omega,\mathcal{L}^*)|\leqslant \Delta P(\mathcal{L})^{-1},$$

which is the analogue of the inequalities used e.g. by Gallagher in [G, Th. A], and by the author in [Ko1]. This inequality also follows from Proposition 2.4 if we take $\mathcal{L}$ containing only singletons (in the arithmetic case, this means using only the primes), since we get the estimate

$$|S(X,\Omega;\mathcal{L}^*)|\leqslant \Delta H^{-1}\qquad \text{with}\qquad H=\sum_{\ell\in\mathcal{L}}\frac{\nu(\Omega_\ell)}{\nu(Y_\ell-\Omega_\ell)}\geqslant\sum_{\ell\in\mathcal{L}}\nu(\Omega_\ell)=P(\mathcal{L})$$

(in fact, by Cauchy's inequality, we have $P(\mathcal{L})^2\leqslant HQ(\mathcal{L})$).

This type of result is also related to Turán's method in probabilistic number theory. In counting primes with the classical setting, or more generally in "small sieve" situations, it may seem quite weak (it only implies $\pi(X)\ll X(\log\log X)^{-1}$). However, it is really a different type of statement, which has additional flexibility: for instance, it still implies that for $X\geqslant 3$ we have

$$|\{n\leqslant X\ \mid\ \omega(n)<\kappa\log\log X\}|\ll\frac{X}{\log\log X}$$

for any $\kappa\in]0,1[$, the implied constant depending only on κ. This estimate is now *of the right order of magnitude*, and this shows in particular that one can not hope to improve (3.1) by using information related to all "squarefree" numbers; in other words, Proposition 2.4 *can not* be extended "as is" to an upper bound for the variance on the left of (3.1).

These remarks indicate that Proposition 3.1 has its own interest in cases where the "stronger" form of the large sieve is in fact not adapted to the type of situation considered. In Section 10, we will describe an amusing use of the inequality (3.1), where the "pure sieve" bound would indeed be essentially trivial.

4. Group and conjugacy sieves

We now come to the description of a more specific type of sieve setting, related to a group structure on Y. Together with the coset sieves of Section 6, this exhausts most examples of applications we know at the moment.

A *group sieve* corresponds to a sieve setting $\Psi=(G,\Lambda,(\rho_\ell))$ where G is a group and the maps $\rho_\ell : G\rightarrow G_\ell$ are homomorphisms onto finite groups. A *conjugacy sieve*, similarly, is a sieve setting $\Psi=(G,\Lambda,(\rho_\ell))$ where $\rho_\ell : G\rightarrow G_\ell^\sharp$ is a surjective map from G to the finite set of conjugacy classes $G_\ell^\sharp$ of a finite group G_ℓ, that factors as

$$G\rightarrow G_\ell\rightarrow G_\ell^\sharp$$

where $G\rightarrow G_\ell$ is a surjective homomorphism. Obviously, if G is abelian, group and conjugacy sieves are identical, and any group sieve induces a conjugacy sieve.

The group structure suggests a natural choice of orthonormal basis $\mathcal{B}_\ell$ for functions on G_ℓ or $G_\ell^\sharp$, as well as natural densities ν_ℓ. We start with the simpler conjugacy sieve.

From the classical representation theory of finite groups (see, e.g., [S2]), we know that for any $\ell\in\Lambda$, the functions

$$y\mapsto \operatorname{Tr}\pi(y),$$

on G_ℓ, where π runs over the set Π_ℓ of (isomorphism classes of) irreducible linear representations $\pi : G_\ell\rightarrow GL(V_\pi)$, form an orthonormal basis of the space $\mathcal{C}(G_\ell)$ of functions on G_ℓ invariant under conjugation, with the inner product

$$\langle f,g\rangle=\frac{1}{|G_\ell|}\sum_{y\in G_\ell}f(y)\overline{g(y)}.$$

Translating this statement to functions on the set $G_\ell^\sharp$ of conjugacy classes, this means that the functions

$$\varphi(y^\sharp)=\operatorname{Tr}\pi(y^\sharp)$$

on $G_\ell^\sharp$ form an orthonormal basis $\mathcal{B}_\ell$ of $L^2(G_\ell^\sharp)$ with the inner product

$$\langle f, g\rangle = \frac{1}{|G_\ell|}\sum_{y^\sharp\in G} |y^\sharp| f(y^\sharp)\overline{g(y^\sharp)}.$$

Moreover, the trivial representation 1 of G_ℓ has for character the constant function 1, so we can use the basis $\mathcal{B}_\ell = (\mathrm{Tr}\,\pi(y^\sharp))_\pi$ for computing the large sieve constant if the density

$$\nu_\ell(y^\sharp) = \frac{|y^\sharp|}{|G_\ell|}$$

is used. Note that this is the image on $G_\ell^\sharp$ of the uniform density on G_ℓ.

Note also that in the abelian case, the representations are one-dimensional, and the basis thus described is the basis of characters of G_ℓ, with the uniform density, i.e., that of group homomorphisms $G_\ell \to \mathbf{C}^\times$ with

$$\langle f, g\rangle = \frac{1}{|G_\ell|}\sum_{y\in G_\ell} f(y)\overline{g(y)}.$$

Coming back to a general group sieve, the basis and densities extended to the sets

$$G_m^\sharp = \prod_{\ell|m} G_\ell^\sharp$$

for $m \in S(\Lambda)$ have a similar interpretation. Indeed, $G_m^\sharp$ identifies clearly with the set of conjugacy classes of the finite group $G_m = \prod G_\ell$. The density ν_m is therefore still given by

$$\nu_m(y^\sharp) = \frac{|y^\sharp|}{|G_m|}.$$

Also, it it well-known that the irreducible representations of G_m are of the form

$$\pi\ :\ g \mapsto \boxtimes_{\ell|m}\pi_\ell(g)$$

for some uniquely defined irreducible representations π_ℓ of G_ℓ, where $\boxtimes$ is the external tensor product defined by

$$g = (g_\ell) \mapsto \bigotimes_{\ell|m} \rho_\ell(g_\ell).$$

In other words, the set Π_m of irreducible linear representations of G_m is identified canonically with $\prod \Pi_\ell$. Moreover, the character of a representation of G_m of this form is simply

$$\mathrm{Tr}\,\pi(g) = \prod_{\ell|m} \mathrm{Tr}\,\pi_\ell(g_\ell),$$

so that the basis $\mathcal{B}_m$ obtained from $\mathcal{B}_\ell$ is none other than the basis of functions $y^\sharp \mapsto \mathrm{Tr}\,\pi(y^\sharp)$ for π ranging over Π_m.

Given a siftable set (X, μ, F) associated to a conjugacy sieve $(G, \Lambda, (\rho_\ell))$, the sums $W(\varphi, \varphi')$ become

$$W(\pi, \tau) = \int_X \mathrm{Tr}\,\pi(\rho_m(F_x))\overline{\mathrm{Tr}\,\tau(\rho_n(F_x))}d\mu(x) \tag{4.1}$$

for irreducible representations π and τ of G_m and G_n respectively, which can usually be interpreted as *exponential sums* (or integrals) over X, since the character values, as traces of matrices of finite order, are sums of finitely many roots of unity.

We summarize briefly by reproducing the general sieve results in this context, phrasing things as related to the conjugacy sieve induced from a group sieve (which seems most natural for applications). In such a situation, the sieving sets Ω_ℓ are naturally given as conjugacy-invariant subsets of G_ℓ, and are identified with subsets $\Omega_\ell^\sharp$ of $G_\ell^\sharp$. Note that we have then $\nu_m(\Omega_\ell^\sharp) = |\Omega_\ell|/|G_\ell|$.

Proposition 4.1. *Let $(G, \Lambda, (\rho_\ell))$ be a group sieve setting, (X, μ, F) an associated siftable set. For any prime sieve support $\mathcal{L}^*$ and an associated sieve support $\mathcal{L}$ satisfying* (2.1)*, and for any conjugacy invariant sifting sets (Ω_ℓ), we have*

$$|S(X, \Omega; \mathcal{L}^*)| \leqslant \Delta H^{-1}$$

where Δ is the smallest non-negative real number such that

$$\sum_{m \in \mathcal{L}} \sum_{\pi \in \Pi_m^*} \Big| \int_X \alpha(x) \operatorname{Tr} \pi(\rho_m(F_x)) d\mu(x) \Big|^2 \leqslant \Delta \int_X |\alpha(x)|^2 d\mu(x)$$

for all square-integrable function $\alpha \in L^2(X, \mu)$, where π ranges over the set Π_m^* of primitive *irreducible linear representations of G_m, i.e., those such that no component π_ℓ for $\ell \mid m$ is trivial, and where*

$$H = \sum_{m \in \mathcal{L}} \prod_{\ell \mid m} \frac{|\Omega_\ell|}{|G_\ell| - |\Omega_\ell|}.$$

Moreover we have

$$\Delta \leqslant \max_{m \in \mathcal{L}} \max_{\pi \in \Pi_m^*} \sum_{n \in \mathcal{L}} \sum_{\tau \in \Pi_n^*} |W(\pi, \tau)|,$$

where

$$W(\pi, \tau) = \int_X \operatorname{Tr} \pi(\rho_m(F_x)) \overline{\operatorname{Tr} \tau(\rho_n(F_x))} d\mu(x).$$

The general sieve setting can also be applied to problems where the sieving sets are not conjugacy-invariant, using the basis of *matrix coefficients* of irreducible representations. Let $(G, \Lambda, (\rho_\ell))$ be a group sieve setting. For each ℓ and each irreducible representation $\pi \in \Pi_\ell$, choose an orthonormal basis $(e_{\pi,i})$ of the space V_π of the representation (with respect to a G_ℓ-invariant inner product $\langle \cdot, \cdot \rangle_\pi$). Then (see, e.g., [Kn, §I.5], which treats compact groups), the family $\mathcal{B}_\ell$ of functions of the type

$$\varphi_{\pi,e,f} \;:\; x \mapsto \sqrt{\dim \pi} \langle \pi(x) e, f \rangle_\pi, \qquad e = e_{\pi,1}, \ldots, e_{\pi,\ell}, \quad f = e_{\pi,1}, \ldots, e_{\pi,\ell}$$

is an orthonormal basis of $L^2(G_\ell)$ for the inner product

$$\langle f, g \rangle = \frac{1}{|G_\ell|} \sum_{x \in G_\ell} f(x) \overline{g(x)},$$

i.e., corresponding to the density $\nu_\ell(x) = 1/|G_\ell|$ for all $x \in G_\ell$. Moreover, for $\pi = 1$, and an arbitrary choice of $e \in \mathbf{C}$ with $|e| = 1$, the function $\varphi_{1,e,e} \in \mathcal{B}_\ell$ is the constant function 1.

If we extend the basis $\mathcal{B}_\ell$ to orthonormal basis $\mathcal{B}_m$ of $L^2(G_m)$ for all $m \in S(\Lambda)$, by multiplicativity, the functions in $\mathcal{B}_m$ are of the type

$$\varphi_{\pi,e,f} \;:\; (x_\ell) \mapsto \sqrt{\dim \pi} \prod_{\ell \mid m} \langle \pi_\ell(x_\ell) e_\ell, f_\ell \rangle_{\pi_\ell}$$

where $e = \otimes e_\ell$ and $f = \otimes f_\ell$ run over elements of the orthonormal basis

$$\Big(\bigotimes_{\ell \mid m} e_{\pi_\ell, i_\ell} \Big), \qquad 1 \leqslant i_\ell \leqslant \dim \pi_\ell,$$

constructed from the chosen bases $(e_{\pi,i})$ of the components, the inner product on the space of $\boxtimes \pi_\ell$ being the natural G_m-invariant one.

The sums $W(\varphi, \varphi')$ occurring in Proposition 2.10 to estimate the large sieve constant are given by

$$W(\varphi_{\pi,e,f}, \varphi_{\tau,e',f'}) = \sqrt{(\dim \pi)(\dim \tau)} \int_X \langle \pi(\rho_m(F_x)) e, f \rangle_\pi \overline{\langle \tau(\rho_n(F_x)) e', f' \rangle_\tau} d\mu(x). \tag{4.2}$$

If we apply Lemma 2.11 to elements $\varphi_{\pi,e,f}$, $\varphi_{\tau,e',f'}$ of the basis $\mathcal{B}_m$ and $\mathcal{B}_n$ of $L^2(G_m)$, the function $[\varphi_\pi, \overline{\varphi_\tau}]$ which is integrated can be written as a matrix coefficient of the representation

$$[\pi, \bar{\tau}] = \pi_{m'} \boxtimes (\pi_d \otimes \overline{\tau_d}) \boxtimes \overline{\tau_{n'}} \tag{4.3}$$

of $G_{[m,n]}$, where we write $\pi = \pi_{m'} \boxtimes \pi_d$, $\tau = \tau_{n'} \boxtimes \tau_d$, with the obvious meaning of the components $\pi_{m'}$, π_d, τ_d, $\tau_{n'}$, and the bar indicates taking the contragredient representation.

Indeed, we have

$$[\varphi_{\pi,e,f}, \overline{\varphi_{\tau,e',f'}}](x_\ell) = \sqrt{(\dim \pi)(\dim \tau)} \langle [\pi, \bar{\tau}](\rho_{[m,n]}(F_x))\tilde{e}, \tilde{f}\rangle_{[\pi,\bar{\tau}]}$$

for $(x_\ell) \in G_{[m,n]}$, with $\tilde{e} = e \otimes e'$, $\tilde{f} = f \otimes f'$.

Concretely, this means that in order to deal with the sums $W(\varphi, \varphi')$ to estimate the large sieve constant using the basis $\mathcal{B}_m$ of matrix coefficients, it suffices to be able to estimate all integrals of the type

$$\int_X \langle \varpi(F_x)e, f\rangle_\varpi d\mu(x). \tag{4.4}$$

where ϖ is a representation of G that factors through a finite product of groups G_ℓ, and e, f are vectors in the space of the representation ϖ (the inner product being G-invariant). See the proof of Theorem 1.3 for an application of this.

Remark 4.2. Another potentially useful sieve setting associated to a group sieve setting (G, Λ, ρ_ℓ) is obtained by replacing ρ_ℓ with the projections $G \to G_\ell \to G_\ell/K_\ell = Y_\ell$ for $\ell \in \Lambda$, where K_ℓ is an arbitrary subgroup of G_ℓ. Considering the density on Y_ℓ which is image of the uniform density on G_ℓ, an orthonormal basis $\mathcal{B}_\ell$ of $L^2(Y_\ell)$ is then obtained by taking the functions

$$\varphi_{\pi,e,f} \,:\, gK_\ell \mapsto \langle \pi(g)e, f\rangle$$

where π runs over irreducible representations of G_ℓ, e runs over an orthonormal basis of the K_ℓ-invariant subspace in the space V_π of π, and f over a full orthonormal basis of V_π.

Indeed, the restriction on e ensures that such functions are well-defined on G_ℓ/K_ℓ (i.e., the matrix coefficient is K_ℓ-invariant), and since those are matrix coefficients, there only remains to check that they span $L^2(Y_\ell)$. However, the total number of functions is

$$\sum_\pi (\dim \pi)\langle \mathrm{Res}^{G_\ell}_{K_\ell} \pi, 1\rangle_{K_\ell} = \sum_\pi (\dim \pi)\langle \pi, \mathrm{Ind}^{G_\ell}_{K_\ell} 1\rangle_{G_\ell} = \dim \mathrm{Ind}^{G_\ell}_{K_\ell} 1 = |Y_\ell|$$

and since they are independent, the result follows.

Because this basis is a sub-basis of the previous one, any estimate for the large sieve constant for the group sieve will give one for this sieve setting.

5. Elementary and classical examples

We first describe how some classical uses of the large sieve are special cases of the group sieve setting of the previous section, and conclude this section with a "new" example of the general case which is particularly easy to analyze (and of little practical use), and hence somewhat enlightening.

Example 5.1. As already mentioned, the classical large sieve arises from the group sieve setting

$$\Psi = (\mathbf{Z}, \{\text{primes}\}, \mathbf{Z} \to \mathbf{Z}/\ell\mathbf{Z})$$

where the condition for an additive character $x \mapsto e(ax/m)$ of $G_m = (\mathbf{Z}/m\mathbf{Z})$ to be primitive is equivalent with the classical condition that $(a, m) = 1$.

In the most typical case, the siftable sets are

$$X = \{n \geqslant 1 \mid N \leqslant n < N + M\}$$

with $F_x = x$, and the abstract sieving problem becomes the "original" one of finding integers in X which lie outside certain residue classes modulo some primes ℓ.

More generally, take

$$\Psi = (\mathbf{Z}^r, \{\text{primes}\}, \mathbf{Z}^r \to (\mathbf{Z}/\ell\mathbf{Z})^r)$$

(the reduction maps) and $X = \{(a_1, \ldots, a_r) \mid N_i \leqslant a_i < N_i + M_i\}$, with F the identity map again. Then what results is the higher-dimensional large sieve (see e.g. [G]).

For completeness, we recall the estimates available for the large sieve constant in those two situations, when we take $\mathcal{L}^*$ to be the set of primes $\leqslant L$, and $\mathcal{L}$ to be the set of squarefree integers $\leqslant L$, for some $L \geqslant 1$. We write $S(X;\Omega,L)$ instead of $S(X;\Omega,\mathcal{L}^*)$.

Theorem 5.2. *With notation as above, we have $\Delta \leqslant N-1+L^2$ for $r=1$ and $\Delta \leqslant (\sqrt{N}+L)^{2r}$ for all $r \geqslant 1$. In particular, for any sieve problem, we have*

$$|S(X;\Omega,L)| \leqslant (N-1+L^2)H^{-1}, \qquad \text{if } r=1,$$
$$|S(X;\Omega,L)| \leqslant (\sqrt{N}+L)^{2r}H^{-1}, \qquad \text{if } r \geqslant 1,$$

where

$$H = \sideset{}{^\flat}\sum_{m \leqslant L} \prod_{\ell \mid m} \frac{|\Omega_\ell|}{\ell^r - |\Omega_\ell|},$$

the notation $\sum^\flat$ indicating a sum restricted to squarefree numbers.

In the one-variable case, this is due essentially to Montgomery, and to Selberg with the constant $N-1+L^2$, see e.g. [IK, §7.5]; the higher-dimensional case as stated is due to Huxley, see [Hu]. Note that modern treatments deduce such estimates from an analytic inequality which is more general than the ones we used, namely, the inequality

$$\sum_r \Bigl| \sum_{M<n\leqslant M+N} a_n e(n\xi_r) \Bigr|^2 \leqslant (N-1+\delta^{-1}) \sum_n |a_n|^2$$

for arbitrary sets (ξ_r) of elements in $\mathbf{R}/\mathbf{Z}$ which are δ-spaced, i.e., the distance $d(\xi_r,\xi_s)$ in $\mathbf{R}/\mathbf{Z}$ is at least δ if $r \neq s$ (this was first considered by Bombieri and Davenport; see, e.g., [IK, Th. 7.7]). This amounts, roughly, to considering sums

$$\sum_{M<n\leqslant M+N} e((\xi_r - \xi_s)n) = W(\pi_r, \pi_s)$$

where $\pi_r : n \mapsto e(n\xi_r)$ and π_s are representations of $G = \mathbf{Z}$ which *do not* factor through a finite index subgroup. This suggests trying to prove similar inequalities for general groups sieves, i.e., essentially, consider integrals (4.4) for arbitrary (unitary) representations ϖ of G.

Note that for $r=1$, the equidistribution assumption (2.9) becomes

$$\sum_{\substack{N\leqslant n<N+M \\ n\equiv y \,(\mathrm{mod}\, d)}} 1 = \frac{M}{d} + r_d(X;y),$$

which holds with $|r_d(X;y)| \leqslant 1$ for any $y \in \mathbf{Z}/d\mathbf{Z}$. From (6.9) we obtain the estimate $\Delta \leqslant N + L^4$, which is by no means ridiculous. (See Section 6 for the computation of the quantity $m([\varphi,\varphi'])$ for group sieves, or do the exercise).

Classical sieve theory is founded on such assumptions as (2.9), usually stated merely for $y=0$, and on further assumptions concerning the resulting level of distribution, i.e., bounds for $r_d(X;0)$ on average over d in a range as large as possible (compared with the size of X). More general bounds for $r_d(X;y)$ do occur however.

Note that, even if this is classical, the general framework clearly shows that to sieve an arbitrary set of integers $X \subset \{n \mid n \geqslant 1\} \subset \mathbf{Z}$, it suffices (at least up to a point!) to have estimates for exponential sums

$$\sum_{x\in X} e\Bigl(\frac{ax}{m} - \frac{bx}{n}\Bigr)$$

with n, m squarefree and $(a,m)=(b,n)=1$. It suffices, in particular, to have equidistribution of X in (all) arithmetic progressions. This means for instance that some measure of large sieve is usually doable for any sequence for which the classical "small" sieves work. This is of particular interest if X is "sparse", in the sense that e.g. $X\subset\{n\mid N<n\leqslant 2N\}$ for some N with $|X|/N$ going to zero.

It would also be interesting, as a problem in itself, to investigate the values of the large sieve constant when using other sieve support than squarefree integers up to L, for instance when the sieve support is the support of a combinatorial (small) sieve.

Example 5.3. Can the multiplicative large sieve inequality for Dirichlet characters be related to our general setting? Indeed, in at least two ways. First, let $q\geqslant 2$ be given, let G be the multiplicative subgroup of $\mathbf{Q}^\times$ generated by primes $p>q$, and take

$$\Psi=(G,\{\text{primes } \ell>q\},G\rightarrow(\mathbf{Z}/\ell\mathbf{Z})^\times=G_\ell).$$

In that context, we can take

$$X=\{n\leqslant N\mid p\mid n\Rightarrow p>q\},$$

and $F_x=x$, and if $\mathcal{L}^*$ is the set of primes $\leqslant L\leqslant q$, and $\mathcal{L}$ is the set of squarefree numbers $\leqslant L$, the sifted sets become

$$S(X;\Omega,L)=\{n\leqslant N\mid (p\mid n\Rightarrow p>q)\text{ and } n\,(\mathrm{mod}\,\ell)\notin\Omega_\ell\text{ for }\ell\leqslant L\leqslant q\},$$

where $\Omega_\ell\subset(\mathbf{Z}/\ell\mathbf{Z})^\times$. A simple check shows that the inequality defining the large sieve constant Δ becomes

$$\sum_{m\leqslant L}{}^{\flat}\ \sum_{\chi\,(\mathrm{mod}\,m)}{}^{*}\Big|\sum_{n\in X}a_n\chi(n)\Big|^2\leqslant\Delta\sum_{n\in X}|a_n|^2 \tag{5.1}$$

for any complex numbers a_n, where χ runs over primitive characters modulo m, and hence $\Delta\leqslant N-1+L^2$ by the multiplicative large sieve inequality (see e.g. [IK, Th. 7.13]).

Alternately, if we allow the density ν_ℓ to have zeros, we may take the classical sieve setting $Y=\mathbf{Z}$, $Y\rightarrow Y_\ell=\mathbf{Z}/\ell\mathbf{Z}$, $X=\{N<n\leqslant M+N\}$, $F_x=x$, with density

$$\nu_\ell(y)=\frac{1}{\ell-1}\quad\text{if } y\neq 0,\qquad \nu_\ell(0)=0,$$

and then check that since the final statements do not involve the inverse of $\nu_\ell(y)$, although the proofs involved division by $\nu_\ell(y)$, it remains true that for $\Omega_\ell\subset\mathbf{Z}_\ell-\{0\}$, we have

$$|S(X,\Omega;L)|\leqslant\Delta H^{-1}$$

where

$$H=\sum_{m\leqslant L}{}^{\flat}\prod_{\ell\mid m}\frac{|\Omega_\ell|}{\ell-1-|\Omega_\ell|}$$

and Δ is the multiplicative large sieve constant defined by (5.1) (e.g., use a positive perturbation $\nu_{\ell,\varepsilon}(y)>0$ of the density so that $H_\varepsilon\rightarrow H$ and $\Delta_\varepsilon\rightarrow\Delta$, as $\varepsilon\rightarrow 0$). Again, we have $\Delta\leqslant N-1+L^2$.

Example 5.4. Serre [S3] has used a variant of the higher-dimensional large sieve where

$$\Psi=(\mathbf{Z}^r,\{\text{primes}\},\mathbf{Z}^r\rightarrow(\mathbf{Z}/\ell^2\mathbf{Z})^r)$$

and

$$X=\{(x_1,\ldots,x_r)\in\mathbf{Z}^r\mid |x_i|\leqslant N\}$$

with $F_x=x$. With suitable sifting sets, this provides estimates for the number of trivial specializations of elements of 2-torsion in the Brauer group of $\mathbf{Q}(T_1,\ldots,T_r)$.

Example 5.5. Here is a new example, which is a number field analogue of the situation of [Ko1] (described also in the Section 11). It is related to Serre's discussion in [S2] of a higher-dimensional Chebotarev density theorem over number fields (see also [P] for an independent treatment with more details). Let $Y/\mathbf{Z}$ be a separated scheme of finite type, and let $Y_\ell \rightarrow Y$ be a family of étale Galois coverings,[5] corresponding to surjective maps $G = \pi_1(Y, \bar{\eta}) \rightarrow G_\ell$. The sieve setting is $(G, \{\text{primes}\}, G \rightarrow G_\ell)$. Now let $|Y|$ denote the set of closed points of Y, which means those where the residue field $k(y)$ is finite, and let

$$X = \{y \in |Y| \mid |k(y)| \leqslant T\}$$

for some $T \geqslant 2$, which is finite. For $y \in X$, denote by $F_x \in G$ the corresponding geometric Frobenius automorphism (or conjugacy class rather) to obtain a siftable set $(X, \text{counting measure}, F)$ associated with the conjugacy sieve. It should be possible to obtain a large sieve inequality in this context, at least assuming GRH and the Artin conjecture.

Note that if Y is the spectrum of the ring of integers in some number field (or even $Y = \mathrm{Spec}(\mathbf{Z})$ itself), this becomes the sieve for Frobenius considered by Zywina [Z], with applications (under GRH) to the Lang-Trotter conjecture, and to Koblitz's conjecture for elliptic curves over number fields.

Example 5.6. The next example illustrates the general sieve setting, showing that it includes (and extends) the inclusion-exclusion familiar in combinatorics and probability theory, and also that the large sieve inequality is sharp in this general context (i.e., there may be equality $|S(X, \Omega; \mathcal{L}^*)| = \Delta H^{-1}$).

Let $(\Omega, \Sigma, \mathbf{P})$ be a probability space and $A_\ell \subset \Sigma$, for $\ell \in \Lambda$, a countable family of events. Consider the event

$$A = \{\omega \in \Omega \mid \omega \notin A_\ell \text{ for any } \ell \in \Lambda\}.$$

For $m \in S(\Lambda)$, denote

$$A_m = \bigcap_{\ell \in m} A_\ell, \qquad A_\emptyset = \Omega.$$

If Λ is finite, which we now assume, the inclusion-exclusion formula is

$$\mathbf{P}(A) = \sum_{m \in S(\Lambda)} (-1)^{|m|} \mathbf{P}(A_m),$$

and in particular, if the events are independent (as a whole), we have

$$\mathbf{P}(A_m) = \prod_{\ell \in m} \mathbf{P}(A_\ell), \qquad \text{and} \qquad \mathbf{P}(A) = \prod_{\ell \in \Lambda} (1 - \mathbf{P}(A_\ell)).$$

Take the sieve setting $(\Omega, \Lambda, \mathbf{1}_{A_\ell})$, where $\mathbf{1}_B$ is the characteristic function of an event B, with $Y_\ell = \{0, 1\}$ for all ℓ, and the siftable set $(\Omega, \mathbf{P}, \mathrm{Id})$. Choose the density $\nu_\ell = \mathbf{1}_{A_\ell}(P)$, i.e., put

$$\nu_\ell(1) = \mathbf{P}(A_\ell), \qquad \nu_\ell(0) = 1 - \mathbf{P}(A_\ell).$$

With sieving sets $\Omega_\ell = \{1\}$ for $\ell \in \Lambda$, we have precisely $S(X, \Omega; \Lambda) = A$.

The large sieve inequality yields

$$\mathbf{P}(A) \leqslant \Delta H^{-1}$$

where

$$H = \sum_{m \in \mathcal{L}} \prod_{\ell \in m} \frac{\mathbf{P}(A_\ell)}{1 - \mathbf{P}(A_\ell)},$$

and Δ is the large sieve constant for the sieve support $\mathcal{L}$, which may be any collection of subsets of Λ such that $\{\ell\} \in \mathcal{L}$ for all $\ell \in \Lambda$.

[5] Or better with "controlled" ramification, if not étale, since this is likely to be needed for some natural applications.

Coming to the large sieve constant, note that $L^2_0(Y_\ell)$ is one-dimensional for all ℓ, hence so is $L^2_0(Y_m)$ for all m (including $m=\emptyset$). The basis function φ_ℓ for $L^2_0(Y_\ell)$ (up to multiplication by a complex number with modulus 1) is given by

$$\varphi_\ell(y) = \frac{y-p_\ell}{\sqrt{p_\ell(1-p_\ell)}}$$

where $p_\ell = \mathbf{P}(A_\ell)$ for simplicity, so that

$$\varphi_\ell(\mathbf{1}_{A_\ell}) = \frac{\mathbf{1}_{A_\ell} - \mathbf{P}(A_\ell)}{\sqrt{\mathbf{V}(\mathbf{1}_{A_\ell})}},$$

and in particular

$$\mathbf{E}(\varphi_\ell(\mathbf{1}_{A_\ell})) = \langle \varphi_\ell, 1\rangle = 0, \qquad \mathbf{E}(\varphi_\ell(\mathbf{1}_{A_\ell})^2) = \|\varphi_\ell\|^2 = 1.$$

Hence, for ℓ, $\ell' \in \Lambda$, $W(\varphi_\ell, \varphi_{\ell'})$ is given by

$$W(\varphi_\ell,\varphi_{\ell'}) = \mathbf{E}(\varphi_\ell(\mathbf{1}_{A_\ell})\varphi_{\ell'}(\mathbf{1}_{A_{\ell'}}))$$

and it is (by definition) the correlation coefficient of the random variables $\mathbf{1}_{A_\ell}$ and $\mathbf{1}_{A_{\ell'}}$, explicitly

$$W(\varphi_\ell,\varphi_{\ell'}) = \begin{cases} 1 & \text{if } \ell=\ell', \\ \dfrac{\mathbf{P}(A_\ell\cap A_{\ell'}) - \mathbf{P}(A_\ell)\mathbf{P}(A_{\ell'})}{\sqrt{p_\ell(1-p_\ell)p_{\ell'}(1-p_{\ell'})}} & \text{otherwise.}\end{cases}$$

If (and only if) the (A_ℓ) form a family of pairwise independent events, we see that $W(\varphi_\ell,\varphi_{\ell'}) = \delta(\ell,\ell')$. More generally, in all cases, for any m, $n\subset\Lambda$, we have

$$W(\varphi_m,\varphi_n) = \mathbf{E}\Bigl(\prod_{\ell\in m}\varphi_\ell(\mathbf{1}_{A_\ell})\prod_{\ell\in n}\varphi_\ell(\mathbf{1}_{A_\ell})\Bigr)$$

which is a multiple normalized centered moment of the $\mathbf{1}_{A_\ell}$.

If the (A_ℓ) are globally independent, we obtain

$$\begin{aligned} W(\varphi_m,\varphi_n) &= \prod_{\substack{\ell\in m\cup n\\ \ell\notin m\cap n}} \frac{\mathbf{E}(\mathbf{1}_{A_\ell}-p_\ell)}{\sqrt{\mathbf{V}(\mathbf{1}_{A_\ell})}} \prod_{\ell\in m\cap n}\frac{\mathbf{E}((\mathbf{1}_{A_\ell}-p_\ell)^2)}{\sqrt{\mathbf{V}(\mathbf{1}_{A_\ell})}} \\ &= \delta(m,n)\end{aligned}$$

(since the third factor vanishes if the product is not empty, i.e., if $m\neq n$, and the third term is 1 by orthonormality of φ_ℓ). It follows by (2.7) that $\Delta \leqslant 1$, and in fact there must be equality. Moreover, in this situation, if $\mathcal{L}$ contains all subsets of Λ, we have

$$H = \prod_{\ell\in\Lambda}\Bigl(1+\frac{p_\ell}{1-p_\ell}\Bigr) = \prod_{\ell\in\Lambda}\frac{1}{1-p_\ell},$$

so that we find

$$\Delta H^{-1} \leqslant \prod_{\ell\in\Lambda}(1-\mathbf{P}(A_\ell)) = \mathbf{P}(A),$$

i.e., the large sieve inequality is an equality here.

Similarly, the inequality (3.1) becomes an equality if the events are pairwise independent, and reflects the formula for the variance of a sum of (pairwise) independent random variables.

In the general case of possibly dependent events, on the other hand, we have a quantitative inequality for $\mathbf{P}(A)$ which may be of some interest (and may be already known!). In fact, we have several possibilities depending on the choice of sieve support. It would be interesting to determine if those inequalities are of some use in probability theory.

To conclude this example, note that *any* sieve, once the prime sieve support $\mathcal{L}^*$ and the sieving sets (Ω_ℓ) are chosen, may be considered as a similar "binary" sieve with $Y_\ell = \{0,1\}$ for all ℓ, by replacing the sieve setting $(Y,\Lambda,(\rho_\ell))$ with $(Y,\mathcal{L}^*,\mathbf{1}_{\Omega_\ell})$.

Example 5.7. There are a few examples of the use of simple sieve methods in combinatorics. An example is a paper of Liu and Murty [LM] (mentioned to us by A. Granville), which explores (with some interesting combinatorial applications) a simple form of the dual sieve. Their sieve setting amounts to taking $\Psi = (A, B, \mathbf{1}_b)$ where A and B are finite sets, and for each $b \in B$, we have a map $\mathbf{1}_b : A \to \{0, 1\}$ (in loc. cit., the authors see (A, B) as a bipartite graph, and $\mathbf{1}_b(a) = 1$ if and only if there is an edge from a to b); the siftable set is A with identity map and counting measure, and the density is determined by $\nu_b(1) = |\mathbf{1}_b^{-1}(1)|/|A|$. In other words, this is also a special case of the previous example, and Theorem 1 and Corollary 1 of loc. cit. can also be trivially deduced from this (though they are simple enough to be better considered separately).

6. Coset sieves

Our next subject is a generalization of group sieves, which is the setting in which the Frobenius sieve over finite fields of [Ko1] and Section 11 operates.

As in Section 4, we start with a group G and a family of surjective homomorphisms $G \to G_\ell$, for $\ell \in \Lambda$, onto finite groups. However, we also assume that there is a normal subgroup G^g of G such that the quotient G/G^g is abelian, and hence we obtain a commutative diagram with exact rows

$$\begin{array}{ccccccccc} 1 & \longrightarrow & G^g & \longrightarrow & G & \xrightarrow{d} & G/G^g & \longrightarrow & 1 \\ & & \big\downarrow & & \big\downarrow{\scriptstyle \rho_\ell} & & \big\downarrow{\scriptstyle p} & & \\ 1 & \longrightarrow & G^g_\ell & \longrightarrow & G_\ell & \xrightarrow{d} & \Gamma_\ell & \longrightarrow & 1, \end{array} \tag{6.1}$$

where the downward arrows are surjective and the quotient groups Γ_ℓ thus defined are finite abelian groups.

After extending the definition of G_m to elements of $S(\Lambda)$ by multiplicativity, we can also define

$$G^g_m = \prod_{\ell | m} G^g_\ell$$

and we still can write commutative diagrams with exact rows

$$\begin{array}{ccccccccc} 1 & \longrightarrow & G^g & \longrightarrow & G & \xrightarrow{d} & G/G^g & \longrightarrow & 1 \\ & & \big\downarrow & & \big\downarrow{\scriptstyle \rho_m} & & \big\downarrow{\scriptstyle p} & & \\ 1 & \longrightarrow & G^g_m & \longrightarrow & G_m & \xrightarrow{d} & \Gamma_m & \longrightarrow & 1, \end{array} \tag{6.2}$$

(but the downward arrows are no longer necessarily surjective).

The sieve setting for a coset sieve is then $(Y, \Lambda, (\rho_\ell))$ where Y is the set of G-conjugacy classes in $d^{-1}(\alpha)$ for some fixed $\alpha \in G/G^g$. Since G^g is normal in G, this set is indeed invariant under conjugation by the whole of G (this is an important point). We let ρ_ℓ be the induced map

$$Y \to Y_\ell = \{y^\sharp \in G^\sharp_\ell \mid d(y^\sharp) = p(\alpha)\} \subset G^\sharp_\ell.$$

The natural density to consider (which arises in the sieve for Frobenius) is still

$$\nu_\ell(y^\sharp) = \frac{|y^\sharp|}{|G^g_\ell|}, \qquad \text{and hence} \qquad \nu_m(y^\sharp) = \frac{|y^\sharp|}{|G^g_m|}$$

for a conjugacy class $y^\sharp$. Note that this means that for any conjugacy-invariant subset $\Omega_\ell \subset G_\ell$, union of a set $\Omega^\sharp_\ell$ of conjugacy classes such that $\Omega^\sharp_\ell \subset d^{-1}(p(\alpha)) = Y_\ell$, we have

$$\nu(\Omega^\sharp_\ell) = \frac{|\Omega_\ell|}{|G^g_\ell|}.$$

We turn to the question of finding a suitable orthonormal basis of $L^2(Y_\ell, \nu_\ell)$. This is provided by the following general lemma, which applies equally to $H = G_\ell$ and to $H = G_m$ for $m \in S(\Lambda)$.

Lemma 6.1. *Let H be a finite group, H^g a subgroup with abelian quotient $\Gamma = H/H^g$. Let $\alpha \in \Gamma$ and Y the set of conjugacy classes of G with image α in Γ.*

For an irreducible linear representation π of H, let φ_π be the function

$$\varphi_\pi \,:\, y^\sharp \mapsto \operatorname{Tr} \pi(y^\sharp)$$

on $H^\sharp$.

(1) *For π, τ irreducible linear representations of H, we have*

$$\langle \varphi_\pi, \varphi_\tau \rangle = \begin{cases} 0, & \text{if either } \varphi_\pi \mid H^g \neq \varphi_\tau \mid H^g \text{ or } \varphi_\pi \mid Y = 0, \\ \overline{\psi(\alpha)} |\hat{\Gamma}^\pi|, & \text{where } \psi \in \hat{\Gamma} \text{ satisfies } \pi \otimes \psi \simeq \tau, \text{ otherwise,} \end{cases} \tag{6.3}$$

where $\hat{\Gamma}$ is the group of characters of Γ, $\hat{\Gamma}^\pi = \{\psi \in \hat{\Gamma} \mid \pi \simeq \pi \otimes \psi\}$, and the inner product is

$$\langle f, g \rangle = \frac{1}{|H^g|} \sum_{y^\sharp \in Y} |y^\sharp| f(y^\sharp) \overline{g(y^\sharp)}.$$

(2) *Let $\mathcal{B}$ be the family of functions*

$$y^\sharp \mapsto \frac{1}{\sqrt{|\hat{\Gamma}^\pi|}} \varphi_\pi(y^\sharp),$$

restricted to Y, where π ranges over the subset of a set of representatives for the equivalence relation

$$\pi \sim \tau \text{ if and only if } \pi \mid H^g \simeq \tau \mid H^g,$$

consisting of those representatives such that $\varphi_\pi \mid Y \neq 0$. Then $\mathcal{B}$ is an orthonormal basis of $L^2(Y)$ for the above inner product.

In the second case of (6.3), the existence of the character ψ will follow from the proof below.

Proof. We have

$$\begin{aligned} \langle \varphi_\pi, \varphi_\tau \rangle &= \frac{1}{|H^g|} \sum_{\substack{y \in H \\ d(y) = \alpha}} \operatorname{Tr} \pi(y) \overline{\operatorname{Tr} \tau(y)} \\ &= \frac{1}{|H^g|} \frac{1}{|\Gamma|} \sum_{y \in H} \Big(\sum_{\psi \in \hat{\Gamma}} \overline{\psi(\alpha)} \psi(y) \Big) \operatorname{Tr} \pi(y) \overline{\operatorname{Tr} \tau(y)} \\ &= \sum_{\psi \in \hat{\Gamma}} \overline{\psi(\alpha)} \langle \pi \otimes \psi, \tau \rangle_H = \sum_{\psi \in \hat{\Gamma}} \overline{\psi(\alpha)} \delta(\pi \otimes \psi, \tau), \end{aligned}$$

by orthogonality of characters of irreducible representations in $L^2(H)$.

First of all, this is certainly zero unless there exists at least one ψ such that $\pi \otimes \psi \simeq \tau$. In such a case we have $\pi \mid H^g \simeq \tau \mid H^g$ since $H^g \subset \ker(\psi)$, so we have shown that the condition $\pi \mid H^g \not\simeq \tau \mid H^g$ implies that the inner product is zero.

Assume now that $\pi \mid H^g \simeq \tau \mid H^g$; then repeating the above with $\alpha = 1$ (i.e., $Y = H^g$), it follows from $\langle \pi, \tau \rangle_{H^g} \neq 0$ that there exists one ψ at least such that $\pi \otimes \psi = \tau$.

Fixing one such character ψ_0, the characters ψ' for which $\pi \otimes \psi' \simeq \tau$ are given by $\psi' = \psi\psi_0$ where $\psi \in \hat{\Gamma}^\pi$. Then we find

$$\langle \varphi_\pi, \varphi_\tau \rangle = \sum_{\psi \in \hat{\Gamma}} \overline{\psi(\alpha)} \delta(\pi \otimes \psi, \pi \otimes \psi_0) = \overline{\psi_0(\alpha)} \sum_{\psi \in \hat{\Gamma}^\pi} \overline{\psi(\alpha)}.$$

For any $\psi \in \hat{\Gamma}^\pi$ and $y^\sharp \in Y$, we have the character relation

$$\operatorname{Tr} \pi(y^\sharp) = \psi(y^\sharp) \operatorname{Tr} \pi(y^\sharp) = \psi(\alpha) \operatorname{Tr} \pi(y^\sharp),$$

hence either $\psi(\alpha) = 1$ for all ψ, or $\operatorname{Tr} \pi(y^\sharp) = 0$ for all $y^\sharp$, i.e., φ_π restricted to Y vanishes. In this last case, we have trivially $\varphi_\tau = 0$ also on Y, and the inner product vanishes.

So we are led to the last case where $\pi \mid H^g = \tau \mid H^g$ but $\psi(\alpha) = 1$ for all $\psi \in \hat{\Gamma}^\pi$. Then the inner product formula is clear from the above.

Now to prove (2) from (1), notice first that the family $\mathcal{B}$ is a generating set of $L^2(Y)$ (indeed, all φ_π generate $L^2(H^\sharp)$, but those π for which $\varphi_\pi = 0$ on Y are clearly not needed, and if $\pi \sim \tau$, we have $\varphi_\tau = \psi(\alpha)\varphi_\pi$ on Y, where ψ satisfies $\tau \simeq \psi \otimes \pi$, so one element of each equivalence class suffices for functions on Y). Then the fact that we have an orthonormal basis follows from the inner product formula, observing that if $\tau \simeq \pi \otimes \psi$, we have in fact $\pi = \tau$ by definition of the equivalence relation, so $\psi = 1$ in (6.3). □

Example 6.2. In this lemma we emphasize that distinct representations of H may give the same restriction on H^g, in which case they correspond to a single element of the basis, and that it is possible that a φ_π vanish on Y, in which case their representative is discarded from the basis.

Take for instance $G = D_n$, a dihedral group of order $2n$. There is an exact sequence

$$1 \to \mathbf{Z}/n\mathbf{Z} \to G \xrightarrow{d} \mathbf{Z}/2\mathbf{Z} \to 1$$

and if $Y = d^{-1}(1) \subset G$ and π is any representation of G of degree 2, we have $\operatorname{Tr}\pi(x) = 0$ for all $x \in Y$ (see e.g. [S2, 5.3]).

In particular, note that even though both cosets of $\mathbf{Z}/n\mathbf{Z}$ in G have four elements, the sets of conjugacy classes in each do not have the same cardinality (there are 5 conjugacy classes, 3 in $\ker d$ and 2 in the other coset). In other words, in a coset sieve, the spaces Y_m usually depend on the value of α (they are usually not even isomorphic).

If we apply Lemma 6.1 to the groups G_m and their subgroups G_m^g, we clearly obtain orthonormal bases of $L^2(Y_m)$ containing the constant function 1, for the density ν_m above. Although it was not phrased in this manner, this is what was used in [Ko1] (with minor differences, e.g., the upper bound κ for the order of $\hat{\Gamma}_m^\pi$ that occurs in loc. cit., and can be removed – as also noticed independently by Zywina in a private email).

As before we summarize the sieve statement:

Proposition 6.3. *Let G be a group, G^g a normal subgroup with abelian quotient, $\rho_\ell : G \to G_\ell$ a family of surjective homomorphisms onto finite groups. Let $(Y, \Lambda, (\rho_\ell))$ be the coset sieve setting associated with some $\alpha \in G/G^g$, and (X, μ, F) an associated siftable set.*

For $m \in S(\Lambda)$, let Π_m be a set of representatives of the set of irreducible representations of G_m modulo equality restricted to G_m^g, containing the constant function 1. Moreover, let Π_m^* be the subset of primitive representations, i.e., those such that when π is decomposed as $\boxtimes_{\ell \mid m}\pi_\ell$, no component π_ℓ is trivial, and $\operatorname{Tr}\pi_\ell$ is not identically zero on Y_ℓ.

Then, for any prime sieve support $\mathcal{L}^*$ and associated sieve support $\mathcal{L}$, i.e., such that (2.1) *holds, and for any conjugacy invariant sifting sets (Ω_ℓ) with $\Omega_\ell \subset Y_\ell$ for $\ell \in \mathcal{L}^*$, we have*

$$|S(X, \Omega; \mathcal{L}^*)| \leqslant \Delta H^{-1}$$

where Δ is the smallest non-negative real number such that

$$\sum_{m \in \mathcal{L}} \sum_{\pi \in \Pi_m^*} \Bigl| \int_X \alpha(x) \operatorname{Tr}\pi(\rho_m(F_x)) d\mu(x) \Bigr|^2 \leqslant \Delta \int_X |\alpha(x)|^2 d\mu(x)$$

for all square-integrable function $\alpha \in L^2(X, \mu)$, and where

$$H = \sum_{m \in \mathcal{L}} \prod_{\ell \mid m} \frac{|\Omega_\ell|}{|G_\ell^g| - |\Omega_\ell|}.$$

Moreover we have

(6.4) $$\Delta \leqslant \max_{m \in \mathcal{L}} \max_{\pi \in \Pi_m^*} \sum_{n \in \mathcal{L}} \sum_{\tau \in \Pi_n^*} |W(\pi, \tau)|,$$

where

$$W(\pi,\tau)=\frac{1}{\sqrt{|\hat{\Gamma}^{\pi}_{m}||\hat{\Gamma}^{\tau}_{n}|}}\int_X \operatorname{Tr}\pi(\rho_m(F_x))\overline{\operatorname{Tr}\tau(\rho_n(F_x))}d\mu(x).$$

We now consider what happens of the equidistribution approach in this context. (Some of this also applies to group conjugacy sieves, where $G^g = G$).

If we apply Lemma 2.11 to the elements φ_π, φ_τ of the basis $\mathcal{B}_m$ and $\mathcal{B}_n$ of $L^2(Y_m)$, we see that the function $[\varphi_\pi,\overline{\varphi_\tau}]$ defined in (2.8) is the character of the representation

$$[\pi,\bar{\tau}]=\pi_{m'}\boxtimes(\pi_d\otimes\overline{\tau_d})\boxtimes\overline{\tau_{n'}}$$

of $G_{[m,n]}$, already defined in (4.3). Hence we have

$$W(\pi,\tau)=\frac{1}{\sqrt{|\hat{\Gamma}^{\pi}_{m}||\hat{\Gamma}^{\tau}_{n}|}}\int_X \operatorname{Tr}([\pi,\bar{\tau}]\rho_{[m,n]}(F_x))d\mu(x). \tag{6.5}$$

In applications, this means that to estimate the integrals $W(\pi,\tau)$ it suffices (and may be more convenient) to be able to deal with integrals of the form

$$\int_X \operatorname{Tr}(\varpi(F_x))d\mu(x)$$

where ϖ is a representation of G that factors through a finite product of groups G_ℓ (see Section 9 for an instance of this).

If we try to approach those integrals using the equidistribution method, then the analogue of (2.9) is the identity

$$|\{\rho_d(F_x)=y^\sharp\}|=\int_{\{\rho_d(F_x)=y^\sharp\}}d\mu(x)=\frac{|y^\sharp|}{|G^g_d|}|X|+r_d(X;y^\sharp), \tag{6.6}$$

defining $r_d(X;y^\sharp)$ for $y^\sharp\in Y_d$. Then (2.10) becomes

$$W(\pi,\tau)=\frac{|X|}{\sqrt{|\hat{\Gamma}^{\pi}_{m}||\hat{\Gamma}^{\tau}_{n}|}}m([\pi,\bar{\tau}])+O\Bigl(\frac{1}{\sqrt{|\hat{\Gamma}^{\pi}_{m}||\hat{\Gamma}^{\tau}_{n}|}}\sum_{y^\sharp\in Y_{[m,n]}}\dim[\pi,\bar{\tau}]|r_{[m,n]}(X;y^\sharp)|\Bigr)$$

where, comparing with Lemma 6.1 with $H=G_{[m,n]}$, we have

$$m([\pi,\bar{\tau}])=\langle\varphi_\pi,\varphi_\tau\rangle \tag{6.7}$$

where the inner product is in $L^2(Y_{[m,n]})$ and both π and τ are extended to (irreducible) representations of $G_{[m,n]}$ by taking trivial components at those $\ell\in[m,n]$ not in m or n respectively. Hence by (6.3), we have $m([\pi,\bar{\tau}])=0$ unless π and τ thus extended are isomorphic restricted to $G^g_{[m,n]}$, which clearly can occur only if $m=n$ and then if $\pi=\tau$ by orthogonality of $\mathcal{B}^*_n$. When $(m,\pi)=(n,\tau)$, the inner product is equal to $|\hat{\Gamma}^{\pi}_{m}|$ by (6.3).

Using this and (2.10), we get

$$W(\pi,\tau)=\delta(\pi,\tau)|X|+O\Bigl(\sum_{y^\sharp\in Y_{[m,n]}}\dim[\pi,\bar{\tau}]|r_{[m,n]}(X;y^\sharp)|\Bigr),$$

where the implied constant is $\leqslant 1$. Hence for any sieve support $\mathcal{L}$, the large sieve bound of Proposition 2.10 holds with

$$\Delta\leqslant|X|+R(X,\mathcal{L}) \tag{6.8}$$

where

$$R(X;\mathcal{L})=\max_{n\in\mathcal{L}}\max_{\pi\in\Pi^*_n}\Bigl\{\sum_{m\in\mathcal{L}}\sum_{\tau\in\Pi^*_m}\sum_{y^\sharp\in Y_{[m,n]}}\dim[\pi,\bar{\tau}]|r_{[m,n]}(X;y^\sharp)|\Bigr\}. \tag{6.9}$$

For later reference, we also note the following fact:

Lemma 6.4. *Let m, n in $S(\Lambda)$, $\pi \in \Pi_m^*$, $\tau \in \Pi_n^*$. The multiplicity of the trivial representation in the restriction of $[\pi, \bar{\tau}]$ to $G^g_{[m,n]}$ is equal to zero if $(m,\pi) \neq (n,\tau)$, and is equal to $|\hat{\Gamma}^\pi_m|$ if $(m,\pi) = (n,\tau)$.*

Proof. This multiplicity is by definition $\langle [\pi, \bar{\tau}], 1\rangle$ computed in $L^2(G^g_{[m,n]})$, i.e., it is $\langle \varphi_\pi, \varphi_\tau\rangle$ in $L^2(Y_{[m,n]})$ in the case $\alpha = 1 \in G/G^g$ (with the same convention on extending π and τ to $G_{[m,n]}$ as before). So the result is a consequence of Lemma 6.1. □

7. Degrees and sums of degrees of representations of finite groups

This section is essentially independent from the rest of the paper, and is devoted to proving some inequalities which are likely to be useful in estimating quantities such as (6.4) or $R(X, \mathcal{L})$ in (6.9). Indeed, we will use them later on in Section 9 and Section 11.

In practice, the bound for the individual exponential sums $W(\pi, \tau)$ is likely to involve the order of the groups G and the degrees of its representations, and their combination in (6.4) will involve sums of the degrees. For instance, in the next sections, we will need to bound

$$\max_{m,\pi} \Big\{ (\dim \pi) \sum_n |G_{[m,n]}| \sum_{\tau \in \Pi_n^*} (\dim \tau) \Big\},$$
$$\max_{m,\pi} \Big\{ (\dim \pi) \sum_n \sum_{\tau \in \Pi_n^*} (\dim \tau) \Big\}.$$

In applications, the groups G_ℓ are often (essentially) classical linear groups over $\mathbf{F}_\ell$, but they are not entirely known (it may only be known that they have bounded index in $GL(n, \mathbf{F}_\ell)$ as ℓ varies, for instance¸ see [Ko1] and Section 11). Our results are biased to this case.

For a finite group G and $p \in [1, +\infty]$, we denote

$$A_p(G) = \Big(\sum_\rho \dim(\rho)^p\Big)^{1/p}, \qquad \text{if } p \neq +\infty, \qquad A_\infty(G) = \max\{\dim(\rho)\}$$

where ρ runs over irreducible linear representations of G (in characteristic zero). For example, we have $A_2(G) = \sqrt{|G|}$ for all G and if G is abelian, then $A_p(G) = |G|^{1/p}$ for all p. Moreover

$$\lim_{p \to +\infty} A_p(G) = A_\infty(G).$$

We are primarily interested in $A_1(G)$ and $A_\infty(G)$, but $A_{5/2}(G)$ will also occur in the proof of Theorem 1.3, and other cases may turn out to be useful in other sieve settings. We start with an easy monotonicity lemma.

Lemma 7.1. *Let G be a finite group and $H \subset G$ a subgroup, $p \in [1, +\infty]$. We have*

$$A_p(H) \leqslant A_p(G).$$

Proof. For any irreducible representation ρ of H, choose (arbitrarily) an irreducible representation $\pi(\rho)$ of G that occurs with positive multiplicity in the induced representation $\mathrm{Ind}_H^G \rho$.

Let π be a representation of G in the image of $\rho \mapsto \pi(\rho)$. For any ρ where $\pi(\rho) = \pi$, we have

$$\langle \rho, \mathrm{Res}_H^G \pi\rangle_H = \langle \mathrm{Ind}_H^G \rho, \pi\rangle_G > 0,$$

by Frobenius reciprocity, i.e., all ρ with $\pi(\rho) = \pi$ occur in the restriction of π to H. Hence for $p \neq +\infty$ we obtain

$$\sum_{\substack{\rho \\ \pi(\rho) = \pi}} \dim(\rho)^p \leqslant \Big(\sum_{\substack{\rho \\ \pi(\rho) = \pi}} \dim(\rho)\Big)^p \leqslant \dim(\pi)^p,$$

and summing over all possible $\pi(\rho)$ gives the inequality

$$A_p(H)^p \leqslant A_p(G)^p$$

by positivity. This settles the case $p \neq +\infty$, and the other case only requires noticing that $\dim(\rho) \leqslant \dim(\pi(\rho)) \leqslant A_\infty(G)$. □

We come to the main result of this section. The terminology, which may not be familiar to all readers, is explained by examples after the proof. We hope that there will be no confusion between p and the characteristic of the finite field $\mathbf{F}_q$ which occurs...

Proposition 7.2. *Let $\mathbf{G}/\mathbf{F}_q$ be a split connected reductive linear algebraic group of dimension d and rank r over a finite field, with connected center. Let W be its Weyl group and $G = \mathbf{G}(\mathbf{F}_q)$ the finite group of rational points of $\mathbf{G}$.*

(1) *For any subgroup $H \subset G$ and $p \in [1, +\infty]$, we have*

$$A_p(H) \leqslant (q+1)^{(d-r)/2+r/p}\Bigl(1+\frac{2r|W|}{q-1}\Bigr)^{1/p},$$

with the convention $r/p = 0$ if $p = +\infty$, in particular the second factor is $= 1$ for $p = +\infty$.

(2) *If $\mathbf{G}$ is a product of groups of type A or C, i.e., of linear and symplectic groups, then*

$$A_p(H) \leqslant (q+1)^{(d-r)/2+r/p}.$$

The proof is based on a simple interpolation argument from the extreme cases $p = 1$, $p = +\infty$. Indeed by Lemma 7.1 we can clearly assume $H = G$ and by writing the obvious inequality

$$A_p(G)^p = \sum_{\rho} \dim(\rho)^p \leqslant A_\infty(G)^{p-1} A_1(G),$$

we see that it suffices to prove the following:

Proposition 7.3. *Let $\mathbf{G}/\mathbf{F}_q$ be a split connected reductive linear algebraic group of dimension d with connected center, and let $G = \mathbf{G}(\mathbf{F}_q)$ be the finite group of its rational points. Let r be the rank of $\mathbf{G}$. Then we have*

$$A_\infty(G) \leqslant \frac{|G|_{p'}}{(q-1)^r} \leqslant (q+1)^{(d-r)/2}, \quad \text{and} \quad A_1(G) \leqslant (q+1)^{(d+r)/2}\Bigl(1+\frac{2r|W|}{q-1}\Bigr), \tag{7.1}$$

where $n_{p'}$ denotes the prime-to-p part of a rational number n, p being the characteristic of $\mathbf{F}_q$. Moreover, if the principal series of G is not empty[6]*, there is equality*

$$A_\infty(G) = \frac{|G|_{p'}}{(q-1)^r}$$

and $\dim \rho = A_\infty(G)$ if and only if ρ is in the principal series.

Finally if $\mathbf{G}$ is a product of groups of type A or C, then the factor $(1 + 2r|W|/(q-1))$ may be removed in the bound for $A_1(G)$.

It seems very possible that the factor $(1 + 2r|W|/(q - 1))$ could always be removed, but we haven't been able to figure this out using Deligne-Lusztig characters, and in fact for groups of type A or C, we simply quote *exact formulas* for $A_1(G)$ due to Gow, Klyachko and Vinroot, which are proved in completely different ways.[7] The extra factor is not likely to be a problem in many applications where $q \to +\infty$, but it may be questionable for uniformity with respect to the rank.

The ideas in the proof were suggested and explained by J. Michel.

Proof. This is based on properties of the Deligne-Lusztig generalized characters. We will mostly refer to [DM] and [Ca] for all facts which are needed (using notation from [DM], except for writing simply G for what is denoted $\mathbf{G}^F$ there). We identify irreducible representations of G (up to isomorphism) with their characters seen as complex-valued functions on G.

First, for a connected reductive group $\mathbf{G}/\mathbf{F}_q$ over a finite field, Deligne and Lusztig have constructed (see e.g. [DM, 11.14]) a family $R_{\mathbf{T}}^{\mathbf{G}}(\theta)$ of generalized representations of $G = \mathbf{G}(\mathbf{F}_q)$ (i.e., linear combinations with integer coefficients of "genuine" representations of G), parametrized by pairs $(\mathbf{T}, \theta)$ consisting of a maximal torus $\mathbf{T} \subset \mathbf{G}$ defined over $\mathbf{F}_q$ and a (one-dimensional)

[6] In particular if q is large enough given $\mathbf{G}$.

[7] The "right" upper bound for the case of groups of type A (i.e, $GL(r)$) may be recovered using the structure of unipotent representations of such groups.

character θ of the finite abelian group $T = \mathbf{T}(\mathbf{F}_q)$. The $R_{\mathbf{T}}^{\mathbf{G}}(\theta)$ are not all irreducible, but any irreducible character occurs (with positive or negative multiplicity) in the decomposition of at least one such character. Moreover, $R_{\mathbf{T}}^{\mathbf{G}}(\theta)$ only depends (up to isomorphism) on the G-conjugacy class of the pair $(\mathbf{T}, \theta)$.

We quote here a useful classical fact: for any $\mathbf{T}$ we have

$$(q-1)^r \leqslant |T| \leqslant (q+1)^r \tag{7.2}$$

(see e.g. [DM, 13.7 (ii)]), and moreover $|T| = (q-1)^r$ if and only if $\mathbf{T}$ is a split torus (i.e., $\mathbf{T} \simeq \mathbf{G}_m^r$ over $\mathbf{F}_q$). Indeed, we have

$$|T| = |\det(q^n - w \mid Y_0)|$$

where $w \in W$ is such that $\mathbf{T}$ is obtained from a split torus $\mathbf{T}_0$ by "twisting with w" (see e.g. [Ca, Prop. 3.3.5]), and $Y_0 \simeq \mathbf{Z}^r$ is the group of cocharacters of $\mathbf{T}_0$. If $\lambda_1, \ldots, \lambda_r$ are the eigenvalues of w acting on Y_0, which are roots of unity, then we have

$$|T| = \prod_{i=1}^{r} (q - \lambda_i),$$

and so $|T| = (q-1)^r$ if and only if each λ_i is equal to 1, if and only if w acts trivially on Y_0, if and only if $w = 1$ (W acts faithfully on Y_0) and $\mathbf{T}$ is split.

As in [DM, 12.12], we denote by $\rho \mapsto p(\rho)$ the orthogonal projection of the space $\mathcal{C}(G)$ of complex-valued conjugacy-invariant functions on G to the subspace generated by Deligne-Lusztig characters, where $\mathcal{C}(G)$ is given the standard inner product

$$\langle f, g \rangle = \frac{1}{|G|} \sum_{x \in G} f(x)\overline{g(x)},$$

and for a representation ρ, we of course denote $p(\rho) = p(\operatorname{Tr} \rho)$ the projection of its character.

For any representation ρ, we have $\dim(\rho) = \dim(p(\rho))$, where $\dim(f)$, for an arbitrary function $f \in \mathcal{C}(G)$ is obtained by linearity from the degree of characters. Indeed, for any f standard character theory shows that

$$\dim(f) = \langle f, \mathrm{reg}_G \rangle$$

where reg_G is the regular representation of G. From [DM, 12.14], the regular representation is in the subspace spanned by the Deligne-Lusztig characters, so by definition of an orthogonal projector we have

$$\dim(\rho) = \langle \rho, \mathrm{reg}_G \rangle = \langle p(\rho), \mathrm{reg}_G \rangle = \dim(p(\rho)).$$

Now because the characters $R_{\mathbf{T}}^{\mathbf{G}}(\theta)$ for distinct conjugacy classes of $(\mathbf{T}, \theta)$ are orthogonal (see e.g. [DM, 11.15]), we can write

$$p(\rho) = \sum_{(\mathbf{T},\theta)} \frac{\langle \rho, R_{\mathbf{T}}^{\mathbf{G}}(\theta) \rangle}{\langle R_{\mathbf{T}}^{\mathbf{G}}(\theta), R_{\mathbf{T}}^{\mathbf{G}}(\theta) \rangle} R_{\mathbf{T}}^{\mathbf{G}}(\theta)$$

(sum over all distinct Deligne-Lusztig characters) and so

$$\dim(p(\rho)) = \sum_{(\mathbf{T},\theta)} \frac{\langle \rho, R_{\mathbf{T}}^{\mathbf{G}}(\theta) \rangle}{\langle R_{\mathbf{T}}^{\mathbf{G}}(\theta), R_{\mathbf{T}}^{\mathbf{G}}(\theta) \rangle} \dim(R_{\mathbf{T}}^{\mathbf{G}}(\theta)).$$

By [DM, 12.9] we have

$$\dim(R_{\mathbf{T}}^{\mathbf{G}}(\theta)) = \varepsilon_{\mathbf{G}} \varepsilon_{\mathbf{T}} |G|_{p'} |T|^{-1}, \tag{7.3}$$

where $\varepsilon_{\mathbf{G}} = (-1)^r$ and $\varepsilon_{\mathbf{T}} = (-1)^{r(\mathbf{T})}$, $r(\mathbf{T})$ being the $\mathbf{F}_q$-rank of $\mathbf{T}$ (see [DM, p. 66] for the definition). This yields the formula

$$\dim(p(\rho)) = |G|_{p'} \sum_{(\mathbf{T},\theta)} \frac{1}{|T|} \frac{\langle \rho, \varepsilon_{\mathbf{G}} \varepsilon_{\mathbf{T}} R_{\mathbf{T}}^{\mathbf{G}}(\theta) \rangle}{\langle R_{\mathbf{T}}^{\mathbf{G}}(\theta), R_{\mathbf{T}}^{\mathbf{G}}(\theta) \rangle}. \tag{7.4}$$

Now we use the fact that pairs $(\mathbf{T},\theta)$ are partitioned in *geometric conjugacy classes*, defined as follows: two pairs $(\mathbf{T},\theta)$ and $(\mathbf{T}',\theta')$ are geometrically conjugate if and only if there exists $g\in\mathbf{G}(\bar{\mathbf{F}}_q)$ such that $\mathbf{T}=g\mathbf{T}'g^{-1}$ and for all n such that $g\in\mathbf{G}(\mathbf{F}_{q^n})$, we have

$$\theta(N_{\mathbf{F}_{q^n}/\mathbf{F}_q}(x))=\theta'(N_{\mathbf{F}_{q^n}/\mathbf{F}_q}(g^{-1}xg))\qquad\text{for } x\in\mathbf{T}(\mathbf{F}_{q^n}),$$

(see e.g. [DM, 13.2]). The point is the following property of geometric conjugacy classes: if the generalized characters $R^{\mathbf{G}}_{\mathbf{T}}(\theta)$ and $R^{\mathbf{G}}_{\mathbf{T}'}(\theta')$ have a common irreducible component, then $(\mathbf{T},\theta)$ and $(\mathbf{T}',\theta')$ are geometrically conjugate (see e.g. [DM, 13.2]).

In particular, for a given ρ, if $\langle\rho,R^{\mathbf{G}}_{\mathbf{T}}(\theta)\rangle$ is non-zero for some $(\mathbf{T},\theta)$, then only pairs $(\mathbf{T}',\theta')$ geometrically conjugate to $(\mathbf{T},\theta)$ may satisfy $\langle\rho,R^{\mathbf{G}}_{\mathbf{T}'}(\theta)\rangle\neq 0$. So we have

$$\dim(p(\rho))=|G|_{p'}\sum_{(\mathbf{T},\theta)\in\kappa}\frac{1}{|T|}\frac{\langle\rho,\varepsilon_{\mathbf{G}}\varepsilon_{\mathbf{T}}R^{\mathbf{G}}_{\mathbf{T}}(\theta)\rangle}{\langle R^{\mathbf{G}}_{\mathbf{T}}(\theta),R^{\mathbf{G}}_{\mathbf{T}}(\theta)\rangle},$$

for some geometric conjugacy class κ, depending on ρ. By Cauchy-Schwarz, we obtain

$$\dim(p(\rho))\leqslant|G|_{p'}\Bigl(\sum_{(\mathbf{T},\theta)\in\kappa}\frac{1}{|T|^2}\frac{1}{\langle R^{\mathbf{G}}_{\mathbf{T}}(\theta),R^{\mathbf{G}}_{\mathbf{T}}(\theta)\rangle}\Bigr)^{1/2}\Bigl(\sum_{(\mathbf{T},\theta)\in\kappa}\frac{|\langle\rho,R^{\mathbf{G}}_{\mathbf{T}}(\theta)\rangle|^2}{\langle R^{\mathbf{G}}_{\mathbf{T}}(\theta),R^{\mathbf{G}}_{\mathbf{T}}(\theta)\rangle}\Bigr)^{1/2}.\tag{7.5}$$

The second term on the right is simply $\langle p(\rho),p(\rho)\rangle\leqslant\langle\rho,\rho\rangle=1$. As for the first term we have

$$\sum_{(\mathbf{T},\theta)\in\kappa}\frac{1}{|T|^2}\frac{1}{\langle R^{\mathbf{G}}_{\mathbf{T}}(\theta),R^{\mathbf{G}}_{\mathbf{T}}(\theta)\rangle}\leqslant\frac{1}{(q-1)^{2r}}\sum_{(\mathbf{T},\theta)\in\kappa}\frac{1}{\langle R^{\mathbf{G}}_{\mathbf{T}}(\theta),R^{\mathbf{G}}_{\mathbf{T}}(\theta)\rangle}$$

by (7.2). Now it is known that for each class κ, the assumption that $\mathbf{G}$ has connected center implies that the generalized character

$$\chi(\kappa)=\sum_{(\mathbf{T},\theta)\in\kappa}\frac{\varepsilon_{\mathbf{G}}\varepsilon_{\mathbf{T}}R^{\mathbf{G}}_{\mathbf{T}}(\theta)}{\langle R^{\mathbf{G}}_{\mathbf{T}}(\theta),R^{\mathbf{G}}_{\mathbf{T}}(\theta)\rangle}$$

is in fact an irreducible character of G (such characters are called *regular* characters; see e.g. [Ca, Prop. 8.4.7]). This implies that

$$\sum_{(\mathbf{T},\theta)\in\kappa}\frac{1}{\langle R^{\mathbf{G}}_{\mathbf{T}}(\theta),R^{\mathbf{G}}_{\mathbf{T}}(\theta)\rangle}=\langle\chi(\kappa),\chi(\kappa)\rangle=1,$$

and so we have

$$\dim p(\rho)\leqslant\frac{|G|_{p'}}{(q-1)^r}.\tag{7.6}$$

Now observe that we will have equality in this argument if ρ is itself of the form $\pm R^{\mathbf{G}}_{\mathbf{T}}(\theta)$, and if $|T|=(q-1)^r$. Those conditions hold for representations of the principal series, i.e., characters $R^{\mathbf{G}}_{\mathbf{T}}(\theta)$ for an $\mathbf{F}_q$-split torus $\mathbf{T}$ and a character θ "in general position" (see e.g. [Ca, Cor. 7.3.5]). Such characters are also, more elementarily, induced characters $\mathrm{Ind}_B^G(\theta)$, where $B=\mathbf{B}(\mathbf{F}_q)$ is a Borel subgroup containing T, for some Borel subgroup $\mathbf{B}$ defined over $\mathbf{F}_q$ containing $\mathbf{T}$ (which exist for a split torus $\mathbf{T}$) and θ is extended to B by setting $\theta(u)=1$ for unipotent elements $u\in B$. For this, see e.g. [Lu, Prop. 2.6].

Conversely, let ρ be such that

$$\dim\rho=\frac{|G|_{p'}}{(q-1)^r}$$

and let κ be the associated geometric conjugacy class. From the above, for any $(\mathbf{T},\theta)$ in κ, we have $|T|=(q-1)^r$, i.e., $\mathbf{T}$ is $\mathbf{F}_q$-split. Now it follows from Lemma 7.4 (probably well-known) that this implies that $R^{\mathbf{G}}_{\mathbf{T}}(\theta)$ is an irreducible representation, so must be equal to ρ.

We now come to $A_1(G)$. To deal with the fact that in (7.4), $|T|$ depends on $(\mathbf{T},\theta)\in\kappa$, we write

$$\dim(p(\rho)) = \frac{|G|_{p'}}{(q-1)^r}\sum_{\kappa}\langle\rho,\chi(\kappa)\rangle + |G|_{p'}\sum_{(\mathbf{T},\theta)}\Big(\frac{1}{|T|}-\frac{1}{(q-1)^r}\Big)\frac{\varepsilon_{\mathbf{G}}\varepsilon_{\mathbf{T}}\langle\rho,R_{\mathbf{T}}^{\mathbf{G}}(\theta)\rangle}{\langle R_{\mathbf{T}}^{\mathbf{G}}(\theta),R_{\mathbf{T}}^{\mathbf{G}}(\theta)\rangle} \tag{7.7}$$

(since by (7.2), the dependency is rather weak).

Now summing over ρ, consider the first term's contribution. Since $\chi(\kappa)$ is an irreducible character, the sum

$$\sum_{\rho}\sum_{\kappa}\langle\rho,\chi(\kappa)\rangle$$

is simply the number of geometric conjugacy classes. This is given by $q^{r'}|Z|$ by [DM, 14.42] or [Ca, Th. 4.4.6 (ii)], where r' is the semisimple rank of $\mathbf{G}$ and $Z=Z(\mathbf{G})(\mathbf{F}_q)$ is the group of rational points of the center of $\mathbf{G}$. For this quantity, note that the center of $\mathbf{G}$ being connected implies that $Z(\mathbf{G})$ is the radical of $\mathbf{G}$ (see e.g. [Sp, Pr. 7.3.1]) so $Z(\mathbf{G})$ is a torus and $r=r'+\dim Z(\mathbf{G})$. So using again the bounds (7.2) for the cardinality of the group of rational points of a torus, we obtain

$$|Z|q^{r'}\leqslant (q+1)^r. \tag{7.8}$$

To estimate the sum of the contributions in the second term, say $\sum t(\rho)$, we write

$$\sum_{\rho}t(\rho)=|G|_{p'}\sum_{(\mathbf{T},\theta)}\Big(\frac{1}{|T|}-\frac{1}{(q-1)^r}\Big)\frac{\varepsilon_{\mathbf{G}}\varepsilon_{\mathbf{T}}\langle\sum_{\rho}\rho,R_{\mathbf{T}}^{\mathbf{G}}(\theta)\rangle}{\langle R_{\mathbf{T}}^{\mathbf{G}}(\theta),R_{\mathbf{T}}^{\mathbf{G}}(\theta)\rangle},$$

and we bound

$$\Big|\langle\sum_{\rho}\rho,R_{\mathbf{T}}^{\mathbf{G}}(\theta)\rangle\Big|\leqslant\langle R_{\mathbf{T}}^{\mathbf{G}}(\theta),R_{\mathbf{T}}^{\mathbf{G}}(\theta)\rangle \tag{7.9}$$

for any $(\mathbf{T},\theta)$, since we can write

$$R_{\mathbf{T}}^{\mathbf{G}}(\theta)=\sum_{\rho}a(\rho)\rho\quad\text{with } a(\rho)\in\mathbf{Z},$$

and therefore

$$\Big|\langle\sum_{\rho}\rho,R_{\mathbf{T}}^{\mathbf{G}}(\theta)\rangle\Big|=\Big|\sum_{\rho}a(\rho)\Big|\leqslant\sum_{\rho}|a(\rho)|^2=\langle R_{\mathbf{T}}^{\mathbf{G}}(\theta),R_{\mathbf{T}}^{\mathbf{G}}(\theta)\rangle. \tag{7.10}$$

Thus

$$\sum_{\rho}t(\rho)\leqslant\frac{|G|_{p'}}{(q-1)^r}\frac{2r}{q-1}|\{(\mathbf{T},\theta)\}|.$$

There are at most $|W|$ different choices of $\mathbf{T}$ up to G-conjugacy, and for each there are at most $|T|\leqslant(q+1)^r$ different characters, and so we have

$$\sum_{\rho}t(\rho)\leqslant\frac{|G|_{p'}}{(q-1)^r}\frac{2r|W|}{q-1}(q+1)^r, \tag{7.11}$$

and

$$\sum_{\rho}\dim\rho\leqslant(q+1)^r\frac{|G|_{p'}}{(q-1)^r}\Big(1+\frac{2r|W|}{q-1}\Big). \tag{7.12}$$

To conclude, we use the classical formula

$$|G|=q^N\prod_{1\leqslant i\leqslant r}(q^{d_i}-1),$$

where N is the number of positive roots of $\mathbf{G}$, and the d_i are the degrees of invariants of the Weyl group (this is because $\mathbf{G}$ is split; see e.g. [Ca, 2.4.1 (iv); 2.9, p. 75]). So

$$|G|_{p'} = \prod_{1 \leqslant i \leqslant r} (q^{d_i} - 1)$$

and

$$\frac{|G|_{p'}}{(q-1)^r} = \prod_{1\leqslant i\leqslant r} \frac{q^{d_i}-1}{q-1} \leqslant \prod_{1\leqslant i\leqslant r} (q+1)^{d_i-1} = (q+1)^{\sum (d_i-1)} = (q+1)^{(d-r)/2}, \tag{7.13}$$

since $\sum (d_i - 1) = N$ and $N = (d-r)/2$ (see e.g. [Ca, 2.4.1], [Sp, 8.1.3]).

Inserting this in (7.6) we derive the first inequality in (7.1), and with (7.12), we get

$$A_1(G) \leqslant (q+1)^{(d+r)/2}\Big(1 + \frac{2r|W|}{q-1}\Big),$$

which is the second part of (7.1).

Now we explain why the extra factor involving the Weyl group can be removed for products of groups of type A and C. Clearly it suffices to work with $\mathbf{G} = GL(n)$ and $\mathbf{G} = CSp(2g)$.

For $\mathbf{G} = GL(n)$, with $d = n^2$ and $r = n$, Gow [Go] and Klyachko [Kl] have proved independently that $A_1(G)$ is equal to the number of symmetric matrices in G. The bound

$$A_1(G) \leqslant (q+1)^{(n^2+n)/2}$$

follows immediately.

For $\mathbf{G} = CSp(2g)$, with $d = 2g^2 + g + 1$ and $r = g + 1$, the exact analog of Gow's theorem is due to Vinroot [V]. Again, Vinroot's result implies $A_1(G) \leqslant (q+1)^{(d+r)/2}$ in this case (see [V, Cor 6.1], and use the formulas for the order of unitary and linear groups to check the final bound). $\square$

Here is the lemma used in the determination of $A_\infty(G)$ when there is a character in general position of a split torus:

Lemma 7.4. *Let $\mathbf{G}/\mathbf{F}_q$ be a split connected reductive linear algebraic group of dimension d and let $G = \mathbf{G}(\mathbf{F}_q)$ be the finite group of its rational points. Let $\mathbf{T}$ be a split torus in $\mathbf{G}$, θ a character of $T = \mathbf{T}(\mathbf{F}_q)$. If $\mathbf{T}'$ is also a split torus for any pair $(\mathbf{T}', \theta')$ geometrically conjugate to $(\mathbf{T}, \theta)$, then $R_{\mathbf{T}}^{\mathbf{G}}(\theta)$ is irreducible.*

Proof. If $R_{\mathbf{T}}^{\mathbf{G}}(\theta)$ is not irreducible, then by the inner product formula for Deligne-Lusztig characters, there exists $w \in W$, $w \neq 1$, such that ${}^w\theta = \theta$ (see e.g. [DM, Cor. 11.15]). Let $\mathbf{T}'$ be a torus obtained from $\mathbf{T}$ by "twisting by w", i.e., $\mathbf{T}' = g\mathbf{T}g^{-1}$ where $g \in \mathbf{G}$ is such that $g^{-1}\operatorname{Fr}(g) = w$ (see e.g. [Ca, 3.3]). Let $Y = \operatorname{Hom}(\mathbf{G}_m, \mathbf{T}) \simeq \mathbf{Z}^r$ (resp. Y') be the abelian group of cocharacters of $\mathbf{T}$ (resp. $\mathbf{T}'$); the conjugation isomorphism $\mathbf{T} \to \mathbf{T}'$ gives rise to a conjugation isomorphism $Y \to Y'$ (loc. cit.). Moreover, there is an action of the Frobenius Fr on Y and a canonical isomorphism $T \simeq Y/(\operatorname{Fr} - 1)Y$ (see e.g. [DM, Prop. 13.7]), hence canonical isomorphisms of the character groups $\hat{T}$ and $\hat{T}'$ as subgroups of the characters groups of Y and Y':

$$\hat{T} \simeq \{\chi \,:\, Y \to \mathbf{C}^\times \mid (\operatorname{Fr} - 1)Y \subset \ker \chi\}, \quad \hat{T}' \simeq \{\chi \,:\, Y' \to \mathbf{C}^\times \mid (\operatorname{Fr} - 1)Y' \subset \ker \chi\}.$$

Unraveling the definitions, a simple calculation shows that the condition ${}^w\theta = \theta$ is precisely what is needed to prove that the character χ of Y associated to θ, when "transported" to a character χ' of Y' by the conjugation isomorphism, still satisfies $\ker \chi' \supset (\operatorname{Fr} - 1)Y'$ (see in particular [Ca, Prop. 3.3.4]), so is associated with a character $\theta' \in \hat{T}'$.

Using the characterization of geometric conjugacy in [DM, Prop. 13.8], it is then clear that $(\mathbf{T}, \theta)$ is geometrically conjugate to $(\mathbf{T}', \theta')$, and since $w \neq 1$, the torus $\mathbf{T}'$ is not split. So by contraposition, the lemma is proved. $\square$

Example 7.5. (1) Let ℓ be prime, $r \geqslant 1$ and let $\mathbf{G} = GL(r)/\mathbf{F}_\ell$. Then $G = GL(r, \mathbf{F}_\ell)$, $\mathbf{G}$ is a split connected reductive of rank r and dimension r^2, with connected center of dimension 1. So from Lemma 7.1 and Proposition 7.2, we get

$$A_p(H) \leqslant (\ell+1)^{r(r-1)/2+r/p}$$

for $p \in [1, +\infty]$ for any subgroup H of G, and in particular

$$A_\infty(H) \leqslant (\ell+1)^{r(r-1)/2} \quad \text{and} \quad A_1(H) \leqslant (\ell+1)^{r(r+1)/2}.$$

It would be interesting to know if there are other values of p besides $p = 1$, 2 and $+\infty$ (the latter when q is large enough) for which $A_p(GL(n, \mathbf{F}_q))$ can be computed exactly.

(2) Let $\ell \neq 2$ be prime, $g \geqslant 1$ and let $\mathbf{G} = CSp(2g)/\mathbf{F}_\ell$. Then $G = CSp(2g, \mathbf{F}_\ell)$ and G is a split connected reductive group of rank $g+1$ and dimension $2g^2+g+1$, with connected center. So from Lemma 7.1 and Proposition 7.2, we get

$$A_p(H) \leqslant (\ell+1)^{g^2+(g+1)/p}$$

for $p \in [1, +\infty]$ for any subgroup H of G, and in particular

$$A_\infty(H) \leqslant (\ell+1)^{g^2} \quad \text{and} \quad A_1(H) \leqslant (\ell+1)^{g^2+g+1}.$$

In the case of $G = SL(r, \mathbf{F}_q)$ or $G = Sp(2g, \mathbf{F}_q)$, which correspond to $\mathbf{G}$ where the center is not connected, the bound for $A_\infty(G)$ given by this example is still sharp if we see G as subgroup of $GL(r, \mathbf{F}_q)$ or $CSp(2g, \mathbf{F}_q)$, because both d and r increase by 1, so $d-r$ doesn't change. However, for $A_1(G)$, the exponent increases by one. Here is a slightly different argument that almost recovers the "right" bound.

Lemma 7.6. *Let $\mathbf{G} = SL(n)$ or $Sp(2g)$ over $\mathbf{F}_q$, d the dimension and r the rank of $\mathbf{G}$, and $G = \mathbf{G}(\mathbf{F}_q)$. Then we have the following bounds*

$$A_p(G) \leqslant \kappa^{1/p}(q+1)^{(d-r)/2+r/p}\Bigl(\frac{q+1}{q-1}\Bigr)^{1/p}.$$

and

$$A_p(G) \leqslant (q+1)^{(d-r)/2+r/p}\Bigl(\frac{q+1}{q-1}\Bigr)^{1/p}\Bigl(1+\frac{2\kappa(r+1)|W|}{q-1}\Bigr)^{1/p}$$

for any $p \in [1, +\infty]$, where $\kappa = n$ for $SL(n)$ and $\kappa = 2$ for $Sp(2g)$.

The first bound is better for fixed q, whereas the second is almost as sharp as the bound for $GL(n)$ or $CSp(2g)$ if q is large.

Proof. As we observed before the statement, this holds for $p = +\infty$, so it suffices to consider $p = 1$ and then use the same interpolation argument as for Proposition 7.2.

Let $\mathbf{G}_1 = GL(n)$ or $CSp(2g)$ for $\mathbf{G} = SL(n)$ or $Sp(2g)$ respectively, $G_1 = \mathbf{G}_1(\mathbf{F}_q)$. We use the exact sequence

$$1 \to G \to G_1 \xrightarrow{m} \Gamma = \mathbf{F}_q^\times \to 1$$

(compare with Section 6) where m is either the determinant or the multiplicator of a symplectic similitude. Let ρ be an irreducible representation of G, and as in the proof of Lemma 7.1, let $\pi(\rho)$ be any irreducible representation of G_1 in the induced representation to G_1. The point is that all "twists" $\pi(\rho) \otimes \psi$, where ψ is a character of $\mathbf{F}_q^\times$ lifted to G_1 through m, are isomorphic restricted to G, and hence each $\pi(\rho) \otimes \psi$ contains ρ when restricted to G, and contains even all ρ with the same $\pi(\rho)$. So if $\pi \sim \pi'$, for representations of G_1, denotes isomorphism when restricted to G, we have

$$A_1(G) \leqslant \sum_{\{\pi\}/\sim} \dim \pi$$

where the sum is over a set of representatives for this equivalence relation. On the other hand, $\dim \pi = \dim \pi'$ for $\pi \sim \pi'$, and for each π there are $|\hat{\Gamma}/\hat{\Gamma}^{\pi}|$ distinct representations equivalent to π, with notation as in Lemma 6.1. Hence,

$$A_1(G) \leqslant \frac{1}{q-1} \sum_{\pi} |\hat{\Gamma}^{\pi}| \dim \pi.$$

From, e.g., [Ko1, Lemma 2.3], we know that $\hat{\Gamma}^{\pi}$ has order at most n (for $SL(n)$) or 2 (for $Sp(2g)$), which by applying Proposition 7.2 yields the first bound[8], namely

$$A_1(G) \leqslant \kappa \frac{(q+1)^{(d+r)/2}}{q-1}, \qquad \text{with } \kappa = 2 \text{ or } n.$$

To obtain the refined bound, observe that in the formula (7.7) for the dimension of an irreducible representation ρ of G_1, the first term is zero unless ρ is a regular representation, and the second $t(\rho)$ is smaller by a factor roughly q. If π is regular, we have $\hat{\Gamma}^{\pi} = 1$ by Lemma 7.7 below. So it follows that

$$\begin{aligned} A_1(G) &\leqslant \frac{1}{q-1} \Big\{ \sum_{\pi \text{ regular}} \dim \pi + \kappa \sum_{\pi \text{ not regular}} \dim \pi \Big\} \\ &\leqslant \frac{A_\infty(G_1)}{q-1} q^r(q-1) + \kappa \sum_{\pi \text{ not regular}} t(\rho) \end{aligned}$$

(in the first term, $q^r(q-1)$ is the number of geometric conjugacy classes for G_1, computed as in (7.8), since r is the semi-simple rank of G_1). We have the analogue of (7.11):

$$\sum_{\pi \text{ not regular}} t(\pi) \leqslant \frac{|G_1|_{p'}}{(q-1)^{r+1}} \frac{2(r+1)|W|}{q-1} (q+1)^{r+1} \leqslant 2(r+1)|W| \frac{(q+1)^{(d+r)/2+1}}{q-1},$$

by (7.13) (because

$$\Big| \langle \sum_{\pi \text{ not regular}} \pi, R_{\mathbf{T}}^{\mathbf{G}_1}(\theta) \rangle \Big| \leqslant \langle R_{\mathbf{T}}^{\mathbf{G}_1}(\theta), R_{\mathbf{T}}^{\mathbf{G}_1}(\theta) \rangle,$$

see (7.10), and the same argument leading to (7.9)). The bound

$$A_1(G) \leqslant (q+1)^{(d+r)/2} \Big(1 + \frac{2\kappa(r+1)|W|}{q-1} \Big)$$

follows. □

Lemma 7.7. *Let* $\mathbf{G} = GL(n)$ *or* $CSp(2g)$ *over* $\mathbf{F}_q$, $G = \mathbf{G}(\mathbf{F}_q)$. *For any regular irreducible representation* ρ *of* G, *we have* $\hat{\Gamma}^{\rho} = 1$.

Proof. As above, let $m : \mathbf{G} \to \mathbf{G}_m$ be the determinant or multiplicator character. Let ρ be a regular representation and ψ a character of $\mathbf{F}_q^{\times}$ such that $\rho \otimes \psi \simeq \rho$, where ψ is shorthand for $\psi \circ m$. We wish to show that ψ is trivial to conclude $\hat{\Gamma}^{\rho} = 1$. For this purpose, write

$$\rho = \sum_{(\mathbf{T},\theta) \in \kappa} \frac{\varepsilon_{\mathbf{G}} \varepsilon_{\mathbf{T}} R_{\mathbf{T}}^{\mathbf{G}}(\theta)}{\langle R_{\mathbf{T}}^{\mathbf{G}}(\theta), R_{\mathbf{T}}^{\mathbf{G}}(\theta) \rangle}$$

for some unique geometric conjugacy class κ. We have $R_{\mathbf{T}}^{\mathbf{G}}(\theta) \otimes \psi = R_{\mathbf{T}}^{\mathbf{G}}(\theta(\psi|T))$ (see, e.g., [DM, Prop. 12.6]), so

$$\rho \otimes \psi = \sum_{(\mathbf{T},\theta) \in \kappa} \frac{\varepsilon_{\mathbf{G}} \varepsilon_{\mathbf{T}} R_{\mathbf{T}}^{\mathbf{G}}(\theta(\psi|T))}{\langle R_{\mathbf{T}}^{\mathbf{G}}(\theta), R_{\mathbf{T}}^{\mathbf{G}}(\theta) \rangle}.$$

Since the distinct Deligne-Lusztig characters are orthogonal, the assumption $\rho \simeq \rho \otimes \psi$ implies that for any fixed $(\mathbf{T}, \theta) \in \kappa$, the pair $(\mathbf{T}, \theta(\psi|T))$ is also in the geometric conjugacy class κ. Consider then the translation of this condition using the bijection between geometric conjugacy

[8] This suffices for the applications in this paper.

classes of pairs $(\mathbf{T},\theta)$ and $\mathbf{F}_q$-rational conjugacy classes of semi-simple elements in $\mathbf{G}^*$, the dual group of $\mathbf{G}$ (see, e.g., [DM, Prop. 13.12]). Denote by s the conjugacy class corresponding to $(\mathbf{T},\theta)$. The pair $(\mathbf{T},\psi|T)$ corresponds to a *central* conjugacy class s', because $\psi|T$ is the restriction of a global character of $\mathbf{G}$ (see the proof of [DM, Prop. 13.30]; alternately, use the fact that both global characters and central conjugacy classes are characterized by being invariant under the action of the Weyl group[9]), and the definition of the correspondance shows that $(\mathbf{T},\theta\psi|T)$ corresponds to the conjugacy class ss' (which is well-defined because s' is central). The assumption that $(\mathbf{T},\theta)$ and $(\mathbf{T},\theta\psi|T)$ are geometrically conjugate therefore means $ss'=s$, i.e, $s'=1$, and clearly this means $\psi=1$, as desired. □

Remark 7.8. Here is a mnemonic device to remember the bounds for $A_\infty(G)$ in (7.1)[10]: among the representations of G, we have the principal series $R(\theta)$, parametrized by the characters of a maximal split torus, of which there are about q^r, and those share a common maximal dimension A. Hence

$$q^r A^2 \asymp \sum_\theta \dim(R(\theta))^2 \asymp |G| \sim q^d,$$

so A is of order $q^{(d-r)/2}$. In other words, we expect that in the formula $\sum \dim(\rho)^2 = |G|$, the principal series contributes a positive proportion.

The bound for $A_1(G)$ is also intuitive: there are roughly q^r conjugacy classes, and as many representations, and for a "positive proportion" of them, the degree of the representation is of the maximal size given by $A_\infty(G)$.

8. Probabilistic sieves

The introduction of a general measure space (X,μ) as component of the siftable set may appear yo be an instance of overenthusiastic French abstraction. However, we believe that the generality involved may be useful and that it suggests new problems in a probabilistic setting.

To start with a simple example, let $\Psi = (\mathbf{Z},\{\text{primes}\},\mathbf{Z}\rightarrow\mathbf{Z}/\ell\mathbf{Z})$ be the classical sieve setting. Consider now a probability space $(X,\Sigma,\mathbf{P})$ (i.e., $\mathbf{P}$ is a probability measure on X with respect to a σ-algebra Σ), and let $F=N\,:\,X\rightarrow\mathbf{Z}$ be an integer-valued random variable. Then the triple $(X,\mathbf{P},N)$ is a siftable set, and given any sieving sets (Ω_ℓ) and prime sieve support $\mathcal{L}$, it is tautological that the measure, or rather probability, of the associated sifted set in X is equal to

$$\mathbf{P}(N\in S(\mathbf{Z},\Omega;\mathcal{L}^*)) = \mathbf{P}(\{\omega\in X\ \mid\ N(\omega)\,(\mathrm{mod}\,\ell)\notin\Omega_\ell,\ \text{for all }\ell\in\mathcal{L}^*\}).$$

In other words, the sieve bounds in that context can give estimates for the probability that the values of some integer-valued random variable satisfy any condition that can be described by sieving sets.

If we are given natural integer-valued random variables, this probabilistic setting gives a precise meaning to such notions as "the probability that an integer is squarefree". If the distribution law of N is uniform on an interval $1\leqslant n\leqslant T$, and we let $T\rightarrow+\infty$, this is just the usual "natural density".

Example 8.1. Let N_λ be a random variable with a Poisson distribution of parameter λ, i.e., we have

$$\mathbf{P}(N_\lambda=k)=e^{-\lambda}\frac{\lambda^k}{k!},\qquad\text{for }k\geqslant 0.$$

Then one can easily show, e.g., that the probability that N_λ is squarefree (excluding 0) tends to $\pi^2/6$ as λ goes to ∞.

[9] Think of $\mathbf{T}$ in $GL(n)$ being the diagonal matrices, with the Weyl group $\mathfrak{S}_n$ permuting the diagonal components.

[10] Which explains why it seemed to the author to be a reasonable statement to look for...

The following setting seems to have some interest as a way to get insight into properties of "random" integers $n \in \mathbf{Z}$.

Consider a simple random walk S_n, $n \geqslant 0$, on $\mathbf{Z}$, i.e., a sequence of random variables S_n on X such that $S_0 = 0$ and $S_{n+1} = S_n + X_{n+1}$ with $(X_n)_{n\geqslant 1}$ a sequence of independent random variables with Bernoulli distribution $\mathbf{P}(X_n = \pm 1) = \frac{1}{2}$ (or one could take general Bernoulli distributions $\mathbf{P}(X_n = 1) = p$, $\mathbf{P}(X_n = -1) = q$, for some p, $q \in]0, 1[$ with $p + q = 1$). These variables (S_n) give a natural sequence of siftable sets $(X, \mathbf{P}, S_n)$. It turns out to be quite easy to estimate the corresponding sieve constants; here the dependency on the random variable component of the siftable set is the most important, so we denote $\Delta(S_n, \mathcal{L})$ the sieve constant.

Proposition 8.2. *Let (S_n) be a simple random walk on $\mathbf{Z}$. With notation as above, we have*

$$\Delta(S_n, \mathcal{L}) \leqslant 1 + \Big|\cos\Big(\frac{2\pi}{L^2}\Big)\Big|^n \sum_{m\in\mathcal{L}} m,$$

for $n \geqslant 1$ and for any sieve support $\mathcal{L}$ consisting entirely of odd *squarefree integers $m \leqslant L$.*

It is natural to exclude even integers, simply because $S_n \pmod 2$ is not equidistributed: more precisely, we have $\mathbf{P}(S_n \text{ is even}) = 0$ or 1 depending on whether n itself is even or odd. In probabilistic terms, the random walk is not aperiodic. The simplest way to avoid this problem would be to assume that the increments X_n have distribution

$$\mathbf{P}(X_n = \pm 1) = \mathbf{P}(X_n = 0) = \frac{1}{3}$$

(i.e., at each step the walker may decide to remain still). The reader will have no trouble adapting the arguments below to this case, without parity restrictions.

Proof. We will estimate the "exponential sums", which in the current context, using probabilistic notation $\mathbf{E}(Y) = \int_X Y dP$ for the integral, are simply

$$W(a, b) = \mathbf{E}\Big(e\Big(\frac{a_1 S_n}{m_1}\Big)e\Big(-\frac{a_2 S_n}{m_2}\Big)\Big)$$

for m_1, $m_2 \in \mathcal{L}$, $a_i \in (\mathbf{Z}/m_i\mathbf{Z})^\times$. Using the expression $S_n = X_1 + \cdots + X_n$ for $n \geqslant 1$, independence, and the distribution of the X_i, we obtain straightforwardly

$$W(a, b) = \mathbf{E}\Big(e\Big(\frac{(a_1 m_2 - a_2 m_1)X_1}{m_1 m_2}\Big)\Big)^n = \Big(\cos 2\pi \frac{a_1 m_2 - a_2 m_1}{m_1 m_2}\Big)^n.$$

The condition that m_i are odd, and that $(a_i, m_i) = 1$, imply that $|W(a, b)| = 1$ if and only if $a_1 = a_2$ and $m_1 = m_2$, and otherwise

$$|W(a, b)| \leqslant \Big|\cos \frac{2\pi}{m_1 m_2}\Big|^n.$$

Hence the sieve constant is bounded by

$$\Delta(S_n, \mathcal{L}) \leqslant \max_{m_1, a_1} \Big\{1 + \sum_{m_2} \sideset{}{^*}\sum_{a_2 \pmod{m_2}} \Big|\cos \frac{2\pi}{m_1 m_2}\Big|^n\Big\} \leqslant 1 + \Big|\cos\Big(\frac{2\pi}{L^2}\Big)\Big|^n \sum_{m\in\mathcal{L}} m.$$

□

Corollary 8.3. *With notation as above, we have:*

(1) *For any sieving sets $\Omega_\ell \subset \mathbf{Z}/\ell\mathbf{Z}$ for ℓ odd, $\ell \leqslant L$, and $L \geqslant 3$, we have*

$$\mathbf{P}(S_n \in S(\mathbf{Z}, \Omega; L)) \leqslant \Big(1 + L^2 \exp\Big(-\frac{n\pi^2}{L^4}\Big)\Big) H^{-1}$$

where

$$H = \sideset{}{^\flat}\sum_{\substack{m\leqslant L\\ m \text{ odd}}} \prod_{\ell | m} \frac{|\Omega_\ell|}{\ell - |\Omega_\ell|}.$$

(2) *Let $\varepsilon > 0$ be given, $\varepsilon \leqslant 1/4$. For any odd $q \geqslant 1$, any a coprime with q, we have*

$$\mathbf{P}(S_n \text{ is prime and } \equiv a\,(\mathrm{mod}\,q)) \ll \frac{1}{\varphi(q)}\frac{1}{\log n}$$

if $n \geqslant 2$, $q \leqslant n^{1/4-\varepsilon}$, the implied constant depending only on ε.

Note that (2) is Theorem 1.1 in the introduction.

Proof. For (1), we take $\mathcal{L}$ to be the set of odd squarefree numbers $\leqslant L$ (so $\mathcal{L}^*$ is the set of odd primes $\leqslant L$), and then since $\cos(x) \leqslant 1 - x^2/4$ for $0 \leqslant x \leqslant 2\pi/9$, the proposition gives

$$\Delta \leqslant 1 + L^2\Bigl(1 - \frac{\pi^2}{L^4}\Bigr)^n \leqslant 1 + L^2 \exp\Bigl(-\frac{n\pi^2}{L^4}\Bigr),$$

and the result is a mere restatement of the large sieve inequality.

For (2), we have to change the sieve setting a little bit. Consider the sieve setting above Ψ, *except* that for primes $\ell \mid q$, we take ρ_ℓ to be reduction modulo $\ell^{\nu(\ell)}$, where $\nu(\ell)$ is the ℓ-valuation of q. Take the siftable set $(X, \mathbf{P}, S_n)$, and the sieve support

$$\mathcal{L} = \{mm' \mid mm' \text{ squarefree}, (m, 2q) = 1,\ m \leqslant L/q \text{ and } m' \mid q\},$$

with $\mathcal{L}^*$ still the set of odd primes $\leqslant L$.

Proceeding as in the proof of Proposition 8.2, the sieve constant is bounded straightforwardly by

$$\Delta \leqslant 1 + \Bigl|\cos\frac{2\pi}{L^2}\Bigr|^n \sum_{\substack{m \leqslant L/q \\ (m,2q)=1}} \sum_{m'|q} mq \leqslant 1 + \tau(q)q^{-1}L^2 \exp\Bigl(-\frac{n\pi^2}{L^4}\Bigr),$$

where $\tau(q)$ is the number of divisors of q.

Finally, take

$$\Omega_\ell = \begin{cases} \{0\} & \text{if } \ell \nmid q, \\ \mathbf{Z}/\ell^{\nu(\ell)}\mathbf{Z} - \{a\} & \text{if } \ell \mid q. \end{cases}$$

If S_n is a prime number congruent to a mod q, then we have $S_n \in S(\mathbf{Z}, \Omega; \mathcal{L}^*)$, hence

$$\mathbf{P}(S_n \text{ is a prime } \equiv a\,(\mathrm{mod}\,q)) \leqslant \mathbf{P}(S_n \in S(\mathbf{Z}, \Omega; \mathcal{L}^*)) \leqslant \Delta H^{-1}$$

where

$$H = \sideset{}{^\flat}\sum_{\substack{m \leqslant L/q \\ (m,2q)=1}} \sideset{}{^\flat}\sum_{m'|q} \frac{\varphi^*(m')}{\varphi(m)}, \qquad \text{with} \qquad \varphi^*(n) = \prod_{\ell^\nu || n} (\ell^\nu - 1).$$

Now the desired estimate follows on taking $q \leqslant n^{1/4-\varepsilon}$ and $L = qn^\varepsilon$, using the classical lower bound (see e.g. [B], [IK, (6.82)])

$$\sideset{}{^\flat}\sum_{\substack{m \leqslant L/q \\ (m,2q)=1}} \frac{1}{\varphi(m)} \geqslant \frac{\varphi(q)}{2q} \log L/q \gg \frac{\varphi(q)}{q} \log n$$

(the implied constant depending only on ε) together with the cute identity

$$\sideset{}{^\flat}\sum_{m'|q} \varphi^*(m') = q$$

which is trivially verified by multiplicativity. □

Remark 8.4. (1) It is important to keep in mind that, by the Central Limit Theorem, $|S_n|$ is usually of order of magnitude $\sqrt{n}$ (precisely, $S_n/\sqrt{n}$ converges weakly to the normal distribution with variance 1 as $n \to +\infty$). So the estimate $\Delta \leqslant 1 + L^2 \exp(-n\pi^2/L^4)$, which gives a non-trivial result in applications as long as, roughly speaking, $L \leqslant n^{1/4}/(\log n)^{1/4}$, compares well with the classical large sieve for integers $n \leqslant N$, where $\Delta \leqslant N - 1 + L^2$, which is non-trivial for $L \leqslant \sqrt{N}$.

(2) The second part is an analogue of the Brun-Titchmarsh inequality, namely (in its original form)

$$\pi(x;q,a) \ll \frac{1}{\varphi(q)}\frac{x}{\log x}$$

for $x \geqslant 2$, $(a,q)=1$ and $q \leqslant x^{1-\varepsilon}$, the implied constant depending only on $\varepsilon > 0$. However, from the previous remark we see that it is weaker than could be expected, namely $q \leqslant n^{1/4-\varepsilon}$ would have to be replaced by $q \leqslant n^{1/2-\varepsilon}$. Here we have exploited the flexibility of the sieve setting and sieve support. For a different use of this flexibility, see Section 11; we want to point out here that the possibility of using a careful non-obvious choice of $\mathcal{L}$ was first exploited by Zywina in his preprint [Z].

It would be quite interesting to know if the extension to $q \leqslant n^{1/2-\varepsilon}$ holds. The point is that if we try to adapt the classical method, which is to sieve for those k, $1 \leqslant k \leqslant x/q$, such that $qk+a$ is prime, we are led to some interesting and non-obvious (for the author) probabilistic issues; indeed, if $S_n \equiv a \,(\mathrm{mod}\, q)$, the (random) integer k such that $S_n = kq+a$ can be described as follows: we have $k = T_N$ where N is a random variable

$$N = |\{m \leqslant n \mid S_m \equiv a \,(\mathrm{mod}\, q)\}|$$

and (T_i) is a random walk with initial distribution given by

$$\mathbf{P}(T_0 = 0) = 1 - \frac{a}{q}, \quad \mathbf{P}(T_0 = -1) = \frac{a}{q},$$

and independent identically distributed increments $V_i = T_i - T_{i-1}$ such that

$$\mathbf{P}(V_i = 0) = 1 - \frac{1}{q}, \qquad \mathbf{P}(V_i = \pm 1) = \frac{1}{2q}.$$

So what is needed is to perform sieve on the siftable set $(\{S_n \equiv a \,(\mathrm{mod}\, q)\}, \mathbf{P}, T_N)$. Since the length N of the auxiliary walk is random, this requires some care, and we hope to come back to this. Note at least that if look at the same problem with $(\Omega, \mathbf{P}, T_i)$ for a fixed i, then we easily get by sieving

$$\mathbf{P}(qT_i + a \text{ is prime}) \ll \frac{1}{\varphi(q)}\frac{1}{\log i}$$

for all $q \leqslant i^{1/2-\varepsilon}$, $\varepsilon > 0$, the implied constant depending only on ε.

(3) Obviously, it would be very interesting to derive lower bounds or asymptotic formulas for $\mathbf{P}(S_n \text{ is prime})$ for instance, and for other analogues of classical problems of analytic number theory. Note that it is tempting to attack the problems with "local" versions of the Central Limit Theorem and summation by parts to reduce to the purely arithmetic deterministic case. Problems where such a reduction is not feasible would of course be more interesting.

In the next section, we will give another example of probabilistic sieve, similar in spirit to the above, although the basic setting will be rather deeper, and the results are not accessible to a simple summation by parts.

Finally, we remark that this probabilistic point of view should not be mistaken with "probabilistic models" of integers (or primes), such as Cramer's model: the values of the random variables we have discussed are perfectly genuine integers.[11]

9. Sieving in arithmetic groups

We now start discussing examples of sieve settings which seem to be either new, or have only been approached very recently. The first example concerns sieving for elements in an arithmetic group G. There are actually a number of different types of siftable sets that one may consider here.

[11] To give a caricatural example, if it were possible to show that, for some sequence of random variables N_n distributed on disjoint subsets of integers, the probability $\mathbf{P}(N_n \text{ and } N_n+2 \text{ are both primes})$ is always strictly positive, then the twin-prime conjecture would follow.

Maybe the most obvious idea for analytic number theorists is to take the group sieve setting defined by

$$\Psi = (SL(n, \mathbf{Z}), \{\text{primes}\}, G \to SL(n, \mathbf{F}_\ell))$$

(where the last reduction map is known to be surjective for all ℓ), and look at the siftable set $X \subset G$ which is the set of those matrices with norm bounded by some quantity T, with $F_x = x$ and counting measure. In other words, instead of sieving integers, we want to sieve integral unimodular matrices. Of course, $SL(n, \mathbf{Z})$ may be replaced with other arithmetic groups, even possibly with infinite-index subgroups.

Here the equidistribution approach leads to hyperbolic lattice point problems (in the case $n = 2$), and generalizations of those for $n \geqslant 3$. The issue of uniformity with respect to q when taking "congruence towers" $\Gamma \cap \Gamma(q)$, where $\Gamma(q)$ is the principal congruence subgroup, is the main issue, compared with the results available in the literature (e.g., the work of Duke, Rudnick and Sarnak [DRS] gives individual equidistribution, with methods that may be amenable to uniform treatments, whereas more recent ergodic-theoretic methods by Eskin, Mozes, McMullen, Shah and other, see e.g. [EMS], seem to be more problematic in this respect).

A tentative and very natural application is the natural fact that "almost all" unimodular matrices with norm $\leqslant T$ have irreducible characteristic polynomial, with an estimate for the number of exceptional matrices (this question was also recently formulated by Rivin [R, Conj. 8], where it is observed that the qualitative form of this statement is likely to follow from the results of Duke, Rudnick and Sarnak). The case of integral matrices with arbitrary determinant can be treated very quickly as a simple consequence of the higher-dimensional large sieve as in [G] (in other words, embed invertible matrices in the additive group $G = M(n, \mathbf{Z}) \simeq \mathbf{Z}^{n^2}$, and use abelian harmonic analysis).

The setting of arithmetic groups suggests other types of siftable sets, which are of a more combinatorial flavor, and the "probabilistic" theme of Section 8 is also a natural fit.[12] Theorem 1.2 gives some first results of this kind.

Let G be a finitely generated group. Assuming a symmetric set of generators S to be fixed (i.e., with $S^{-1} = S$), three siftable sets (X, μ, F) of great interest arise naturally:
– the set X of elements $g \in G$ with word-length metric $\ell_S(g)$ at most N, for some integer $N \geqslant 1$, i.e, the set of those elements $g \in G$ that can be written as

$$g = s_1 \cdots s_k$$

with $k \leqslant N$, $s_i \in S$ for $1 \leqslant i \leqslant k$. Here we take $F_x = x$ for $x \in X$, and of course μ is the counting measure.
– the set W of words of length N in the alphabet S, for some integer $N \geqslant 1$, with F_w the "value" in G of the word $w \in W$, i.e., the image of w by the natural (surjective) homomorphism $F(S) \to G$ from the free group generated by S to G. Again μ is the counting measure.
– as in Section 8, we may consider a probabilistic siftable set $(\Omega, \mathbf{P}, X_N)$, where Ω is some probability space, $\mathbf{P}$ the associated probability measure, and $X_N = \xi_1 \cdots \xi_N$, where (ξ_k) is a N-uple of S-valued random variables. The simplest case is when (ξ_k) is an independent vector, and the distribution of each ξ_k is uniform: $\mathbf{P}(\xi_k = s) = 1/|S|$. In other words, X_N is then the N-th step in the simple left-invariant random walk on G given by S. If $G = \mathbf{Z}$ and $S = \{\pm 1\}$, we considered this in Section 8.

Remark 9.1. Note that the last two examples are in fact equivalent: we have

$$\mathbf{P}(X_N \in A) = \frac{1}{|W|} |\{w \in W \mid F_w \in A\}|$$

for any subset $A \subset G$. (Since $|W| = |S|^N$, this explains why the two statements of Theorem 1.2 are equivalent). Although this reduces one particular probabilistic case to a "counting" sieve, we may indeed wish to vary the distribution of the factors ξ_k of the random walk, and doing so

[12] A useful survey on combinatorial and geometry group theory is given in the book of de la Harpe [Ha], and a survey of random walks on groups is that of Saloff-Coste [SC].

would not in general lead to such a reduction; even when possible, this may not be desirable, because it would involve rather artificial constructs. For instance, another natural type of random walk is the random walk given by factors ξ_k where

$$\mathbf{P}(\xi_k = s) = \frac{1}{|S|+1} \quad \text{for } s \in S, \qquad \mathbf{P}(\xi_k = 1) = \frac{1}{|S|+1}.$$

This is also equivalent to replacing S by $S \cup \{1\}$ (if $1 \notin S$ at least), but the set of words w where each component w_i may be the identity is not vary natural.

We now provide a concrete example by proving Theorems 1.2 and 1.3, indeed in a slightly more general case. Let $\mathbf{G}$ be either $SL(n)$ or $Sp(2g)$ for some $n \geqslant 3$ or $g \geqslant 2$, let $G = \mathbf{G}(\mathbf{Z})$, and let S be a symmetric set of generators for G (for instance, the elementary matrices with ± 1 off the diagonal are generators of $SL(n, \mathbf{Z})$, see Remark 9.9).[13] We consider either the group sieve setting

$$\Psi = (G, \{\text{primes}\}, G \to G_\ell)$$

where $G_\ell = \mathbf{G}(\mathbf{Z}/\ell\mathbf{Z})$ is $SL(n, \mathbf{F}_\ell)$ or $Sp(2g, \mathbf{F}_\ell)$, and the maps are simply reduction modulo ℓ, or the induced conjugacy sieve setting. It is well-known that the reduction maps are onto for all ℓ (see e.g. [Shi, Lemma 1.38] for the case of $SL(n)$).

We will look here at the second type of siftable set $\Upsilon = (W, \mu, F)$, i.e., W is the set of words of length N in S, and F_w is the "value" of a word w in G. Equivalently, we consider the simple left-invariant random walk (X_n). In that case, the qualitative form of Theorem 1.2 was proved by Rivin [R], and the latest version of Rivin's preprint also discusses quantitative forms of equidistribution in G_ℓ, using Property (T) in a manner analogous to what we do.

We will obtain a bound for the large sieve constant by appealing to (6.4) and its analogue for the group sieve setting, estimating the exponential sums $W(\pi, \tau)$ or $W(\varphi_{\pi,e,f}, \varphi_{\tau,e',f'})$ of (4.2).[14] The crucial ingredient is the so-called "Property (τ)".

Proposition 9.2. *Let Γ be a finitely generated group, I an arbitrary index set and $\rho_i : \Gamma \to G_i$ for $i \in I$ a family of surjective homomorphisms onto finite groups, such that Γ has Property (τ) with respect to the family $(\ker \rho_i)$ of finite index subgroups of Γ.*

Let $S = S^{-1}$ be a symmetric finite generating set of Γ, and for $N \geqslant 1$, let $W = W_N$ denote the set of words of length N in the alphabet S, and let F_w denote the value of the word w in Γ. Assume that there exists a word r in the alphabet S of odd *length c such that $F_r = \mathrm{Id} \in \Gamma$.*

Then there exists $\alpha > 0$ such that for any $i \in I$, any representation $\pi : \Gamma \to GL(V)$ that factors through G_i and does not contain the trivial representation, any vectors e, f in the space of π, we have

$$\Big| \sum_{w \in W} \langle \pi(F_w) e, f \rangle \Big| \leqslant \|e\| \|f\| |W|^{1-\alpha}, \tag{9.1}$$

for $N \geqslant 1$, where $\langle \cdot, \cdot \rangle$ is a Γ-invariant inner product on V, and hence

$$\Big| \sum_{w \in W} \mathrm{Tr}\, \pi(F_w) \Big| \leqslant (\dim \pi) |W|^{1-\alpha}. \tag{9.2}$$

The constant α depends only on Γ, $|S|$, the (τ)-constant for $(S, \Gamma, \ker \rho_i)$ and the length c of the relation r.

We will recall briefly the definition of Property (τ) and the associated (τ)-constant in the course of the proof; see e.g. [Lu, §4.3] or [LZ] for more complete surveys. This should also be compared with [SC, Th. 6.15].

[13] We will also comment briefly on what happens for $G = SL(2, \mathbf{Z})$.

[14] Of course the equidistribution approach may also be used, but it is less efficient and not really quicker or simpler.

Proof. Let $i \in I$ and let π be a representation that factors through G_i and does not contain the trivial representation. Clearly (9.2) follows from (9.1) since the trace of a matrix is the sum of the diagonal matrix coefficients in an orthonormal basis.

Let

$$M = \frac{1}{|S|}\sum_{s\in S}\pi(s), \qquad M' = \mathrm{Id} - M,$$

which are both self-adjoint elements of the endomorphism ring $\mathrm{End}(V)$, since $S = S^{-1}$. We then find by definition

$$\frac{1}{|W|}\sum_{w\in W}\langle \pi(F_w)e, f\rangle = \langle M^N e, f\rangle.$$

Let $\rho \geqslant 0$ be the spectral radius of M, or equivalently the largest of absolute values of the eigenvalues of M, which are real since M is self-adjoint. Then by Cauchy's inequality we have

$$|\langle M^N e, f\rangle| \leqslant \|e\|\|f\|\rho^N,$$

so that it only remains to prove that there exists $\delta > 0$, independent of i and π, such that $\rho \leqslant 1 - \delta$.

Clearly $\rho = \max(\rho_+, \rho_-)$, where $\rho_+ \in \mathbf{R}$ (resp. ρ_-) is the largest eigenvalue and ρ_- is the opposite of the smallest eigenvalue (if it is negative) and 0 otherwise. We bound each $\rho_\pm$ separately, proving $\rho_\pm \leqslant 1 - \delta_\pm$ with $\delta_\pm$ independent of i and π.

For ρ_+, it is equivalent (by the variational characterization of the smallest eigenvalue) to prove that there exists $\delta_+ > 0$, independent of i and π, such that

$$\frac{\langle M'(v), v\rangle}{\langle v, v\rangle} \geqslant \delta_+$$

for any non-zero vector $v \in V$. But a simple and familiar computation yields

$$\frac{1}{|S|}\sum_{s\in S}\|\pi(s)v - v\|^2 = 2\langle M'(v), v\rangle$$

and therefore tautologically we have

$$\frac{\langle M'(v), v\rangle}{\langle v, v\rangle} \geqslant \frac{1}{2|S|}\inf_{\varpi}\inf_{v\neq 0}\max_{s\in S}\frac{\|\varpi(s)v - v\|^2}{\|v\|^2}, \tag{9.3}$$

where ϖ ranges over *all* unitary representations of Γ that factor through some $\ker\rho_i$ and do not contain the trivial representation (and $\|\cdot\|$ on the right-hand side is the unitary norm for each such representation). But it is precisely the content of Property (τ) for Γ with respect to $(\ker\rho_i)$ that this triple extremum is > 0 (see e.g. [Lu, Def. 4.3.1]).

So we come to ρ_-. Here a suitable lower-bound follows from Theorem 6.6 of [SC] (due to Diaconis, Saloff-Coste, Stroock), using the fact that any eigenvalue of M is also an eigenvalue of M_{reg}, where M_{reg} is the analogue of M for the regular representation of Γ on $L^2(\Gamma/\ker\rho_i)$.

For completeness, we prove what is needed here, adapting the arguments to the case of a general representation. It suffices to prove that there exists $\delta_- > 0$ independent of i and π such that

$$\frac{\langle M''(v), v\rangle}{\langle v, v\rangle} \geqslant \delta_- \tag{9.4}$$

for all non-zero $v \in V$, where now $M'' = \mathrm{Id} + M$. We have

$$2\langle M''(v), v\rangle = \frac{1}{|S|}\sum_{s\in S}\|\pi(s)v + v\|^2.$$

Now let $r = s_1\cdots s_c$ be a word of odd length c in the alphabet S such that r is trivial in Γ; denote

$$r_k = s_1\cdots s_k \text{ for } 1 \leqslant k \leqslant r, \quad r_0 = 1.$$

For $v \in V$, we can write

$$v = \frac{1}{2}\Big((v + \pi(s_1)v) - (\pi(s_1)v + \pi(s_1 s_2 v)) + \cdots + (\pi(s_1 \cdots s_{c-1})v + \pi(1)v)\Big)$$

(the odd length is used here), hence by Cauchy's inequality we get

$$\|v\|^2 \leqslant \frac{c}{4} \sum_{i=0}^{c-1} \|\pi(r_i)v + \pi(r_i s_{i+1})v\|^2 = \frac{c}{4} \sum_{i=0}^{c-1} \|v + \pi(s_{i+1})v\|^2$$

(the representation is unitary). By positivity, since at worst all s_i are equal to the same generator in S, we get

$$\|v\|^2 \leqslant \frac{c^2}{4} \sum_{s \in S} \|\pi(s)v + v\|^2 = \frac{c^2|S|}{2} \langle M''(v), v\rangle, \tag{9.5}$$

which implies (9.4) with $\delta_- = \frac{2}{c^2|S|} > 0$. □

Remark 9.3. The odd-looking assumption on the existence of r is indeed necessary for such a general statement, because of periodicity issues. Namely, if (and in fact only if) all S-relations in Γ are of even length, the Cayley graph[15] $C(\Gamma, s)$ of Γ with respect to S is bipartite[16], and so are its finite quotients $C(\Gamma/\ker \rho_i, S)$. In that case, it is well-known and easy to see that -1 is an eigenvalue of M_{reg} (the operator M for the regular representation; take the function such that $f(x)$ equals ± 1 depending on whether the point is at even or odd distance from the origin) and the argument above fails. Alternately, this can be seen directly with the exponential sums: the relations being of even length implies that there is a well-defined surjective homomorphism $\varepsilon : \Gamma \to \mathbf{Z}/2\mathbf{Z}$ with $\varepsilon(s) = -1$ for $s \in S$. Viewing ε as a representation $\Gamma \to \{\pm 1\} \subset \mathbf{C}^\times$, we have

$$\sum_{w \in W} \varepsilon(F_w) = \begin{cases} |W| & \text{if } N \text{ is even} \\ -|W| & \text{if } N \text{ is odd.} \end{cases}$$

We will describe an example of this for $\Gamma = SL(2, \mathbf{Z})$ below.

The simplest way of ensuring that r exists is to assume that $1 \in S$; geometrically, this means each vertex of the Cayley graph has a self-loop, and probabilistically, this means that one considers a "lazy" random walk on the Cayley graph, with probability $1/|S|$ of staying at the given element.

In fact, if we consider the effect of replacing S by $S' = S \cup \{1\}$ (in the case where $1 \notin S$), we have

$$M_{S'} = \Big(1 - \frac{1}{|S'|}\Big) M_S + \frac{1}{|S'|},$$

with obvious notation, and so we obtain

$$\rho_- \geqslant -1 + \frac{2}{|S'|}$$

directly (which is the same lower bound as the one we proved, in the case $c = 1$).

With the estimate of Proposition 9.2, we can perform some sieve.

Theorem 9.4. *Let $\mathbf{G} = SL(n)$, $n \geqslant 3$, or $Sp(2g)$, $g \geqslant 2$, be as before, $G = \mathbf{G}(\mathbf{Z})$, and let*

$$\Psi = (G, \{primes\}, G \to G_\ell = \mathbf{G}(\mathbf{Z}/\ell\mathbf{Z}))$$

be the group sieve setting. Let $S = S^{-1}$ be a symmetric generating set for G, (W, F) the siftable set of random products of length N of elements of S.

[15] Recall $C(\Gamma, S)$ has vertex set Γ and as many edges from g_1 to g_2 as there are elements $s \in S$ such that $g_2 = g_1 s$; this allows both loops and multiple edges, and those will occur if $1 \in S$ or, in the Cayley graphs of quotients of Γ, if two generators have the same image.

[16] I.e., the vertex set Γ is partitioned in two pieces $\Gamma_\pm$ and edges always go from one piece to another.

(1) *For any sieve support $\mathcal{L}$, the large sieve constant for the induced conjugacy sieve satisfies*

$$\Delta(W, \mathcal{L}) \leqslant |W| + |W|^{1-\alpha} R(\mathcal{L}), \tag{9.6}$$

where $\alpha > 0$ is a constant depending only on $\mathbf{G}$ and S and [17]

$$R(\mathcal{L}) = \max_{m \in \mathcal{L}} \Bigl\{ A_\infty(G_m) \Bigr\} \times \sum_{n \in \mathcal{L}} A_1(G_n).$$

(2) *There exists $\eta > 0$ such that*

$$|\{w \in W \mid \det(F_w - T) \in \mathbf{Z}[T] \text{ is reducible } \}| \ll |W|^{1-\eta} \tag{9.7}$$

where η and the implied constant depend only on $\mathbf{G}$ and S.

(3) *For any sieve support $\mathcal{L}$, the large sieve constant for the group sieve satisfies*

$$\Delta(W, \mathcal{L}) \leqslant |W| + |W|^{1-\alpha} \tilde{R}(\mathcal{L}), \tag{9.8}$$

where $\alpha > 0$ is as above and

$$\tilde{R}(\mathcal{L}) = \max_{m \in \mathcal{L}} \Bigl\{ \sqrt{A_\infty(G_m)} \Bigr\} \times \sum_{n \in \mathcal{L}} A_{5/2}(G_n)^{5/2}.$$

(4) *There exists $\beta > 0$ such that*

$$|\{w \in W \mid \text{one entry of } F_w \text{ is a square } \}| \ll |W|^{1-\beta} \tag{9.9}$$

where β and the implied constant depend only on $\mathbf{G}$ and S.

It is clear that the fourth part implies Theorem 1.3.

Lemma 9.5. *Let $\mathbf{G}$ be as above.*

(1) *Property (τ) holds for the group $G = \mathbf{G}(\mathbf{Z})$ with respect to the family of congruence subgroups* $(\ker(G \to \mathbf{G}(\mathbf{Z}/d\mathbf{Z})))_{d \geqslant 1}$.

(2) *For any symmetric generating set $S = S^{-1}$, there exists an S-relation of odd length.*

Proof. (1) This is well-known; in fact, the group G is a lattice in a semisimple real Lie group with $\mathbf{R}$-rank $\geqslant 2$, and hence it satisfies the stronger Property (T) of Kazhdan, which means that in (9.3), the infimum may be taken on *all* unitary representations of G not containing the trivial representation and remains > 0 (see, e.g., [HV, Cor. 3.5], [Lu, Prop.3.2.3, Ex. 3.2.4, §4.4])).

(2) If all S-relations are of even length, the homomorphism

$$F(S) \to \{\pm 1\}$$

defined by $s \mapsto -1$ induces a non-trivial homomorphism $G \to \{\pm 1\}$. However, there is no such homomorphism for the groups under consideration (e.g., because its kernel H will be a finite index normal subgroup, hence by the Congruence Subgroup Property, due to Mennicke and Bass-Lazard-Serre in this case, see [BMS, p. 64] for references, will factor through a principal congruence subgroup $\ker(G \to \mathbf{G}(\mathbf{Z}/d\mathbf{Z}))$ for some integer $d \geqslant 1$, defining a non-trivial homomorphism[18] $\mathbf{G}(\mathbf{Z}/d\mathbf{Z}) \to \{\pm 1\}$, which is impossible since $\mathbf{G}(\mathbf{Z}/d\mathbf{Z})$ is its own commutator group). ◻

Proof of Theorem 9.4. (1) Let m, $n \in \mathcal{L}$, π, $\tau \in \Pi_m^*$, Π_n^* respectively. Since the maps $G \to \mathbf{G}(\mathbf{Z}/d\mathbf{Z})$ are onto for all d (e.g., because the family (ρ_d) is linearly disjoint in the sense of Definition 2.3, by Goursat's lemma, as in [Ch, Prop. 5.1]), we have in fact $G_{[m,n]} = \mathbf{G}(\mathbf{Z}/[m,n]\mathbf{Z})$.

By Lemma 6.4, the representation $[\pi, \bar{\tau}]$ of $G_{[m,n]}$ defined in (4.3) contains the identity representation if and only if $(m, \pi) = (n, \tau)$, and then contains it with multiplicity one. Let $[\pi, \bar{\tau}]_0$ denote the orthogonal of the trivial component in the second case, and $[\pi, \bar{\tau}]_0 = [\pi, \bar{\tau}]$ otherwise.

[17] With notation as in Section 7.

[18] Here we use the fact that $G \to \mathbf{G}(\mathbf{Z}/d\mathbf{Z})$ is surjective, see the first line of the next proof.

We can now appeal to Proposition 9.2 applied to the representation $[\pi,\bar{\tau}]_0\circ\rho_{[m,n]}$ of G, using the family $(\rho_d : G \to \mathbf{G}(\mathbf{Z}/d\mathbf{Z}))$ of congruence subgroups (since $\ker[\pi,\bar{\tau}]_0\circ\rho_{[m,n]} \supset G_{[m,n]}$). The previous lemma ensures that all required assumptions on S and this family are valid, and by (9.2), the conclusion is the estimate

$$\Bigl|W(\pi,\tau)-\delta(\pi,\tau)|W|\Bigr| \leqslant (\dim\pi)(\dim\tau)|W|^{1-\alpha}$$

for the exponential sum (4.1), where α depends only on $\mathbf{G}$, S and the relevant (τ) or (T) constant.

By Proposition 2.10, we obtain

$$\Delta(W,\mathcal{L}) \leqslant |W| + |W|^{1-\alpha}\max_{m\in\mathcal{L}} A_\infty(G_m)\sum_{n\in\mathcal{L}} A_1(G_n),$$

as stated.

(3) This is exactly similar, except that now we use the basis of matrix coefficients for the group sieve setting, and correspondingly we appeal to (9.1) and the fact (see the final paragraphs of Section 4) that the sums $W(\varphi_{\pi,e,f},\varphi_{\tau,e',f'})$ are (up to the factor $\sqrt{(\dim\pi)(\dim\tau)}$) of the type considered in (9.1).

In the case where $[\pi,\tau]$ contains the trivial representation (i.e., if $(m,\pi)=(n,\tau)$), we also use the fact that, when identified with $\mathrm{End}(V_\pi)$, the one-dimensional space of invariant vectors in $\pi\otimes\bar{\pi}=[\pi,\pi]$ is spanned by homotheties and the orthogonal projection of a linear map $u\in\mathrm{End}(V_\pi)$ is multiplication by $\mathrm{Tr}(u)/\sqrt{\dim\pi}$ (this is a corollary of the orthogonality relations; note that $\|\mathrm{Id}\|^2=\dim\pi$). This means that for a rank 1 linear map of the form $u=e\otimes\bar{e}'$ (where e is in the space of π, and e' in that of the contragredient), the projection is the multiplication by $\langle e,e'\rangle/\sqrt{\dim\pi}$. Since the vectors are part of an orthonormal basis, we get

$$\begin{aligned}\langle(\pi\otimes\bar{\pi})(e\otimes e'),f\otimes f'\rangle &= \frac{\langle e,e'\rangle\langle f,f'\rangle}{\dim\pi}+\langle[\pi,\pi]_0(e\otimes e'),f\otimes f'\rangle\\ &= \frac{\delta((e,f),(e',f'))}{\dim\pi}+\langle[\pi,\pi]_0(e\otimes e'),f\otimes f'\rangle.\end{aligned}$$

Altogether, we obtain

$$\Bigl|W(\varphi_{\pi,e,f},\varphi_{\tau,e',f'})-\delta((\pi,e,f),(\tau,e',f'))|W|\Bigr| \leqslant \sqrt{(\dim\pi)(\dim\tau)}|W|^{1-\alpha}$$

and hence

$$\begin{aligned}\Delta(W,\mathcal{L}) &\leqslant |W|+|W|^{1-\alpha}\max_{m,\pi,e,f}\sqrt{\dim\pi}\sum_{n\in\mathcal{L}}\sum_{\tau,e',f'}\sqrt{\dim\tau}\\ &\leqslant |W|+|W|^{1-\alpha}\max_{m\in\mathcal{L}}\sqrt{A_\infty(G_m)}\sum_{n\in\mathcal{L}}\sum_{\tau}(\dim\tau)^{5/2}\\ &= |W|+|W|^{1-\alpha}\max_{m\in\mathcal{L}}\sqrt{A_\infty(G_m)}\sum_{n\in\mathcal{L}}A_{5/2}(G_n)^{5/2}.\end{aligned}$$

(2) To obtain (9.7), we apply the large sieve inequality for group sieves of Proposition 4.1, using (9.6). This is completely standard; without trying to get the sharpest result (see Section 11 for more refined arguments in a similar end-game), we select the prime sieve support $\mathcal{L}^*=\{\ell\leqslant L\}$ for some $L\geqslant 2$, and take $\mathcal{L}=\mathcal{L}^*$ (pedantically, the singletons of elements of $\mathcal{L}^*$...). Letting $d=n^2-1$, $r=n-1$ (for $\mathbf{G}=SL(n)$) or $d=2g^2+g$, $r=g$ (for $\mathbf{G}=Sp(2g)$), we have (using Lemma 7.6 of Section 7)

$$R(\mathcal{L})\ll L^{d+1},$$

for $L\geqslant 2$, the implied constant depending only on $\mathbf{G}$.

We take for sieving sets the conjugacy classes in the set $\Omega_\ell\subset\mathbf{G}(\mathbf{F}_\ell)$ of matrices with irreducible characteristic polynomial in $\mathbf{F}_\ell[T]$. From (1) of Proposition B.1 in Appendix B, we obtain

$$\frac{|\Omega_\ell|}{|G_\ell|} \gg 1 \tag{9.10}$$

for $\ell \geqslant 3$, where the implied constant depends on $\mathbf{G}$ (compare with [Ch, §3], [Ko1, Lemma 7.2]).

Since those $w \in W$ for which $\det(F_w - T)$ is reducible are contained in the sifted set $S(W; \Omega, \mathcal{L}^*)$, we have by Proposition 4.1

$$|\{w \in W \mid \det(F_w - T) \in \mathbf{Z}[T] \text{ is reducible }\}| \leqslant \Delta H^{-1} \ll (|W| + |W|^{1-\alpha} L^{d+1}) H^{-1},$$

where $H \gg \pi(L)$ by (9.10). Taking $L = |W|^{\alpha/(d+1)}$, we get the bound stated.

(4) Clearly, it suffices to prove the estimate for the number of $w \in W$ for which the (i,j)-th component of F_w is a square, where i and j are fixed integers from 1 to n or $2g$ in the $SL(n)$ and $Sp(2g)$ cases respectively. The principle is similar, using (2) to estimate the large sieve constant for the sieve where $\mathcal{L}^* = \{\ell \leqslant L\}$, $\mathcal{L} = \mathcal{L}^*$, with

$$\Omega_\ell = \{g = (g_{\alpha,\beta}) \in \mathbf{G}(\mathbf{F}_\ell) \mid g_{i,j} \text{ is not a square in } \mathbf{F}_\ell\}.$$

We get by Lemma 7.6 the bound

$$\tilde{R}(\mathcal{L}) \ll L^{1+(3d-r)/2}$$

for $L \geqslant 2$, where the implied constant depends only on $\mathbf{G}$.

Next by (2) of Proposition B.1, we have

$$\frac{|\Omega_\ell|}{|G_\ell|} \gg 1$$

for $L \geqslant 3$ (for $L = 2$, the left-hand side may vanish for $SL(2)$), where the implied constant depends only on $\mathbf{G}$. (The proof in Appendix B uses the Riemann Hypothesis over finite fields; the reader may find it interesting to see whether a more elementary argument may be found).

Hence the sieve bound is

$$|\{w \in W \mid \text{ the } (i,j)\text{-th entry of } F_w \text{ is a square}\}| \leqslant (|W| + \tilde{R}(L)) H^{-1}$$

with $H \gg \pi(L)$ for $L \geqslant 3$, the implied constant depending on $\mathbf{G}$. We take $L = |W|^{\alpha/(1+(3d-r)/2)}$ if this is $\geqslant 3$ and then obtain (9.9). To deal with those N for which this L is < 3, we just enlarge the implied constant in (9.9). ☐

From part (2) of this theorem, we can easily deduce Theorem 1.2.

Corollary 9.6. *Let $\mathbf{G} = SL(n)$, $n \geqslant 2$, or $Sp(2g)$, $g \geqslant 1$, let $G = \mathbf{G}(\mathbf{Z})$ and let $S = S^{-1}$ be a symmetric generating set of G. Let (X_k) be the associated simple left-invariant random walk on G. Then almost surely there exist only finitely many k such that $\det(X_k - T)$ is a reducible polynomial.*

Part of the point of this statement is that it *requires* some quantitative estimate for the probability that X_k has reducible characteristic polynomial.

Proof. For $n \geqslant 3$ (resp. $g \geqslant 2$), it suffices to apply the "easy" Borel-Cantelli lemma[19], since the estimate (2) above for $N = k$ is equivalent with

$$\mathbf{P}(\det(X_k - T) \text{ is reducible}) \ll \exp(-\alpha k)$$

with $\alpha = \eta \log |S| > 0$, and this shows that the series

$$\sum_{k \geqslant 0} \mathbf{P}(\det(X_k - T) \text{ is reducible})$$

converges. From the weaker bound (9.11) in Remark 9.10 below, we see that this series remains convergent for $SL(2,\mathbf{Z}) = Sp(2,\mathbf{Z})$. ☐

[19] If A_n are events in a probability space such that the series $\sum \mathbf{P}(A_n)$ converges, then almost surely ω belongs to only finitely many A_n.

The next corollary is a geometric application which answers a question of Maher [Ma, Question 1.3], and was suggested by Rivin's paper [R]. See [Iv] for a survey of the mapping class group of surfaces, [FLP, Exp. 1, 9] for information on pseudo-Anosov diffeomorphisms of surfaces.

Corollary 9.7. *Let G be the mapping class group of a closed orientable surface of genus $g \geqslant 1$, let S be a finite symmetric generating set of G and let (X_k), $k \geqslant 1$, be the simple left-invariant random walk on G. Then the set $X \subset G$ of non-pseudo-Anosov elements is transient for this random walk.*

Proof. We follow the arguments of Rivin. First of all, the mapping class group G may be defined as the group of diffeomorphisms of a fixed compact connected surface Σ_g of genus g preserving the orientation, up to isotopy (i.e., homotopy in the diffeomorphism group). The main point is that the induced action on the integral homology $H_1(\Sigma_g, \mathbf{Z})$, which preserves the intersection pairing, yields a surjective map

$$\rho \,:\, G \rightarrow Sp(2g, \mathbf{Z}).$$

Let S be a generating set as above[20], and let $S' = \rho(S)$, a finite symmetric generating set for $Sp(2g, \mathbf{Z})$. The image $Y_k = \rho(X_k)$ of the random walk on G is a random walk on $Sp(2g, \mathbf{Z})$.

Note that the steps ξ_k are independent and identically distributed, but is not necessarily true that each ξ_k is uniformly distributed on S', which means that we are not exactly in the setting of Theorem 9.4. However, we can easily prove the analogue of Proposition 9.2 for any random walk on $Sp(2g, \mathbf{Z})$ defined by identically distributed independent steps ξ_k with the property that $\mathbf{P}(\xi_k = s') = p(s') > 0$ for all $s' \in S'$, simply by replacing the self-adjoint operator M with

$$M = \sum_{s \in S'} p(s')\pi(s'),$$

and using the identities

$$\sum_{s' \in S'} p(s')\|\pi(s')v \pm v\|^2 = 2\langle (\mathrm{Id} \pm M)v, v\rangle$$

to obtain the bounds

$$\frac{\langle (\mathrm{Id} - M)v, v\rangle}{\langle v, v\rangle} \geqslant \frac{\min p(s')}{2} \inf_{\varpi} \inf_{v \neq 0} \max_{s \in S} \frac{\|\varpi(s)v - v\|^2}{\|v\|^2},$$

and

$$\|v\|^2 \leqslant \frac{c^2}{4} \frac{1}{\min p(s')} \sum_{s' \in S'} p(s')\|v + \pi(s')v\|^2 = \frac{c^2}{2 \min p(s')} \langle (\mathrm{Id} + M)v, v\rangle$$

analogues of (9.3) and (9.5). From this, sieve bounds for the random walk (Y_k) follow, comparable to those for the simple random walk.

Now, we need only use the fact (the "homological criterion for pseudo-Anosov diffeomorphism") that it *suffices* that the following three conditions on the characteristic polynomial $P = \det(T - \rho(X_k))$ hold for X_k to be pseudo-Anosov:

– P is irreducible;
– there is no root of unity which is a zero of P;
– there is no $d \geqslant 2$ and polynomial Q such that $P(X) = Q(X^d)$.

Accordingly we have

$$\mathbf{P}(X_k \text{ is not pseudo-Anosov}) \leqslant p_1 + p_2 + p_3$$

where p_1, p_2, p_3 are the probabilities that $\det(T - \rho(X_k))$ satisfy those three conditions. Assume first $g \geqslant 2$. Then, according to (2) of Theorem 9.4 (adapted to a non-simple random walk), there exists $\alpha > 0$ such that

$$p_1 \ll \exp(-\alpha k)$$

[20] The existence of such finite generating set is not obvious, of course, and is known as the Dehn-Lickorish Theorem; see e.g. [Iv, Th. 4.2.D].

for $k \geqslant 1$. To estimate p_2 and p_3 we can use simpler sieves to obtain comparable bounds. For p_2, since P is an integral polynomial of degree $2g$ and hence may have only finitely many roots of unity as zeros, we need only estimate the sifted sets for the sifting sets

$$\Omega_\ell = \{g \in Sp(2g, \mathbf{F}_\ell) \mid (\Phi_d \,(\mathrm{mod}\, \ell)) \nmid \det(T - g)\}$$

where $\Phi_d \in \mathbf{Z}[X]$ is the d-th cyclotomic polynomial for some fixed $d \geqslant 1$. It is clear (by the same local density arguments of Appendix B that were already used) that $|\Omega_\ell| \gg |Sp(2g, \mathbf{F}_\ell)|$, and hence the sieve again yields $p_2 \ll \exp(-\alpha k)$ for $k \geqslant 1$.

For p_3, we consider similarly

$$\Omega'_\ell = \{g \in Sp(2g, \mathbf{F}_\ell) \mid \det(T - g) \text{ is not of the form } Q(X^d)\}$$

for some fixed $d \geqslant 2$. We also trivially have $|\Omega'_\ell| \gg |Sp(2g, \mathbf{F}_\ell)|$, and $p_3 \ll \exp(-\alpha k)$.

Now we conclude that $\mathbf{P}(X_k$ is not pseudo-Anosov$) \ll \exp(-\alpha k)$, and we can again apply the Borel-Cantelli lemma. □

Remark 9.8. Maher [Ma] proved that the probability that X_k is pseudo-Anosov tends to 1 as $k \to +\infty$ using rather more of the geometry and structure of the mapping class group, and asked about the possible transience of the set of non-pseudo-Anosov elements. However, his methods are also more general, and work for random walks on any subgroup of G that is not "too small" in some sense. It should be emphasized that his condition encompasses groups which seem utterly inapproachable by sieve as above, in particular the so-called Torelli group which is the kernel of the homology action ρ above. It may seem surprising that pseudo-Anosov should exist in this subgroup, but Maher's result shows that they remain "generic" (see [FLP, p. 250] for a construction which gives some examples, and the observation that Nielsen had conjectured they did not exist). It would be interesting to know if the random walk on the Torelli group is still transient on the set of pseudo-Anosov elements.

Remark 9.9. In the most classical sieves, estimating either the analogue of $R(\mathcal{L})$ or H is not a significant part of the work, the latter because once Ω_ℓ is known, which is usually not a problem there, it boils down to estimates for sums of multiplicative functions, which are well understood.

The results we have proved, and an examination of Appendix B show that when performing a sieve in some group setting, sharp estimates for $R(\mathcal{L})$ or for H involve deeper tools. For the large sieve constant, this involves the representation theory of the group in non-trivial ways. For H, the issue of estimating $|\Omega_\ell|$ may quickly become a difficult counting problem over finite fields. It is not hard to envision situations where the full force of Deligne's work on exponential sums over finite fields becomes really crucial, and not merely a convenience.

Note that the use of the sharp upper bounds of Proposition 7.3 in the proof of Theorem 9.4 is not necessary if one wishes merely to find a bound for the large sieve constant of the type $|W| + |W|^{1-\alpha} L^A$ for some $\alpha > 0$ and $A \geqslant 0$: trivial bounds for $A_p(G)$ are sufficient.

If no exact value of the (τ)-constant for $G = \mathbf{G}(\mathbf{Z})$ and S is known, the value of α coming from Proposition 9.2 is not explicit, so knowing a specific value of A is not particularly rewarding. However, in some cases explicit Kazhdan, hence (τ), constants are known for the groups we are considering. The question of such explicit bounds was first raised by Serre, de la Harpe and Valette; the arguments above show it is indeed a very natural question with concrete applications, such as explicit sieve bounds. The first results for arithmetic groups are due to Burger for $SL(3, \mathbf{Z})$ (see [HV, Appendice]).

To give an idea, we quote a result of Kassabov [Ka], improving an earlier one of Shalom [Sha, Cor. 1]: let $G = SL(n, \mathbf{Z})$ with $n \geqslant 3$, and let S be the symmetric generating set (of $2(n^2 - n)$ elements) of elementary matrices $E_{i,j}(\pm 1)$ with ± 1 in the (i, j)-th entry. Then, for any unitary representation π of G not containing the trivial representation, and any non-zero vector v in the space of π, there exists $s \in S$ such that $\|\pi(s)v - v\| \geqslant \varepsilon_n \|v\|$, with $\varepsilon_n = (42\sqrt{n} + 920)^{-1}$.

The standard commutator relation

$$E_{1,2}(1)[E_{1,3}(1), E_{3,2}(1)]^{-1} = 1$$

(which uses that $n \geqslant 3$...) shows that there are relations of odd length $\leqslant 5$ in terms of S. Looking at the proof of Proposition 9.2, we see that we can take $\delta_+ = \frac{1}{2}\varepsilon_n^2|S|^{-1}$ and $\delta_- = \frac{1}{25(n^2-n)} > \delta_+$. This means that for this particular generating set, α in Theorem 9.4 can be taken to satisfy

$$\alpha = -\frac{\log(1-\varepsilon_n^2|S|^{-1})}{\log|S|} \geqslant \frac{1}{8(n^2-n)(21\sqrt{n}+460)^2\log(2(n^2-n))}$$

for $n \geqslant 3$. So we have

$$|\{w \in W \mid \det(F_w - T) \text{ is reducible}\}| \ll |W|^{1-\eta}$$

for $N \geqslant 1$, the implied constant depending on n, with η given by

$$\eta = \frac{\alpha}{n^2} \geqslant \frac{1}{8n^3(n-1)(21\sqrt{n}+460)^2\log(2(n^2-n))} \gg \frac{1}{n^5\log n}.$$

Coming back to the probabilistic interpretation (which is more suited to what follows), this means in particular that if k is of order of magnitude larger than $n^5\log n$, the probability that $\det(X_k - T)$ is irreducible becomes close to 1. It would be interesting to have a more precise knowledge of this "transition time"

$$\tau_n = \min\{k \geqslant 1 \mid \det(X_k - T) \text{ is irreducible}\},$$

(which, of course, depends also on S).

Note that, at the very least, with this particular generating set, $\det(X_k - T)$ is reducible for $k \leqslant t_n$ where t_n is the first time when all basis vectors have been moved at least once. Since multiplying by ξ_k means moving one of the n basis vectors chosen uniformly, t_n is the stopping time for the "coupon collector problem". Besides the obvious bound $t_n \geqslant n$, it is well-known (see, e.g., [F, IX.3.d]) that

$$\mathbf{E}(t_n) = n(\log n + \gamma) + O(1), \text{ for } n \geqslant 1, \quad \mathbf{V}(t_n) \sim \zeta(2)n^2 \text{ as } n \to +\infty.$$

The gap between upper and lower bounds for τ_n is quite large, and numerical experiments strongly suggest that the lower bound is closer to the truth (in fact, it suggests that $\mathbf{E}(\tau)$ might be $\sim c\mathbf{E}(t_n)$ for some constant $c > 1$ as $n \to +\infty$). In terms of possible improvements, it is interesting to note that the order of magnitude of Kassabov's estimate of the Kazhdan constant ε_n for this generating set S is optimal, since Zuk has pointed out that it must be $\geqslant \sqrt{2/n}$ at least (see [Sha, p. 149]).

Remark 9.10. If $G = SL(2,\mathbf{Z})$, although G does not have Property (T), it is still true that Property (τ) holds for the congruence subgroups $\ker(SL(2,\mathbf{Z}) \to SL(2,\mathbf{Z}/d\mathbf{Z}))$, by Selberg's $\lambda_1 \geqslant 1/4$ theorem on the smallest eigenvalue of the hyperbolic laplacian acting on congruence subgroups of $SL(2,\mathbf{Z})$. However, the second condition of Lemma 9.5 is not true. Indeed, there is a well-known homomorphism

$$SL(2,\mathbf{Z}) \to SL(2,\mathbf{F}_2) \simeq \mathfrak{S}_3 \xrightarrow{\varepsilon} \{\pm 1\}$$

(where the isomorphism in the middle is obtained by looking at the action on the three lines in $\mathbf{F}_2^2$, and ε is the signature), and (for instance) the generators

$$S = \left\{\begin{pmatrix} 1 & \pm 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \pm 1 & 1 \end{pmatrix}\right\}$$

(which are the analogues of the generating set for $SL(n,\mathbf{Z})$ considered in the previous remark) all map to transpositions in $\mathfrak{S}_3$. So $\varepsilon(r) = -1$ for any word of odd length in the alphabet S.

Still, while this shows that Proposition 9.2 can not be applied, it remains true that for an arbitrary symmetric set of generators S of $SL(2,\mathbf{Z})$ and for odd primes ℓ, the Cayley graph of $SL(2,\mathbf{F}_\ell)$ with respect to S is not bipartite (because any homomorphism $SL(2,\mathbf{F}_\ell) \to \{\pm 1\}$ is still trivial for $\ell \geqslant 3$). Hence this Cayley graph contains some cycle of odd length, which is easily checked to be $\leqslant 2d(\ell)+1$, where $d(\ell)$ is the diameter of the Cayley graph.[21] Since we have an expander family (by Property (τ)), there is a bound

$$d(\ell) \ll \log \ell$$

for $\ell \geqslant 3$, where the implied constant depends only on S (since the (τ)-constant, hence the expanding ratio, is fixed); see, e.g., [SC, §6.4]. After a look at the character table of $SL(2,\mathbf{F}_\ell)$, it is not difficult to check that this leads to sieve constants such that

$$\Delta \leqslant |W| + |W| \exp\Bigl(-\frac{cN}{\log L}\Bigr)\Bigl\{\max_{n\in\mathcal{L}} \psi(n) \sum_{m\in\mathcal{L}} \psi(m)^2\Bigr\}$$

where $L = \max \mathcal{L}$ and $c > 0$ depends only on S (see (11.7) for the definition of $\psi(m)$).

For the problem of irreducibility (which is not the most interesting question about quadratic polynomials, perhaps...) taking $\mathcal{L}$ the odd primes $\leqslant L$, this leads to

$$|\{w \in W \mid \det(F_w - T) \in \mathbf{Z}[T] \text{ is reducible}\}| \ll |W| \exp(-c'\sqrt{N}) \tag{9.11}$$

where $c' > 0$ depends only on S, and as observed already, this proves Corollary 9.6 and Theorem 1.2 for $SL(2,\mathbf{Z})$.

Remark 9.11. The work of Bourgain, Gamburd and Sarnak (see [BGS] and Sarnak's slides [Sa3]) is based on another type of sieve settings, which amounts to the following. First, we have a finitely generated group Γ which is a discrete subgroup of a matrix group over $\mathbf{Z}$, acting on an affine algebraic variety $V/\mathbf{Z}$. Then the sieve setting is $(\Gamma \cdot v, \{\text{primes}\}, \rho_\ell)$ where $\Gamma \cdot v$ is the orbit of a fixed element $v \in V(\mathbf{Z})$, and ρ_ℓ is the reduction map to the finite orbit of the reduction in $V(\mathbf{F}_\ell)$ (with uniform density). The siftable set if a subset Y of the orbit defined by the images of elements of Γ of bounded word-length or bounded norm, with counting measure and identity map.

Remark 9.12. If we consider an abstract finitely generated group Γ, and wish to investigate by sieve methods some of its properties, the family of reduction maps modulo primes makes no sense. We want to point out a family (ρ_ℓ) that may be of use, inasmuch as it satisfies the linear disjointness condition (Definition 2.3).

Let $\tilde{\Lambda}$ be the set of surjective homomorphisms

$$\rho : G \rightarrow \rho(G) = H$$

where H is a non-abelian finite simple group, and let $\Lambda \subset \tilde{\Lambda}$ be a set of representatives for the equivalence relation $\rho_1 \sim \rho_2$ if and only if there exists an isomorphism $\rho_1(G) \rightarrow \rho_2(G)$ such that the triangle

$$\begin{tikzcd} & G \arrow[dl, "\rho_1"'] \arrow[dr, "\rho_2"] & \\ \rho_1(G) \arrow[rr] & & \rho_2(G) \end{tikzcd} \tag{9.12}$$

commutes.

Lemma 9.13. *The system $(\rho)_{\rho\in\Lambda}$ constructed in this manner is linearly disjoint.*

This is an easy adaptation of classical variants of the Goursat-Ribet lemmas, and is left as an exercise (see e.g. [Ri, Lemma 3.3]).

[21] Fix some vertex x and find two vertices y and z which are neighbors but satisfy $d(x,y) \equiv d(x,z) \pmod 2$ (those exist, because otherwise the graph would be bipartite; note that $d(x,y) = d(x,z)$); then follow a path γ_1 of length $d(x,y) \leqslant d(\ell)$ from x to y, take the edge from y to z, then follow a path of length $d(z,x) = d(x,y)$ from z to x to obtain a loop of length $2d(x,y)+1 \leqslant 2d(\ell)+1$. The example of a cycle of odd length, i.e., of the Cayley graph of $\mathbf{Z}/m\mathbf{Z}$ with m odd with respect to $S = \{\pm 1\}$ shows that this is best possible for arbitrary graphs, and the Ramanujan graphs of Lubotzky-Philips-Sarnak give examples of expanding families where diameter and length of shortest loop (not necessarily of odd length) are of the same order of magnitude, see [Sa2, Th. 3.3.1].

To make an efficient sieve, it would be necessary, in practice, to have some knowledge of Λ, such as the distribution of the orders of the finite simple quotient groups of G. This is of course in itself an interesting problem (see, e.g., the book [LS]).

10. The elliptic sieve

The next application is also apparently new, although it concerns a sieve which is a sort of "twisted" version of the classical large sieve. Let $E/\mathbf{Q}$ be an elliptic curve given by a Weierstrass equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6, \qquad \text{where } a_i \in \mathbf{Z}.$$

Assuming the rank of E is positive, let Λ_E be the set of primes ℓ of good reduction, and for $\ell \in \Lambda_E$, let $\rho_\ell : E(\mathbf{Q}) \to E(\mathbf{F}_\ell)$ be the reduction map.

The natural sets $X \subset E(\mathbf{Q})$ for sieving are the finite sets of rational points $x \in E(\mathbf{Q})$ with (canonical or naïve) height $h(x) \leqslant T$ for some $T \geqslant 0$ (with again counting measure and $F_x = x$ for $x \in X$). There are interesting potential applications of such sieves, because of the following interpretation: a rational point $x = (r, s) \in E(\mathbf{Q})$ (in affine coordinates, so x is non-zero in $E(\mathbf{Q})$) maps to a non-zero point $E(\mathbf{F}_\ell)$ if and only if ℓ does not divide the denominator of the affine coordinates r and s of the point. This shows that integral or S-integral points (in the affine model above) appear naturally as (subsets of) sifted sets.

We use such ideas to prove Theorem 1.4, showing that most rational points have denominators divisible by many (small) primes. Recall that $\omega_E(x)$ is the number of primes, without multiplicity, dividing the denominator of the coordinates of x, with $\omega_E(0) = +\infty$. We also recall the statement:

Theorem 10.1. *Let $E/\mathbf{Q}$ be an elliptic curve with rank $r \geqslant 1$. Then we have*

$$|\{x \in E(\mathbf{Q}) \mid h(x) \leqslant T\}| \sim c_E T^{r/2} \tag{10.1}$$

as $T \to +\infty$, for some constant $c_E > 0$, and moreover for any fixed real number κ with $0 < \kappa < 1$, we have

$$|\{x \in E(\mathbf{Q}) \mid h(x) \leqslant T \text{ and } \omega_E(x) < \kappa \log\log T\}| \ll T^{r/2}(\log\log T)^{-1},$$

for $T \geqslant 3$, where the implied constant depends only on E and κ.

Proof. Let $M \simeq \mathbf{Z}^r$ be a subgroup of $E(\mathbf{Q})$ such that

$$E(\mathbf{Q}) = M \oplus E(\mathbf{Q})_{tors},$$

and let $(x_1, \ldots, x_r)$ be a fixed $\mathbf{Z}$-basis of M. Moreover, let M' be the group generated by $(x_2, \ldots, x_r)$. We will in fact perform sieving only on "lines" directed by x_1.

But first of all, since the canonical height is a positive definite quadratic form on $E(\mathbf{Q}) \otimes \mathbf{R} = M \otimes \mathbf{R}$, the asymptotic formula (10.1) is clear: it amounts to nothing else but counting integral points in $M \otimes \mathbf{R} \simeq \mathbf{R}^r$ with norm $\sqrt{h(x)} \leqslant \sqrt{T}$ (this being repeated as many times as there are torsion cosets).

Moreover, we may (for convenience) measure the size of elements in $E(\mathbf{Q})$ using the squared L^∞-norm

$$\|x\|_\infty^2 = \max |a_i|^2, \qquad \text{for } x = \sum a_i x_i + t \text{ with } t \in E(\mathbf{Q})_{tors},$$

i.e., we have $h(x) \asymp \|x\|_\infty^2$ for all $x \in M$, the implied constants depending only on E.

Now we claim the following:

Lemma 10.2. *For any fixed $\kappa \in]0, 1[$, any fixed $x' \in M'$, any fixed torsion point $t \in E(\mathbf{Q})_{tors}$, we have*

$$|\{x \in (t + x') \oplus \mathbf{Z}x_1 \mid \|x\|_\infty^2 \leqslant T \text{ and } \omega_E(x) < \kappa \log\log T\}| \ll \sqrt{T}(\log\log T)^{-1},$$

for $T \geqslant 3$, the implied constant depending only on E, κ and x_1, but not on x' or t.

Taking this for granted, we conclude immediately that

$$|\{x \in E(\mathbf{Q}) \mid h(x) \leqslant T \text{ and } \omega_E(x) < \kappa \log\log T\}| \ll T^{r/2}(\log\log T)^{-1},$$

by summing the inequality of the lemma over all $x' \in M'$ with $\|x'\|_\infty^2 \leqslant T$ and over all $t \in E(\mathbf{Q})_{tors}$ (the number of pairs (t, x') is $\ll T^{(r-1)/2}$), the implied constant depending only on E and the choice of basis of M.

Next we come to the proof of this lemma. Fix $x' \in M'$, $t \in E(\mathbf{Q})_{tors}$. The left-hand side of the lemma being zero unless $\|t + x'\|_\infty^2 \leqslant T$, we assume that this is the case. We will use the following group sieve setting:

$$\Psi = (\mathbf{Z}x_1, \Lambda_E, \mathbf{Z}x_1 \to \rho_\ell(\mathbf{Z}x_1) \subset \rho_\ell(E(\mathbf{Q})))$$
$$X = \{mx_1 \in G \mid \|t + x' + mx_1\|_\infty^2 = m^2 \leqslant T\}, \qquad F_x = x.$$

For any prime $\ell \in \Lambda_E$, the finite group G_ℓ is a quotient of $\mathbf{Z}x_1$ and is isomorphic to $\mathbf{Z}/\nu(\ell)\mathbf{Z}$ where $\nu(\ell)$ is the order of the reduction of x_1 modulo ℓ. (So this sieve is really an ordinary-looking one for integers, except for the use of reductions modulo $\nu(\ell)$ instead of reductions modulo primes).

Lemma 10.3. *Let x_1 be an infinite order point on $E(\mathbf{Q})$, and $\nu(\ell)$ the order of x_1 modulo ℓ. Then all but finitely many primes p occur as the value of $\nu(\ell)$ for some ℓ of good reduction.*

Proof. For a prime p, consider $px_1 \in E(\mathbf{Q})$. A prime ℓ of good reduction divides the denominator of the coordinates of px_1 if and only if $p \equiv 0 \pmod{\nu(\ell)}$, which means that $\nu(\ell)$ is either 1 or p. So if p is *not* of the form $\nu(\ell)$, it follows that px_1 is an S-integral point, where S is the union of the set of primes of bad reduction and the finite set of primes where $x_1 \equiv 0 \pmod{\ell}$. By Siegel's finiteness theorem (see, e.g., [Si3, Th. IX.4.3]), there are only finitely many S-integral points in $E(\mathbf{Q})$, and therefore only finitely many p for which p is not of the form $\nu(\ell)$. $\square$

(Note that this lemma is also a trivial consequence of a result of Silverman [Si1, Prop. 10] according to which all but finitely many *integers* are of the form $\nu(\ell)$ for some ℓ; in fact Silverman's result depends on a stronger form of Siegel's theorem).

The lemma allows us to sieve X using as prime sieve support $\mathcal{L}^*$ the set of $\ell \in \Lambda_E$ is such that $\nu(\ell)$ is a prime number $p \leqslant L$ (where, in case the same prime p occurs as values of $\nu(\ell)$ for two or more primes, we keep only one), and with $\mathcal{L} = \mathcal{L}^*$ (with the usual identification of elements with singletons in $S(\Lambda)$).

From the lemma, it follows that the inequality defining the large sieve constant here, namely

$$\sum_{\ell \in \mathcal{L}} \sideset{}{^*}\sum_{a \,(\mathrm{mod}\, \nu(\ell))} \Big| \sum_{|m| \leqslant \sqrt{T}} \alpha(m) e\Big(\frac{am}{\nu(\ell)}\Big)\Big|^2 \leqslant \Delta \sum_{|m| \leqslant \sqrt{T}} |\alpha(m)|^2, \tag{10.2}$$

for all $(\alpha(m))$, can be reformulated as

$$\sideset{}{^*}\sum_{p \leqslant L} \sideset{}{^*}\sum_{a \,(\mathrm{mod}\, p)} \Big| \sum_{|m| \leqslant \sqrt{T}} \alpha(m) e\Big(\frac{am}{p}\Big)\Big|^2 \leqslant \Delta \sum_{|m| \leqslant \sqrt{T}} |\alpha(m)|^2.$$

where $\sum^*$ in the sum over p indicates that only those p which occur as $\nu(\ell)$ for some ℓ are taken into account. We recognize the most standard large sieve inequality, and by positivity, it follows that

$$\Delta \leqslant 2\sqrt{T} + L^2$$

for $L \geqslant 2$. We now apply Proposition 3.1: we have

$$\sum_{x \in X} \Big(P(x, \mathcal{L}) - P(\mathcal{L})\Big)^2 \leqslant \Delta Q(\mathcal{L}) \tag{10.3}$$

where $P(x, \mathcal{L})$, $P(\mathcal{L})$ and $Q(\mathcal{L})$ are defined in (3.2), for any given choice of sets $\Omega_\ell \subset G_\ell$ for $\ell \in \Lambda_E$.

We let $\Omega_\ell = \{-\rho_\ell(t+x')\}$. By the remark before the statement of Theorem 10.1, we have $\rho_\ell(mx_1) \in \Omega_\ell$ if and only if ℓ divides the denominator of the coordinates of $t+x'+mx_1$, and therefore for $x = mx_1 \in X$, we have

$$P(mx_1, \mathcal{L}) \leqslant \omega_E(t+x'+mx_1).$$

On the other hand, we have

$$P(\mathcal{L}) = \sum_{\ell\in\mathcal{L}} \frac{1}{|G_\ell|} = \sum_{\ell\in\mathcal{L}} \frac{1}{\nu(\ell)} = \sum_{p\leqslant L} \frac{1}{p} + O(1) = \log\log L + O(1)$$

for any $L \geqslant 3$, because, by Lemma 10.3, the values $\nu(\ell) \leqslant L$ range over all primes $\leqslant L$, with only finitely many exceptions (independently of L).

Hence there exists L_0 depending on E, x_1 and κ only, such that if $L \geqslant L_0$, we have

$$P(\mathcal{L}) \geqslant \frac{1+\kappa}{2} \log\log L.$$

Putting together these two inequalities, we see that if we assume $T \leqslant L^2$, say, and $L \geqslant L_0'$ for some other constant L_0' (depending on E, x_1 and κ), then for any $mx_1 \in X$ such that $t+x'+mx_1$ satisfies $\omega_E(t+x'+mx_1) < \kappa \log\log T$, we have

$$\Bigl(P(x,\mathcal{L}) - P(\mathcal{L})\Bigr)^2 \gg (\log\log T)^2,$$

the implied constant depending only on E, x_1 and κ. So it follows by positivity from (10.3) and the inequality $Q(\mathcal{L}) \leqslant P(\mathcal{L}) \ll \log\log T$ that

$$\begin{aligned}|\{x \in t+x' \oplus \mathbf{Z}x_1 \mid \|x\|_\infty^2 \leqslant T \text{ and } \omega_E(x) < \kappa\log\log T\}| &\ll \Delta(\log\log T)^{-1} \\ &\ll (\sqrt{T} + L^2)(\log\log T)^{-1}\end{aligned}$$

for any $L \geqslant L_0'$. If $T^{1/2} \geqslant L_0'$, we take $L = T^{1/2}$ and prove the inequality of the lemma directly, and otherwise we need only increase the resulting implied constant to make it valid for all $T \geqslant 3$, since L_0' depends only on E, x_1 and κ. □

It would be interesting to know whether there is some "regular" distribution for the function $\omega_E(x)$. Notice the similarity between the above discussion and the Hardy-Ramanujan results concerning the normal order of the number of prime divisors of an integer (see e.g. [HW, 22.11]), but note that since the denominators of rational points x are typically of size $\exp h(x)$, they should have around

$$\log\log\exp(h(x)) = \log(h(x))$$

prime divisors in order to be "typical" integers. However, note also that the prime divisors accounted for in the proof above are all $\leqslant T^{1/2} \simeq \sqrt{h(x)} \simeq \sqrt{\log n}$; it *is* typical behavior for an integer $n \leqslant T$ to have roughly $\log\log\log T$ prime divisors of this size (much more precise results along those lines are known, due in particular to Turán, Erdös and Kac).

Note also that, as mentioned during the discussion of Proposition 3.1, applying the (apparently stronger) form of the large sieve involving squarefree numbers would only give a bound for the number of points which are $\mathcal{L}$-integral. Since (for any finite set S), there are only finitely many S-integral points, and moreover this is used in the proof of Lemma 10.3, this would not be a very interesting conclusion.

We can relate this sieve, more precisely Lemma 10.2, to so-called *elliptic divisibility sequences*, a notion introduced by M. Ward and currently the subject of a number of investigations by Silverman, T. Ward, Everest, and others (see e.g. [Si2], [W], [EEW]). This shows that the proposition above has very concrete interpretations.

Proposition 10.4. *Let* $(W_n)_{n\geqslant 0}$ *be an* unbounded *sequence of integers such that*

$$W_0=0, \quad W_1=1, \quad W_2W_3\neq 0, \quad W_2 \mid W_4$$
$$W_{m+n}W_{m-n}=W_{m+1}W_{m-1}W_n^2-W_{n+1}W_{n-1}W_m^2, \quad \textit{for } m\geqslant n\geqslant 1,$$
$$\Delta = W_4W_2^{15}-W_3^3W_2^{12}+3W_4^2W_2^{10}-20W_4W_3^3W_2^7$$
$$+4W_4^3W_2^5+16W_3^6W_2^4+8W_4^2W_3^3W_2^2+W_4^4\neq 0.$$

Then for any κ *such that* $0<\kappa<1$*, we have*

$$|\{n\leqslant N \mid \omega(W_n)<\kappa\log\log N\}| \ll \frac{N}{\log\log N}$$

for $N\geqslant 3$*, where the implied constant depends only on* κ *and* (W_n)*.*

Proof. This depends on the relation between elliptic divisibility sequences and pairs (E,x_1) of an elliptic curve $E/\mathbf{Q}$ and a point $x_1\in E(\mathbf{Q})$. Precisely (see e.g. [EEW, §2]) there exists such a pair (E,x_1) with x_1 of infinite order such that if (a_n), (b_n), (d_n) are the (unique) sequences of integers with $d_n\geqslant 1$, $(a_n,d_n)=(b_n,d_n)=1$ and

$$nx_1=\Bigl(\frac{a_n}{d_n^2},\frac{b_n}{d_n^3}\Bigr),$$

then we have

$$d_n \mid W_n \text{ for } n\geqslant 1$$

(without the condition $\Delta=0$, this is still true provided *singular* elliptic curves are permitted; the condition that (W_n) be unbounded implies that x_1 is of infinite order).

Now the d_n are precisely the denominators of the coordinates of the points in $\mathbf{Z}x_1$, and we have therefore

$$\omega(W_n)\geqslant\omega(d_n)=\omega_E(nx_1).$$

Hence Lemma 10.2 gives the desired result. □

The "simplest" example is the sequence (W_n) given by

$$W_0=0, \quad W_1=1, \quad W_2=1, \quad W_3=-1, \quad W_4=1,$$
$$W_n=\frac{W_{n-1}W_{n-3}+W_{n-2}^2}{W_{n-4}}, \qquad \text{for } n\geqslant 4$$

(sequence `A006769` in the Online Encyclopedia of Integer Sequences), which corresponds to case of $E: y^2-y=x^3-x$ and $x_0=(0,0)$.

Finally, it will be noticed that the same reasoning and similar results hold for elements of non-degenerate divisibility sequences (u_n) defined by linear recurrence relations of order 2, e.g., $u_n=a^n-1$ where $a\geqslant 2$ is an integer. (The analogue of Silverman's theorem here is a result of Schinzel, and the rest is easy).

11. Sieving for Frobenius over finite fields

The final example of large sieve we discuss concerns the distribution of geometric Frobenius conjugacy classes in finite monodromy groups, refining the arguments and methods in [Ko1]. It is a good example of a coset sieve as in Section 6.

The precise setting is as follows (see also [Ko1]). Let q be a power of a prime p, let $U/\mathbf{F}_q$ be a smooth affine geometrically connected algebraic variety of dimension $d\geqslant 1$ over $\mathbf{F}_q$. Put $\bar{U}=U\times\bar{\mathbf{F}}_q$, the extension of scalars to an algebraic closure of $\mathbf{F}_q$.

Let $\bar{\eta}$ denote a geometric generic point of U. We consider the coset sieve with

(11.1) $$G=\pi_1(U,\bar{\eta}), \quad G^g=\pi_1(\bar{U},\bar{\eta}), \quad G/G^g\simeq \mathrm{Gal}(\bar{\mathbf{F}}_q/\mathbf{F}_q)\simeq\hat{\mathbf{Z}},$$

so that we have the exact sequence

$$1\to G^g\to G\xrightarrow{d}\hat{\mathbf{Z}}\to 1,$$

the last arrow being the "degree".

We assume given a family of representations

$$\rho_\ell \,:\, \pi_1(U,\bar\eta) \to GL(r,k_\ell)$$

for ℓ in a subset Λ of the set of prime numbers, where k_ℓ is a finite field of characteristic ℓ and r is independent of ℓ. By the equivalence of categories between lisse sheaves of k-modules and continuous actions of $\pi_1(U,\bar\eta)$ on finite dimensional k vector spaces, this corresponds equivalently to a system $(\mathcal{F}_\ell)$ of étale k_ℓ-vector spaces. We then put $G_\ell = \mathrm{Im}(\rho_\ell)$, the *arithmetic monodromy group* of ρ_ℓ, so that we have surjective maps $G = \pi_1(U,\bar\eta) \to G_\ell$.

The siftable set we are interested in is given by $X = U(\mathbf{F}_q)$, with counting measure, with the map $x \mapsto F_x \in G_\ell^\sharp$ given by the geometric Frobenius conjugacy class at the rational point $x \in U(\mathbf{F}_q)$ (relative to the field $\mathbf{F}_q$). Since, in the exact sequence above, we have $d(F_x) = -1 \in \hat{\mathbf{Z}}$ for all $x \in U(\mathbf{F}_q)$, this corresponds to the sieve setting $(Y,\Lambda,(\rho_\ell))$ where Y, as in Section 6, is the set of conjugacy classes in $\pi_1(U,\bar\eta)$ with degree -1.

Then, concerning the exponential sums of Proposition 6.3, we have two basic bounds.

Proposition 11.1. *Assume that the representations (ρ_m) for $m \in S(\Lambda)$ are such that, for all squarefree numbers m divisible only by primes in Λ, the map*

$$\pi_1(\bar U,\bar\eta) \to G_m^g = \prod_{\ell\mid m} G_\ell^g$$

is onto. With notation as before and as in Proposition 6.3, we have:

(1) *If G_ℓ is a group of order prime to p for all $\ell \in \Lambda$, then*

$$W(\pi,\tau) = \delta((m,\pi),(n,\tau))q^d + O\Big(q^{d-1/2}|G_{[m,n]}|(\dim\pi)(\dim\tau)\Big)$$

for m, $n \in S(\Lambda)$, $\pi \in \Pi_m^*$, $\tau \in \Pi_n^*$, where the implied constant depends only on $\bar U$.

(2) *If $d = 1$ (U is a curve) and if the sheaves $\mathcal{F}_\ell$ are of the form $\mathcal{F}_\ell = \tilde{\mathcal{F}}_\ell/\ell\tilde{\mathcal{F}}_\ell$ for some* compatible family *of torsion-free $\mathbf{Z}_\ell$-adic sheaves $\tilde{\mathcal{F}}_\ell$, then*

$$W(\pi,\tau) = \delta((m,\pi),(n,\tau))q + O\Big(q^{1/2}(\dim\pi)(\dim\tau)\Big)$$

where the implied constant depends only on the compactly-supported Euler-Poincaré characteristics of $\bar U$ and of the compatible system $(\tilde{\mathcal{F}}_\ell)$ on $\bar U$.

Recall that a system $(\tilde{\mathcal{F}}_\ell)$ of étale sheaves of torsion-free $\mathbf{Z}_\ell$-modules is *compatible* if, for every ℓ, every extension field $\mathbf{F}_{q^r}$ of $\mathbf{F}_q$, any $v \in U(\mathbf{F}_{q^r})$, the characteristic polynomial $\det(1-TF_v \mid \tilde{\mathcal{F}}_\ell)$ has integer coefficients and is independent of ℓ. Then the Euler-Poincaré characteristic $\chi_c(\bar U,\tilde{\mathcal{F}}_\ell)$ is independent of ℓ, being the degree of the L-function

$$\prod_{x\in|U|} \det(1-TF_x \mid \tilde{\mathcal{F}}_\ell)^{-1}$$

of the sheaf as a rational function ($|U|$ is the set of all closed points of U).

Proof. This is essentially Proposition 5.1 of [Ko1], in the case $\ell = m$, $\ell' = n$ at least. We repeat the proof since it is quite short.

By (6.5) and the definition of X, we have

$$W(\pi,\tau) = \frac{1}{\sqrt{|\hat\Gamma_m^\pi||\hat\Gamma_n^\tau|}} \sum_{u\in U(\mathbf{F}_q)} \mathrm{Tr}([\pi,\bar\tau]\rho_{[m,n]}(F_u))$$

where the sum is the sum of local traces of Frobenius for a continuous representation of $\pi_1(U,\bar\eta)$. We can view $[\pi,\bar\tau]$ as a representation acting on a $\bar{\mathbf{Q}}_\ell$-vector space for some prime $\ell \neq p$, and then this expression may be interpreted as the sum of local traces of Frobenius at points in $U(\mathbf{F}_q)$ for some lisse $\bar{\mathbf{Q}}_\ell$-adic sheaf $\mathcal{W}(\pi,\tau)$ on U.

By the Grothendieck-Lefschetz Trace Formula (see, e.g., [Gr], [D2], [Mi, VI.13]), we have then

$$W(\pi,\tau)=\frac{1}{\sqrt{|\hat{\Gamma}_m^\pi||\hat{\Gamma}_n^\tau|}}\sum_{i=0}^{2d}(-1)^i\,\mathrm{Tr}(\mathrm{Fr}\mid H_c^i(\overline{U},\mathcal{W}(\pi,\tau)))$$

where Fr denotes the global geometric Frobenius on $\bar{U}$.

Since the representation corresponding to $\mathcal{W}(\pi,\tau)$ factors through a finite group, this sheaf is pointwise pure of weight 0. Therefore, by Deligne's Weil II Theorem [D1, p. 138], the eigenvalues of the geometric Frobenius automorphism Fr acting on $H_c^i(\overline{U},\mathcal{W}(\pi,\tau))$ are algebraic integers, all conjugates of which are of absolute value $\leqslant q^{i/2}$.

This yields

$$W(\pi,\tau)=\frac{1}{\sqrt{|\hat{\Gamma}_m^\pi||\hat{\Gamma}_n^\tau|}}\mathrm{Tr}(\mathrm{Fr}\mid H_c^{2d}(\overline{U},\mathcal{W}(\pi,\tau)))+O\Big(\sigma_c'(\bar{U},\mathcal{W}(\pi,\tau))q^{d-1/2}\Big),$$

with an absolute implied constant, where

$$\sigma_c'(\bar{U},\mathcal{W}(\pi,\tau)))=\sum_{i=0}^{2d-1}\dim H_c^i(\overline{U},\mathcal{W}(\pi,\tau)).$$

For the "main term", we use the formula

$$H_c^{2d}(\bar{U},\mathcal{W}(\pi,\tau))=V_{\pi_1(\bar{U},\bar{\eta})}(-d)$$

where $V=\mathcal{W}_{\overline{\eta}}(\pi,\tau)$ is the space on which the representation which "is" the sheaf acts. But, by assumption, when we factor the representation (restricted to the geometric fundamental group) as follows

$$\pi_1(\bar{U},\bar{\eta})\overset{\rho_{[m,n]}}{\longrightarrow}G_{[m,n]}^g\overset{[\pi,\bar{\tau}]}{\longrightarrow}GL((\dim\pi)(\dim\tau),\bar{\mathbf{Q}}_\ell),$$

the first map is *surjective*. Hence we have

$$V_{\pi_1(\bar{U},\bar{\eta})}(-d)=W_{G_{[m,n]}^g}(-d)$$

with W denoting the space of $[\pi,\bar{\tau}]$. As we are dealing with linear representations of finite groups in characteristic 0, this coinvariant space is the same as the space of invariants, and its dimension is the multiplicity of the trivial representation in $[\pi,\bar{\tau}]$ (acting on W, i.e., restricted to $G_{[m,n]}^g$). By Lemma 6.4, we have therefore

$$H_c^{2d}(\bar{U},\mathcal{W}(\pi,\tau))=0$$

if $(m,\pi)\neq(n,\tau)$. Otherwise the dimension is $|\hat{\Gamma}_m^\pi|$ and the Tate twist means the global Frobenius acts on the invariant space by multiplication by q^d (the eigenvalue is exactly q^d, not a root of unity times q^d because in the latter case would correspond to a situation where $[\pi,\bar{\tau}]$ vanishes identically on $Y_{[m,n]}$, which is excluded by the choice of Π_m^*, Π_n^* in Proposition 6.3). This gives

$$W(\pi,\tau)=\delta((m,\pi),(n,\tau))q^d+O\Big(\sigma_c'(\bar{U},\mathcal{W}(\pi,\tau))q^{d-1/2}\Big),$$

with an absolute implied constant.

To conclude, in Case (1), we appeal to Proposition 4.7 of [Ko1], which gives the desired estimate directly. In Case (2), we will apply Proposition 4.1 of [Ko1], but however we argue a bit differently[22]. Namely, we claim that

$$\sigma_c'(\bar{U},\mathcal{W}(\pi,\tau))\leqslant(\dim[\pi,\bar{\tau}])(1-\chi+w|S|),\tag{11.2}$$

where $\chi=\chi_c(\bar{U},\mathbf{Q}_\ell)$, and w is the sum of Swan conductors of $\tilde{\mathcal{W}}_\ell$ at the "points at infinity" $x\in S\subset U(\bar{\mathbf{F}}_q)$, which is independent of ℓ, being equal to

$$\chi\,\mathrm{rank}\,\tilde{\mathcal{W}}_\ell-\chi_c(\bar{U},\tilde{\mathcal{W}}_\ell)$$

[22] The result there yields $\sigma_c'(\bar{U},\mathcal{W}(\pi,\tau))\leqslant(\dim[\pi,\bar{\tau}])(1-\chi+\omega([m,n])w)$ and we do not want the term $\omega([m,n])$, which would lead to a loss of $\log\log L$ below...

where both terms are independent of ℓ. This provides the stated estimate (2).

To check (11.2), we look at the proof of loc. cit. with the current notation, and extract the bound

$$\sigma'_c(\bar{U}, \mathcal{W}(\pi,\tau)) \leqslant (\dim[\pi,\bar{\tau}])\Big(1-\chi+\sum_{x\in S}\max_{\ell\mid[m,n]}\mathrm{Swan}_x(\tilde{\mathcal{W}}_\ell)\Big).$$

Then, by positivity of the Swan conductors, we note that

$$\mathrm{Swan}_x(\tilde{\mathcal{W}}_\ell) \leqslant \sum_{x\in S}\mathrm{Swan}_x(\tilde{\mathcal{W}}_\ell) = w$$

for each x and $\ell \mid [m,n]$ (we use here the compatibility of the system), so that

$$\max_{\ell\mid[m,n]}\mathrm{Swan}_x(\tilde{\mathcal{W}}_\ell) \leqslant w,$$

and

$$\sum_{x\in S}\max_{\ell\mid[m,n]}\mathrm{Swan}_x(\tilde{\mathcal{W}}_\ell) \leqslant w|S|$$

which concludes the proof. □

To apply the bounds for the exponential sums to the estimation of the large sieve constants, we need bounds for the quantities

$$\max_{m,\pi}\Big\{q^d + C(\dim\pi)\sideset{}{^\flat}\sum_{n\leqslant L}|G_{[m,n]}|\sum_{\tau\in\Pi_n^*}(\dim\tau)\Big\} \tag{11.3}$$

in the first case and

$$\max_{m,\pi}\Big\{q^d + C(\dim\pi)\sideset{}{^\flat}\sum_{n\leqslant L}\sum_{\tau\in\Pi_n^*}(\dim\tau)\Big\} \tag{11.4}$$

in the second case.

For this purpose, we make the following assumptions: for all $\ell\in\Lambda$, and $\pi\in\Pi_\ell^*$, we have

$$|G_\ell| \leqslant (\ell+1)^s, \quad \dim\pi \leqslant (\ell+1)^v, \quad \sum_{\pi\in\Pi_\ell^*}(\dim\pi) \leqslant (\ell+1)^t, \tag{11.5}$$

where s, t and v are non-negative integers. In the notation of Section 7, the second and third are implied by

$$A_\infty(G_\ell) \leqslant (\ell+1)^v, \quad A_1(G_\ell) \leqslant (\ell+1)^t$$

respectively.

Here are some examples; the first two are results proved in Section 7 (see Example 7.5).

– if G_ℓ is a subgroup of $GL(r,\mathbf{F}_\ell)$, we can take $s=r^2$, $v=r(r-1)/2$, $t=r(r+1)/2$.

– if G_ℓ is a subgroup of symplectic similitudes for some non-degenerate alternating form of rank $2g$, we can take $s=g(2g+1)+1$, $v=(s-(g+1))/2=g^2$, $t=g^2+g+1$.

– in particular, if $G_\ell \subset GL(2,\mathbf{F}_\ell)$ and $G^g = SL(2,\mathbf{F}_\ell)$, we have

$$|G_\ell| \leqslant \ell^4, \quad \max(\dim\pi) = \ell+1, \quad \sum_{\pi\in\Pi_\ell^*}(\dim\pi) \leqslant (\ell+1)^3. \tag{11.6}$$

This particular case can be checked easily by looking at the character table for $GL(2,\mathbf{F}_\ell)$ and $SL(2,\mathbf{F}_\ell)$. See also the character tables of $GL(3,\mathbf{F}_\ell)$ and $GL(4,\mathbf{F}_\ell)$ in [St] for those cases.

Remark 11.2. In [Ko1], we used different assumptions on the size of the monodromy groups and the degrees of their representations. The crucial feature of (11.5) is that $A_1(G_\ell)$ and $A_\infty(G_\ell)$ are bounded by *monic* polynomials. Having polynomials with constant terms >1 would mean, after multiplicativity is applied, that $A_\infty(G_m)$ and $A_1(G_m)$ would be bounded by polynomials times a divisor function; on average over m, this would mean a loss of a power of logarithm, which in large sieve situation (as above with irreducibility of zeta functions of curves) is likely to overwhelm the saving coming from using squarefree numbers in the sieve. In "small sieve" settings, the loss from divisor functions is reduced to a power of $\log\log|X|$, which may remain reasonable, and may be sufficient justification for using simpler but weaker polynomial bounds.

Let

$$\psi(m)=\prod_{\ell\mid m}(\ell+1). \tag{11.7}$$

It follows by multiplicativity from (11.5) that we have

$$|G_m|\leqslant\psi(m)^s,\quad \dim\pi\leqslant\psi(m)^v,\quad \sum_{\pi\in\Pi_m^*}(\dim\pi)\leqslant\psi(m)^t, \tag{11.8}$$

for all squarefree m.

We wish to sieve with the prime sieve support $\mathcal{L}^*=\{\ell\in\Lambda\mid \ell\leqslant L\}$ for some L. The first idea for the sieve is to use the traditional sieve support $\mathcal{L}_1$ which is the set of squarefree integers $m\leqslant L$ divisible only by primes in Λ. However, since we have $\psi(m)\ll m\log\log m$, and this upper bound is sharp (if m has many small prime factors), the use of $\mathcal{L}_1$ leads to a loss of a power of a power of $\log\log L$ in the second term in the estimation of (11.3) and (11.4). As described by Zywina [Z], this can be recovered using the trick of sieving using only squarefree integers m free of small prime factors, in the sense that $\psi(m)\leqslant L+1$ instead of $m\leqslant L$ (which for primes ℓ remain equivalent with $\ell\leqslant L$). This means we use the sieve support

$$\mathcal{L}=\{m\in S(\Lambda)\mid m\text{ is squarefree and }\psi(m)\leqslant L+1\}.$$

We quote both types of sieves:

Corollary 11.3. *With the above data and notation, let $\Omega_\ell\subset G_\ell$, for all primes $\ell\in\Lambda$, be a conjugacy-invariant subset of G_ℓ such that $d(\Omega_\ell)=-1$. Then we have both*

$$|\{u\in U(\mathbf{F}_q)\mid \rho_\ell(F_u)\notin\Omega_\ell\text{ for }\ell\leqslant L\}|\leqslant(q^d+Cq^{d-1/2}(L+1)^A)H^{-1} \tag{11.9}$$

and

$$|\{u\in U(\mathbf{F}_q)\mid \rho_\ell(F_u)\notin\Omega_\ell\text{ for }\ell\leqslant L\}|\leqslant(q^d+Cq^{d-1/2}L^A(\log\log L)^v)K^{-1}, \tag{11.10}$$

where

$$H=\sideset{}{^\flat}\sum_{\psi(m)\leqslant L+1}\prod_{\ell\mid m}\frac{|\Omega_\ell|}{|G_\ell^g|-|\Omega_\ell|},\quad K=\sideset{}{^\flat}\sum_{m\leqslant L}\prod_{\ell\mid m}\frac{|\Omega_\ell|}{|G_\ell^g|-|\Omega_\ell|},$$

and

(i) *If $p\nmid|G_\ell|$ for all $\ell\in\Lambda$, we can take $A=v+2s+t+1$, and the constant C depends only on $\bar{U}$.*

(ii) *If $d=1$ and the system (ρ_ℓ) arises by reduction of a compatible system of $\mathbf{Z}_\ell$-adic sheaves on U, then we can take $A=t+v+1$, and the constant C depends only on the Euler-Poincaré characteristic of U, the compactly-supported Euler-Poincaré characteristic of the compatible system $(\tilde{\mathcal{W}}_\ell)$ on $\bar{U}$, and on s, t, v in the case of* (11.10).

Proof. From Proposition 2.10, we must estimate

$$\Delta=\max_{m,\pi}\sideset{}{^\flat}\sum_{n}\sum_{\tau\in\Pi_n^*}|W(\pi,\tau)|,$$

where m and n run over $\mathcal{L}$ and $\mathcal{L}_1$, respectively. By Proposition 11.1, this is bounded by the quantities (11.3) and (11.4). Using (11.8), the result is now straightforward, using (in the case of (11.10)) the simple estimate

$$\sideset{}{^\flat}\sum_{n\leqslant L}\psi(n)^A\ll L^{A+1}$$

for $L\geqslant 1$, $A\geqslant 0$, the implied constant depending on A. □

This theorem can be used to get a slight improvement of the "generic irreducibility" results for numerators of zeta functions of curves of [Ko1] (see Section 6 of that paper for some context and in particular Theorem 6.2): a small power of $\log q$ is gained in the upper bound, as in Gallagher's result [G, Th. C]. We only state one special case, for a fixed genus (see the remark following the statement for an explanation of this restriction).

Theorem 11.4. *Let $\mathbf{F}_q$ be a finite field of characteristic $p \neq 2$, let $f \in \mathbf{F}_q[X]$ be a squarefree monic polynomial of degree $2g$, $g \geqslant 1$. For $t \in \mathbf{F}_q$ which is not a zero of f, let $P_t \in \mathbf{Z}[T]$ be the numerator of the zeta function of the smooth projective model of the hyperelliptic curve*

$$C_t \,:\, y^2 = f(x)(x-t), \tag{11.11}$$

and let K_t be the splitting field of P_t over $\mathbf{Q}$, which has degree $[K_t : \mathbf{Q}] \leqslant 2^g g!$. Then we have

$$|\{t \in \mathbf{F}_q \mid f(t) \neq 0 \text{ and } [K_t : \mathbf{Q}] < 2^g g!\}| \ll q^{1-\gamma}(\log q)^{1-\delta}$$

where $\gamma = (4g^2+2g+4)^{-1}$ and $\delta > 0$, with $\delta \sim 1/(4g)$ as $g \to +\infty$. The implied constant depends only on g.

Proof. Let $S(f) \subset \mathbf{F}_q$ be the set in question. Proceeding as in Section 8 of [Ko1], we set up a sieve using the sheaves $\mathcal{F}_\ell = R^1\pi_!\mathbf{F}_\ell$ for $\ell > 2$, π denoting the projection from the family of curves (11.11) to the parameter space $U = \{t \mid f(t) \neq 0\} \subset \mathbf{A}^1$. Those sheaves are tame, obtained by reducing modulo ℓ a compatible system, and the geometric monodromy of $\mathcal{F}_\ell$ is $Sp(2g, \mathbf{F}_\ell)$ by a result of J.K. Yu (which also follows from a recent more general result of C. Hall [H]; see [Ko3] for a write-up of this special case of Hall's result). Using (11.3), and the proof of Proposition 11.1 to bound explicitly the implied constant, we obtain

$$|S(f)| \leqslant (q + 4g\sqrt{q}L^A)H^{-1}$$

where $A = 2g^2+g+2$ (see Example 7.5) and

$$H = \min_{1 \leqslant i \leqslant 4}\Bigl\{ \sideset{}{^\flat}\sum_{\psi(m) \leqslant L} \prod_{\ell \mid m} \frac{|\Omega_{i,\ell}|}{|G_\ell^g| - |\Omega_{i,\ell}|}\Bigr\},$$

the sets $\Omega_{i,\ell}$ being defined as in [Ko1, §7,8]. For each of these we have

$$\frac{|\Omega_{i,\ell}|}{|G_\ell^g| - |\Omega_{i,\ell}|} = \frac{\delta_i}{1-\delta_i} + O\Bigl(\frac{1}{\ell}\Bigr) \tag{11.12}$$

for $\ell \geqslant 3$, for some $\delta_i \in]0,1[$ which is a "density" of conjugacy classes satisfying certain conditions, either in the group of permutations on g letters or the group of signed permutations of $2g$ letters (this follows easily from [G, §2] and Sections 7, 8 of [Ko1]). The implied constant depends only on g. Precisely, we have

$$\delta_1 \sim \frac{1}{2g}, \quad \delta_2 \sim \frac{1}{4g}, \quad \delta_3 \sim \frac{\log 2}{\log g}, \quad \delta_4 \sim \frac{1}{\sqrt{2\pi g}}$$

as $g \to +\infty$ (see [Ko1, §8]).

Thus, we need lower bounds for sums of the type

$$\sideset{}{^\flat}\sum_{\psi(m) \leqslant L} \beta(m) = \sum_{\psi(m) \leqslant L} \beta(m)\mu^2(m)$$

where β is a multiplicative function, roughly constant at the primes. This is a well-studied area of analytic number theory. We can appeal for instance to Theorem 1 of [LW]; in the notation of loc. cit., we have $f(m) = \mu^2(m)\beta(m)$, $g(m) = \psi(m)$, with $\kappa = \delta_i/(1-\delta_i)$, $\eta = 1$, $\alpha = 1$, $\theta = 1$, $\alpha' = 1$, $\theta' = 0$, $t(p) = 0$, $C_3 = 0$. We obtain

$$\sideset{}{^\flat}\sum_{\psi(m) \leqslant L} \beta(m) \gg L(\log L)^{-1+\delta_i/(1-\delta_i)}, \tag{11.13}$$

for $L \geqslant 3$, where the implied constant depends only on g.

Taking $L = q^{1/2A}$, the upper bound for $S(f)$ then follows. $\square$

Remark 11.5. In [Ko1], we obtained a result uniform in terms of g. Here it is certainly possible to do the same, by checking the dependency of the estimate (11.13) on g. However, notice that the gain compared to [Ko1][23] is of size $(\log q)^{\delta}$ with $\delta \sim 1/(4g)$, and this becomes trivial as soon as g is of size $\log\log q$. This is a much smaller range than the (already restricted) range where the estimate of [Ko1] is non-trivial, namely g somewhat smaller than $\sqrt{\log q}$.

Now we prove Theorem 1.5 stated in the introduction.

Proof of Theorem 1.5. We can certainly afford to be rather brief here. The sieve setting and siftable set are the same as in Theorem 11.4. The number of points of C_t and J_t are given by

$$|C_t(\mathbf{F}_q)| = q+1-\mathrm{Tr}(\mathrm{Fr} \mid H^1(\bar{C}_t,\mathbf{Z}_\ell)), \qquad |J_t(\mathbf{F}_q)| = |\det(1-\mathrm{Fr} \mid H^1(\bar{C}_t,\mathbf{Z}_\ell))|,$$

(for any prime $\ell \nmid p$). Thus, defining sieving sets

$$\Omega_\ell^J = \{g \in CSp(2g,\mathbf{F}_\ell) \mid g \text{ is } q\text{-symplectic and } \det(g-1) \text{ is a square in } \mathbf{F}_\ell\},$$
$$\Omega_\ell^C = \{g \in CSp(2g,\mathbf{F}_\ell) \mid g \text{ is } q\text{-symplectic and } q+1-\mathrm{Tr}(g) \text{ is a square in } \mathbf{F}_\ell\},$$

(where q-symplectic similitudes are those with multiplicator q), we have for any prime sieve support $\mathcal{L}^*$ the inclusion

$$\{t \in \mathbf{F}_q \mid f(t) \neq 0 \text{ and } |\mathsf{S}_t(\mathbf{F}_q)| \text{ is a square}\} \subset S(U,\Omega^{\mathsf{S}};\mathcal{L}^*),$$

for $\mathsf{S} \in \{C,J\}$. By (3) and (4), respectively, of Proposition B.1 in Appendix B, we have

$$\frac{|\Omega_\ell^{\mathsf{S}}|}{|Sp(2g,\mathbf{F}_\ell)|} \geqslant \frac{1}{2}\Bigl(\frac{\ell}{\ell+1}\Bigr)^g.$$

for $\ell \geqslant 3$. Thus if $\mathcal{L}$ is the set of odd prime $\leqslant L$, we obtain

$$|\{t \in \mathbf{F}_q \mid f(t) \neq 0 \text{ and } |\mathsf{S}_t(\mathbf{F}_q)| \text{ is a square}\}| \leqslant (q+4g\sqrt{q}L^A)H^{-1}$$

where $A = 2g^2+g+2$, and

$$H = \sum_{3\leqslant \ell \leqslant L} \frac{|\Omega_\ell^{\mathsf{S}}|}{|Sp(2g,\mathbf{F}_\ell)|} \geqslant \frac{1}{2}\sum_{3\leqslant\ell\leqslant L}\Bigl(\frac{\ell}{\ell+1}\Bigr)^g.$$

By the mean-value theorem we have

$$\Bigl(\frac{\ell}{\ell+1}\Bigr)^g = 1-\frac{g}{\ell+1}+O(g^2(\ell+1)^{-2})$$

for $\ell \geqslant 3$, $g \geqslant 1$, with an absolute implied constant, and thus by the Prime Number Theorem we have

$$H \geqslant \frac{1}{2}\pi(L)+O(g\log\log L+g^2)$$

with an absolute implied constant. For $L \gg g^2\log 2g$, this gives

$$H \gg \frac{1}{2}\frac{L}{\log L}$$

with an absolute implied constant, and

$$|\{t \in \mathbf{F}_q \mid f(t) \neq 0 \text{ and } |\mathsf{S}_t(\mathbf{F}_q)| \text{ is a square}\}| \ll g^2(q+q^{1/2}L^A)L^{-1}(\log L),$$

this time with no condition on g and L as this is trivial when $L \leqslant Cg^2\log 2g$ (which explains the poorer dependency on g than follows from what we said). So choosing $L = q^{1/(2A)}$, we obtain the uniform estimate

$$|\{t \in \mathbf{F}_q \mid f(t) \neq 0 \text{ and } |\mathsf{S}_t(\mathbf{F}_q)| \text{ is a square}\}| \ll q^{1-\gamma}(\log q)$$

with $\gamma = 1/(4g^2+2g+4)$ and where the implied constant is absolute. ∎

See also the end of Appendix A for a lower bound sieve result on the same families of curves.

[23] It seems that the exponent of q is better, but this reflects the use of the "right" bounds for degrees of representations of finite symplectic groups, and this exponent can be obtained with the method of [Ko1] also.

If we are in a general sieving situation as described in Section 2, we may in many cases be interested in a lower bound, in addition to the upper bounds that the large sieve naturally provides. For this purpose we can hope to appeal to the usual principles of small sieves, at least when Λ is the set of prime numbers and for some specific sieve supports. We describe this for completeness, with no claim to originality, and refer to books such as [HR], the forthcoming [IF] or [IK, §7] for more detailed coverage of the principles of sieve theory.

We assume that our sieve setting is of the type

$$\Psi = (Y, \{\text{primes}\}, (\rho_\ell)),$$

and our sieve support will be the set $\mathcal{L}$ of squarefree numbers $d < L$ for some parameter L. We write $S(X;\Omega, L)$ for the sifted set $S(X;\Omega,\mathcal{L})$. The siftable set is (X,μ,F) as before.

Let

$$\Omega_m = \prod_{\ell \mid m} \Omega_\ell$$

for m squarefree, and for an arbitrary integrable function $x \mapsto \alpha(x)$, denote

$$S_d(X;\alpha) = \int_{\{\rho_d(F_x)\in\Omega_d\}} \alpha(x) d\mu(x).$$

For $x \in X$, let $n(x) \geqslant 1$ be the integer defined by

$$n(x) = \prod_{\substack{\ell \leqslant L \\ \rho_d(F_x)\in\Omega_\ell}} \ell$$

so that for squarefree $d \in \mathcal{L}$, we have $\rho_d(F_x) \in \Omega_d$ if and only if $d \mid n(x)$.

Then we have

$$\begin{aligned}\int_{S(X;\Omega,L)} \alpha(x)d\mu(x) &= \int_{\{(n(x),P(L))=1\}} \alpha(x)d\mu(x)\\ &= \sum_{(n,P(L))=1} \Bigl(\int_{\{n(x)=n\}} \alpha(x)d\mu(x)\Bigr) = \sum_{(n,P(L))=1} a_n\end{aligned}$$

where $P(L)$ is the product of primes $\ell < L$ and

$$a_n = \int_{\{n(x)=n\}} \alpha(x) d\mu(x).$$

Note that

$$\sum_{n\equiv 0 \,(\mathrm{mod}\, d)} a_n = S_d(X;\alpha).$$

Let now $(\lambda_d^{\pm})$ be two sequences of real numbers supported on $\mathcal{L}$ such that $\lambda_1^{\pm} = 1$ and

$$\sum_{d\mid n} \lambda_d^- \leqslant 0 \leqslant \sum_{d\mid n} \lambda_d^+$$

for $n \geqslant 2$. Then, if $\alpha(x) \geqslant 0$ for all x, we have

$$\sum_{(n,P(L))=1} a_n \leqslant \sum_n \Bigl(\sum_{d\mid (n,P(L))} \lambda_d^+\Bigr) a_n = \sum_{d<L} \lambda_d^+ \Bigl(\sum_{n\equiv 0\,(\mathrm{mod}\, d)} a_n\Bigr) = \sum_{d<L} \lambda_d^+ S_d(X;\alpha)$$

and similarly

$$\sum_{(n,P(L))=1} a_n \geqslant \sum_{d<L} \lambda_d^- S_d(X;\alpha).$$

It is natural to introduce the approximations (compare (2.9))

$$S_d(X;\alpha) = \nu_d(\Omega_d) H + r_d(X;\alpha), \tag{A.1}$$

(where ν_d is the a density as in Section 2), which is really a definition of $r_d(X;\alpha)$, where the "expected main term" is

$$H = \int_X \alpha(x) d\mu(x).$$

Then, in effect, we have proved:

Proposition A.1. *Assume $\alpha(x) \geqslant 0$ for all $x \in X$. Let $\lambda_d^{\pm}$ be arbitrary upper and lower-bound sieve coefficients which vanish for $d \geqslant L$. We have then*

$$V^-(\Omega)H - R^-(X;L) \leqslant \int_{S(X;\Omega,L)} \alpha(x)d\mu(x) \leqslant V^+(\Omega)H + R^+(X;L)$$

where

$$V^{\pm}(\Omega) = \sum_{d<L} \lambda_d^{\pm} \nu_d(\Omega_d) \qquad \text{and} \qquad R^{\pm}(X;L) = \sum_{d<L} |\lambda_d^{\pm} r_d(X;\alpha)|.$$

In fact this is not quite what is needed for applications, because $V^{\pm}(X)$ are not yet in a form that makes them easy to evaluate. This next crucial step (called a "fundamental lemma") depends on the choice of $\lambda_d^{\pm}$ (which is by no means obvious) and on properties of Ω_d. For instance, we have the following (see e.g. [IK, Cor. 6.2]; note this by no means the most general or best result known).

Proposition A.2. *Let $\kappa > 0$ and $y > 1$. There exist upper and lower-bound sieve coefficients $(\lambda_d^{\pm})$, depending only on κ and y, supported on squarefree integers $< y$, bounded by one in absolute value, with the following properties: for all $s \geqslant 9\kappa + 1$ and $L^{9\kappa+1} < y$, we have*

$$\int_{S(X;\Omega,L)} \alpha(x)d\mu(x) < \Bigl(1 + e^{9\kappa+1-s}K^{10}\Bigr) \prod_{\ell<L} (1 - \nu_\ell(\Omega_\ell))H + R^+(X;L^s),$$

$$\int_{S(X;\Omega,L)} \alpha(x)d\mu(x) > \Bigl(1 - e^{9\kappa+1-s}K^{10}\Bigr) \prod_{\ell<L} (1 - \nu_\ell(\Omega_\ell))H + R^-(X;L^s),$$

provided the sieving sets (Ω_ℓ) satisfy the condition

$$\text{(A.2)} \qquad \prod_{w\leqslant \ell<L} (1 - \nu_\ell(\Omega_\ell))^{-1} \leqslant K\Bigl(\frac{\log L}{\log w}\Bigr)^{\kappa}, \qquad \textit{for all } w \textit{ and } L,\ 2 \leqslant w < L < y,$$

for some $K \geqslant 0$.

In standard applications, $r_d(X;\alpha)$ should be "small", as the remainder term in some equidistribution theorem. Note again that this can only be true if the family (ρ_d) is linearly disjoint. If this remainder is well-controlled on average over $d < D$, for some D (as large as possible) we can apply the above for L such that $L^s < D$ (with $s \geqslant 9\kappa + 1$). Note that when s is large enough (i.e., L small enough), the coefficient $1 \pm e^{9\kappa+1-s}K^{10}$ will be close to 1, in particular it will be *positive* in the lower bound.

Note that the condition (A.2) holds if $\nu_\ell(\Omega_\ell)$ is of size $\kappa\ell^{-1}$ on average. This is the traditional context of a "small sieve" of dimension κ; we see that in the abstract framework, this means rather that the sieving sets Ω_ℓ are "of codimension 1" in a certain sense. The important case $\kappa = 1$ (the classical "linear sieve") corresponds intuitively to sieving sets defined by a single irreducible algebraic condition.

Note also that the factor

$$\prod_{\ell<L} (1 - \nu_\ell(\Omega_\ell))$$

is the natural one to expect intuitively if $\nu_\ell(\Omega_\ell)$ is interpreted as the probability of $\rho_\ell(F_x)$ being in Ω_ℓ, and the various ℓ being independent. To see the connection with the quantity H^{-1} in the large sieve bound (2.4), note that if $\mathcal{L}$ is the full power set of the prime sieve support $\mathcal{L}^*$, then multiplicativity gives

$$H = \sum_{m \in \mathcal{L}} \prod_{\ell | m} \frac{\nu(\Omega_\ell)}{\nu(Y_\ell - \Omega_\ell)} = \prod_{\ell \in \mathcal{L}^*} \Bigl(1 + \frac{\nu(\Omega_\ell)}{\nu(Y_\ell - \Omega_\ell)}\Bigr) = \prod_{\ell \in \mathcal{L}^*} \frac{1}{1 - \nu_\ell(\Omega_\ell)}$$

(recall $\nu_\ell(Y_\ell) = 1$). So H^{-1} has exactly the same shape as the factor above. Of course, as in small sieves, if $\mathcal{L}$ is as large as the power set of $\mathcal{L}^*$, the large sieve constant will be much too big for the large sieve inequality to be useful, and so "truncation" is needed.

We conclude with a simple application, related to Theorem 1.5 and Section 11.

Proposition A.3. *Let q be a power of a prime number $p \geqslant 5$, $g \geqslant 1$ an integer and let $f \in \mathbf{F}_q[T]$ be a squarefree polynomial of degree $2g$. For t not a zero of f, let C_t denote the smooth projective model of the hyperelliptic curve $y^2 = f(x)(x-t)$, and let J_t denote its Jacobian variety. There exists an absolute constant $\alpha \geqslant 0$ such that*

$$|\{u \in \mathbf{F}_q \mid f(t) \neq 0 \text{ and } |C_t(\mathbf{F}_q)| \text{ has no odd prime factor} < q^\gamma\}| \gg \frac{q}{\log q}$$
$$|\{u \in \mathbf{F}_q \mid f(t) \neq 0 \text{ and } |J_t(\mathbf{F}_q)| \text{ has no odd prime factor} < q^\gamma\}| \gg \frac{q}{\log q}$$

for any γ such that

$$\gamma^{-1} > \alpha(2g^2 + g + 1)(\log\log 3g),$$

where the implied constants depends only on g and γ.

In particular, for any fixed g, there are infinitely many points $t \in \bar{\mathbf{F}}_q$ such that $|C_t(\mathbf{F}_{q^{\deg(t)}})|$ has at most $\alpha(2g^2 + g + 1)(\log\log 3g) + 2$ prime factors, and similarly for $|J_t(\mathbf{F}_{q^{\deg(t)}})|$.

Remark A.4. (1) It may well be that $|J_t(\mathbf{F}_q)|$ is even for all t, since if f has a root x_0 in $\mathbf{F}_q$, it will define a non-zero point of order 2 in $J_t(\mathbf{F}_q)$.

(2) See e.g. [Co] for results on almost prime values of group orders of elliptic curves over $\mathbf{Q}$ modulo primes; except for CM curves, they are conditional on GRH.

Proof. Obviously we use the same coset sieve setting and siftable set as in Theorems 11.4 and Theorem 1.5, and consider the sieving sets

$$\Omega_\ell^J = \{g \in CSp(2g, \mathbf{F}_\ell) \mid g \text{ is } q\text{-symplectic and } \det(g-1) = 0 \in \mathbf{F}_\ell\}\},$$
$$\Omega_\ell^C = \{g \in CSp(2g, \mathbf{F}_\ell) \mid g \text{ is } q\text{-symplectic and } \mathrm{Tr}(g) = q+1\}\},$$

for $\ell \geqslant 3$, where $\mathsf{S} \in \{C, J\}$. By (5) and (6) of Proposition B.1, we have

$$\nu_\ell(\Omega_\ell^{\mathsf{S}}) = \frac{|\Omega_\ell^{\mathsf{S}}|}{|Sp(2g, \mathbf{F}_\ell)|} \leqslant \min\Bigl(1, \frac{\ell^{g-1}}{(\ell-1)^g}\Bigr),$$

from which (A.2) can be checked to hold with $\kappa = 1$ and $K \ll \log g$ (consider separately primes $\ell < g$ and $\ell \geqslant g$).

Coming to the error term $R^-(X; L)$, individual estimates for $r_d(X; \alpha)$ with $\alpha(x) = 1$ amount to estimates for the error term in the Chebotarev density theorem. Using Proposition 11.1 in the standard way we obtain

$$r_d(X; \alpha) \ll gq^{1/2} |\Omega_d^{\mathsf{S}}|^{1/2} \ll gq^{1/2} \Bigl(\psi(d)^{2g^2+g} d^{g-1} \varphi(d)^{-g}\Bigr)^{1/2},$$

with absolute implied constants (see also [Ko2, Th. 1.3]), and hence

$$R^-(X; L^s) \ll gq^{1/2} L^{s(2g^2+g+1)/2} (\log\log L^s)^{g^2+g},$$

for any $s \geqslant 1$, with an absolute implied constant.

Let $s = \log 2 + 10 \log K \ll \log\log 3g$, and let $\varepsilon > 0$ be arbitrarily small. Then we can take

$$L = q^{(s(2g^2+g+1))^{-1} - \varepsilon}$$

in the lower bound sieve, which gives

$$|\{t \in \mathbf{F}_q \mid f(t) \neq 0 \text{ and } |\mathsf{S}_t(\mathbf{F}_q)| \text{ has no odd prime factor } < L\}|$$
$$\gg q \prod_{\substack{\ell < L \\ \nu_\ell(\Omega_\ell^{\mathsf{S}})<1}} (1-\nu_\ell(\Omega_\ell^{\mathsf{S}})) \gg q \prod_{3g<\ell<L} \Bigl(1-\frac{\ell^{g-1}}{(\ell+1)^g}\Bigr)$$

provided $L > 3g$, say, the implied constant being absolute. Putting all together, the theorem follows now easily. □

Appendix B: local density computations over finite fields

In Sections 9, 11, and in the previous Appendix, we have quoted various estimates for the "density" of certains subsets of matrix groups over finite fields, which are required to prove lower (or upper) bounds for the saving factor H in certain applications of the large sieve inequalities. We prove those statements here, relying mostly on the work of Chavdarov [Ch] to link such densities with those of polynomials of certain types which are much easier to compute. In one case, however, we use the Riemann Hypothesis over finite fields to estimate a multiplicative exponential sum.

Proposition B.1. *Let $\ell \geqslant 3$ be a prime number.*

(1) *Let $G = SL(n, \mathbf{F}_\ell)$ or $G = Sp(2g, \mathbf{F}_\ell)$, with $n \geqslant 2$ or $g \geqslant 1$. Then we have*

$$\frac{1}{|G|}|\{g \in G \mid \det(g-T) \in \mathbf{F}_\ell[X] \text{ is irreducible}\}| \gg 1$$

where the implied constant depends only on n or g.

(2) *Let $G = SL(n, \mathbf{F}_\ell)$ or $G = Sp(2g, \mathbf{F}_\ell)$, with $n \geqslant 2$ or $g \geqslant 1$, let i, j be integers with $1 \leqslant i, j \leqslant n$ or $1 \leqslant i, j \leqslant 2g$ respectively. Then we have*

$$\frac{1}{|G|}|\{g = (g_{\alpha,\beta}) \in G \mid g_{i,j} \in \mathbf{F}_\ell \text{ is not a square}\}| \gg 1$$

where the implied constant depends only on n or g.

(3) *Let $G = CSp(2g, \mathbf{F}_\ell)$ with $g \geqslant 1$, and denote by $m(g) \in \mathbf{F}_\ell^\times$ the multiplicator of a symplectic similitude $g \in G$. Then for any $q \in \mathbf{F}_\ell^\times$, we have*

$$\frac{1}{|Sp(2g, \mathbf{F}_\ell|}|\{g \in G \mid m(g) = q \text{ and } \det(g-1) \text{ is a square in } \mathbf{F}_\ell\}| \geqslant \frac{1}{2}\Bigl(\frac{\ell}{\ell+1}\Bigr)^g.$$

(4) *Let $G = CSp(2g, \mathbf{F}_\ell)$ with $g \geqslant 1$. Then for any $q \in \mathbf{F}_\ell^\times$, we have*

$$\frac{1}{|Sp(2g, \mathbf{F}_\ell|}|\{g \in G \mid m(g) = q \text{ and } q+1-\mathrm{Tr}(g) \text{ is a square in } \mathbf{F}_\ell\}| \geqslant \frac{1}{2}\Bigl(\frac{\ell}{\ell+1}\Bigr)^g.$$

(5) *Let $G = CSp(2g, \mathbf{F}_\ell)$ with $g \geqslant 1$. Then for any $q \in \mathbf{F}_\ell^\times$, we have*

$$\frac{1}{|Sp(2g, \mathbf{F}_\ell|}|\{g \in G \mid m(g) = q \text{ and } \det(g-1) = 0\}| \leqslant \min\Bigl(1, \frac{\ell^{g-1}}{(\ell-1)^g}\Bigr).$$

(6) *Let $G = CSp(2g, \mathbf{F}_\ell)$ with $g \geqslant 1$. Then for any $q \in \mathbf{F}_\ell^\times$, we have*

$$\frac{1}{|Sp(2g, \mathbf{F}_\ell|}|\{g \in G \mid m(g) = q \text{ and } q+1-\mathrm{Tr}(g) = 0\}| \leqslant \min\Bigl(1, \frac{\ell^{g-1}}{(\ell-1)^g}\Bigr).$$

Proof. (1) (Compare with [Ch, §3], [Ko1, Lemma 7.2]). Take the case $G = SL(n, \mathbf{F}_\ell)$, for instance. We need to compute

$$\frac{1}{|G|}\sum_{f \in \tilde{\Omega}_\ell} |\{g \in G \mid \det(g-T) = (-1)^n f\}|,$$

where f runs over the set $\tilde{\Omega}_\ell$ of irreducible monic polynomials $f \in \mathbf{F}_\ell[T]$ of degree n with $f(0) = 1$. For each f, we have

$$|\{g \in G \mid \det(g - T) = f\}| \gg \frac{|G|}{\ell^{n-1}}$$

by the argument in [Ch, Th. 3.5] (note that the algebraic group $SL(n)$ is connected and simply connected), and the number of f is close to $\frac{1}{n}\ell^{n-1}$ as $\ell \to +\infty$, by identifying the set of f with the set of Galois-orbits of elements of norm 1 in $\mathbf{F}_{\ell^n}$ which are of degree n and not smaller. All this implies the result for G, the case of the symplectic group being similar.

(2) By detecting squares using the Legendre character, we need to compute

$$\frac{1}{2|G|} \sum_{\substack{g \in G \\ g_{i,j} \neq 0}} \Bigl(1 + \Bigl(\frac{g_{i,j}}{\ell}\Bigr)\Bigr)$$

where $(\frac{\cdot}{\ell})$ is the non-trivial quadratic character of $\mathbf{F}_\ell^\times$. Let $\mathbf{G}$ be the algebraic group $SL(n)$ or $Sp(2g)$ over $\mathbf{F}_\ell$, d its dimension (either $n^2 - 1$ or $2g^2 - g$). Since $\mathbf{G} \cap \{g_{i,j} = 0\}$ is obviously a proper closed subset of the geometrically connected affine variety $\mathbf{G}$, the affine variety

$$\mathbf{G}_{i,j} = \mathbf{G} - \mathbf{G} \cap \{g_{i,j} = 0\}$$

over $\mathbf{F}_\ell$ is geometrically connected of dimension d, and we have

$$|\mathbf{G}_{i,j}(\mathbf{F}_\ell)| = |\{g \in \mathbf{G}(\mathbf{F}_\ell) \mid g_{i,j} \neq 0\}| \gg |\mathbf{G}(\mathbf{F}_\ell)|,$$

for $\ell \geqslant 3$. This means that it is enough to prove

$$\sum_{g \in \mathbf{G}_{i,j}(\mathbf{F}_\ell)} \Bigl(\frac{g_{i,j}}{\ell}\Bigr) \ll \ell^{d-1/2}$$

for $\ell \geqslant 3$, the implied constant depending only on $\mathbf{G}$. Such a bound follows (for instance) from the fact that this sum is a multiplicative character sum over the $\mathbf{F}_\ell$-rational points of the geometrically connected affine algebraic variety $\mathbf{G}_{i,j}$ of dimension d.

Instead of looking for an elementary proof (which may well exist), we invoke the powerful ℓ-adic cohomological formalism (see e.g. [IK, 11.11] for an introduction, and compare with the proof of Proposition 11.1). Using the (rank 1) Lang-Kummer sheaf $\mathcal{K} = \mathcal{L}_{(\frac{g_{i,j}}{\ell})}$, we have by the Grothendieck-Lefschetz trace formula

$$\sum_{g \in \mathbf{G}_{i,j}(\mathbf{F}_\ell)} \Bigl(\frac{g_{i,j}}{\ell}\Bigr) = \sum_{g \in \mathbf{G}_{i,j}(\mathbf{F}_\ell)} \mathrm{Tr}(\mathrm{Fr}_{g,\ell} \mid \mathcal{K}) = \sum_{k=0}^{2d} \mathrm{Tr}(\mathrm{Fr} \mid H^k_c(\overline{\mathbf{G}}_{i,j}, \mathcal{K}))$$

where $\mathrm{Fr}_{g,\ell}$ (resp. Fr) is the local (resp. global) geometric Frobenius for g seen as defined over $\mathbf{F}_\ell$ (resp. acting on the cohomology of the base-changed variety to an algebraic closure of $\mathbf{F}_\ell$). By Deligne's Riemann Hypothesis (see, e.g., [IK, Th. 11.37]), we have

$$\sum_{g \in \mathbf{G}_{i,j}(\mathbf{F}_\ell)} \Bigl(\frac{g_{i,j}}{\ell}\Bigr) \ll q^d \dim H^{2d}_c(\overline{\mathbf{G}}_{i,j}, \mathcal{K}) + q^{d-1/2} \sum_{k<2d} \dim H^k_c(\overline{\mathbf{G}}_{i,j}, \mathcal{K})$$

$$\ll q^d \dim H^{2d}_c(\overline{\mathbf{G}}_{i,j}, \mathcal{K}) + q^{d-1/2}$$

for $\ell \geqslant 3$, by results of Bombieri or Adolphson–Sperber that show that the sum of dimensions of cohomology groups is bounded independently of ℓ (see, e.g., [IK, Th. 11.39]).

It therefore remains to prove that $H^{2d}_c(\overline{\mathbf{G}}_{i,j}, \mathcal{K}) = 0$. However, this space is isomorphic (as vector space) to the space of coinvariants of the geometric fundamental group of $\mathbf{G}_{i,j}$ acting on a one-dimensional space through the character which "is" the Lang-Kummer sheaf $\mathcal{K}$. This means that either the coinvariant space is zero, and we are done, or otherwise the sheaf is geometrically trivial. The latter translates to the fact that the traces on $\mathcal{K}$ of the local Frobenius Fr_{g,ℓ^ν} of rational points $g \in \mathbf{G}_{i,j}(\mathbf{F}_{\ell^\nu})$ over all extensions fields $\mathbf{F}_{\ell^\nu}/\mathbf{F}_\ell$ depend only on ν, i.e., the map

$$g \mapsto \Bigl(\frac{N_{\mathbf{F}_{\ell^\nu}/\mathbf{F}_\ell} g_{i,j}}{\ell}\Bigr)$$

on $\mathbf{G}_{i,j}(\mathbf{F}_{\ell^\nu})$ depends only on ν. But this is clearly impossible for $SL(n)$ or $Sp(2g)$ with $n \geqslant 2$, $g \geqslant 1$ (but not for $SL(1)$ or for $SL(2,\mathbf{F}_2)$...), because we can explicitly write down matrices even in $\mathbf{G}(\mathbf{F}_\ell)$ both with $g_{i,j}$ a non-zero square and $g_{i,j}$ not a square (taking $\ell \geqslant 3$ for $SL(2,\mathbf{F}_\ell)$).

(3) and (4): those are similar to (1). Namely, define first a q-symplectic polynomial f in $\mathbf{F}_\ell[X]$ to be one of degree $2g$ such that[24]

$$f(0) = 1, \qquad \text{and} \qquad (qT)^{2g} f(1/(qT)) = f(T).$$

We can express such a q-symplectic polynomial uniquely in the form

$$f(T) = 1 + a_1(f)T + \cdots + a_{g-1}(f)T^{g-1} + a_g(f)T^g + \\ qa_{g-1}(g)T^{g+1} + \cdots + q^{g-1}a_1(f)T^{2g-1} + q^g T^{2g},$$

with $a_i(f) \in \mathbf{F}_\ell$, and this expression gives a bijection

$$f \mapsto (a_1(f), \ldots, a_g(f))$$

between the set of q-symplectic polynomials and $\mathbf{F}_\ell^g$.

Then we need to bound

$$\frac{1}{|Sp(2g,\mathbf{F}_\ell)|} \sum_{f \in \Omega^\gamma} |\{g \in G \mid \det(1 - Tg) = f\}| \tag{B.1}$$

where we have put (in case (3) and (4) respectively)

$$\Omega^{(3)} = \{f \in \mathbf{F}_\ell[T] \mid f \text{ is } q\text{-symplectic and } f(1) \text{ is a square in } \mathbf{F}_\ell\},$$
$$\Omega^{(4)} = \{f \in \mathbf{F}_\ell[T] \mid f \text{ is } q\text{-symplectic and } q + 1 - a_1(f) \text{ is a square in } \mathbf{F}_\ell\}.$$

Now it is easy to check that we have

$$|\Omega^\gamma| = \frac{\ell^g + \ell^{g-1}}{2} \geqslant \frac{\ell^g}{2} \tag{B.2}$$

for $\gamma = 3$ or 4 (recall ℓ is odd). Indeed, treating the case $\gamma = 3$ (the other is similar), we have

$$|\Omega^{(3)}| = |\{f \mid f(1) = 0\}| + \frac{1}{2} \sum_{f(1) \neq 0} \Bigl(1 + \Bigl(\frac{f(1)}{\ell}\Bigr)\Bigr).$$

The first term is ℓ^{g-1} since $f \mapsto f(1)$ is a non-zero linear functional on $\mathbf{F}_\ell^g$. The first part of the second sum is $(\ell^g - \ell^{g-1})/2$, and the last is

$$\sum_{(a_2,\ldots,a_g)} \sum_{a_g \neq -\tilde{f}(1)} \Bigl(\frac{a_g + \tilde{f}(1)}{\ell}\Bigr)$$

where $\tilde{f}(1)$ is defined by $f(1) = a_g + \tilde{f}(1)$ (note that $\tilde{f}(1)$ depends only on $(a_2, \ldots, a_g)$). Because of the summation over the free variable a_g, this expression vanishes.

Now appealing to Lemma 7.2 of [Ko1] (itself derived from the work of Chavdarov), we obtain

$$\frac{1}{|Sp(2g,\mathbf{F}_\ell)|} |\{g \in G \mid \det(1 - Tg) = f\}| \geqslant \frac{1}{(\ell+1)^g} \tag{B.3}$$

for all q-symplectic polynomials f, and hence the stated bound follows by (B.1), (B.2), (B.3).

(5) and (6): this is again similar to (3) and (4), where we now deal with

$$\frac{1}{|Sp(2g,\mathbf{F}_\ell)|} \sum_{f \in \Omega^\gamma} |\{g \in G \mid \det(1 - Tg) = f\}|$$

[24] Unfortunately, this is not stated correctly in [Ko1], although none of the results there are affected by this slip...

with now

$$\Omega^{(5)} = \{f \in \mathbf{F}_\ell[T] \mid f \text{ is } q\text{-symplectic and } f(1) = 0\},$$
$$\Omega^{(6)} = \{f \in \mathbf{F}_\ell[T] \mid f \text{ is } q\text{-symplectic and } q+1 = a_1(f)\}.$$

We have in both cases $|\Omega^\gamma| = \ell^{g-1}$ since the condition is a linear one on the coefficients. By the proof of Lemma 7.2 of [Ko1] we also have

$$|\{g \in G \mid \det(1 - Tg) = f\}| \leqslant \frac{1}{(\ell-1)^g}$$

for all f, and therefore

$$\frac{1}{|Sp(2g, \mathbf{F}_\ell)|} \sum_{f \in \Omega^\gamma} |\{g \in G \mid \det(1 - Tg) = f\}| \leqslant \frac{\ell^{g-1}}{(\ell-1)^g}.$$

Since the quantity to estimate is also at most 1 for trivial reasons, we have the desired result. □

References

[BMS] H. Bass, J. Milnor and J-P. Serre: *Solution of the congruence subgroup problem for SL_n ($n \geqslant 3$) and Sp_{2n} ($n \geqslant 2$)*, Publ. Math. I.H.É.S 33 (1967), 59–137.

[B] E. Bombieri: *Le grand crible dans la théorie analytique des nombres*, Astérisque 18, S.M.F (1974).

[BGS] J. Bourgain, A. Gamburd and P. Sarnak: *Sieving and expanders*, C. R. Acad. Sci. Paris Sér. I Math. 343 (2006), 155–159.

[BF] J. Brüdern and E. Fouvry: *Lagrange's four squares theorem with almost prime variables*, J. reine angew. Math. 454 (1994), 59–96.

[Ca] R.W. Carter: *Finite groups of Lie type*, Wiley Interscience (1985).

[Ch] N. Chavdarov: *The generic irreducibility of the numerator of the zeta function in a family of curves with large monodromy*, Duke Math. J. 87 (1997), 151–180.

[Co] A. Cojocaru: *Reductions of an elliptic curve with almost prime orders*, Acta Arithmetica 119 (2005), 265–289.

[D1] P. Deligne: *La conjecture de Weil, II*, Publ. Math. I.H.É.S 52 (1981), 313–428.

[D2] P. Deligne: *Cohomologie étale*, S.G.A $4\frac{1}{2}$, L.N.M 569, Springer Verlag (1977).

[DM] F. Digne and J. Michel: *Representations of finite groups of Lie type*, L.M.S Student Texts 21, (1991).

[D] W. Duke: *Elliptic curves with no exceptional primes*, C. R. Acad. Sci. Paris 325 (1997), no. 8, 813–818.

[DRS] W. Duke, Z. Rudnick and P. Sarnak: *Density of integer points on affine homogeneous varieties*, Duke Math. J. 71 (1993), 143–179.

[EEW] M. Einsiedler, G. Everest and T. Ward: *Primes in elliptic divisibility sequences*, LMS J. Comput. Math. 4 (2001), 1–13.

[EMS] A. Eskin, S. Mozes and N. Shah: *Unipotent flows and counting lattice points on homogeneous varieties*, Ann. of Math. 143 (1996), 253–299.

[FLP] A. Fathi, F. Laudenbach and V. Poénaru: *Travaux de Thurston sur les surfaces*, Astérisque 66/67, S.M.F (1979).

[F] W. Feller: *An introduction to probability theory and its applications*, Vol. I, 3rd edition, John Wiley & Sons (1968).

[FM] É. Fouvry and P. Michel: *Sur les changements de signe des sommes de Kloosterman*, Annals of Math., to appear.

[G] P.X. Gallagher: *The large sieve and probabilistic Galois theory*, in Proc. Sympos. Pure Math., Vol. XXIV, Amer. Math. Soc. (1973), 91–101.

[Go] R. Gow: *Properties of the characters of the finite general linear group related to the transpose-inverse involution*, Proc. London Math. Soc. 47 (1983), 493–506.

[Gr] A. Grothendieck: *Formule de Lefschetz et rationalité des fonctions L*, Séminaire Bourbaki (1964–65), North Holland, 1968, exp. 279, 1–15.

[H] C. Hall: *Big orthogonal or symplectic monodromy mod ℓ*, preprint (2006), `arXiv:math.NT/0608718`

[HR] H. Halberstam and H. Richert: *Sieve methods*, Academic Press (1974).

[HW] G.H. Hardy and E.M. Wright: *An introduction to the theory of numbers*, Fifth Edition, Oxford 1979.

[Ha] P. de la Harpe: *Topics in Geometric Group Theory*, Chicago Lectures in Math., Univ.of Chicago Press (2000).

[HV] P. de la Harpe and A. Valette: *La propriété (T) de Kazhdan pour les groupes localement compacts*, Astérisque 175, Soc. Math. France (1989).

[Hu] M.N. Huxley: *The large sieve inequality for algebraic number fields*, Mathematika 15 (1968), 178–187.

[Iv] N. Ivanov: *Mapping class groups*, Handbook of geometric topology, 523–633, North-Holland, Amsterdam, 2002.

[IF] H. Iwaniec and J. Friedlander: *Sieve theory*, in preparation.

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

[Ka] M. Kassabov: *Kazdhan constants for $SL_n(\mathbf{Z})$*, Int. J. Algebra Comput. 15 (2005), 971–995.

[Kl] A.A. Klyachko: *Models for complex representations of the groups $GL(n,q)$*, Mat. Sb. (1983), 371–386.

[Kn] A. Knapp: *Representation theory of semisimple groups*, Princeton Math. Series 36, Princeton Univ. Press, 1986.

[Ko1] E. Kowalski: *The large sieve, monodromy and zeta functions of curves*, J. reine angew. Math, to appear, `arXiv:math.NT/0503714`

[Ko2] E. Kowalski: *On the rank of quadratic twists of elliptic curves over function fields*, International J. of Number Theory 2 (2006), 267–288.

[Ko3] E. Kowalski: *Big symplectic monodromy: a theorem of C. Hall*, note available at `http://www.math.u-bordeaux1.fr/~kowalski/short-notes.html`

[LW] Y.K. Lau and J. Wu: *Sums of some multiplicative functions over a special set of integers*, Acta Arith. 101 (2002), 365–394.

[LM] Y-R. Liu and R. Murty: *Sieve methods in combinatorics*, Journal of Combinatorial Theory, Series A 111 (2005) 1–23.

[Lu] A. Lubotzky: *Discrete groups, expanding graphs and invariant measures*, Progr. Math. 125, Birkhäuser, 1994.

[LS] A. Lubotzky and D. Segal: *Subgroup growth*, Progr. Math. 212, Birkhäuser, 2003.

[LZ] A. Lubotzky and A. Żuk: *On Property (τ)*, draft, available online.

[Lu] G. Lusztig: *Representations of Finite Chevalley Groups*, CBMS 39, A.M.S (1978).

[Ma] J. Maher: *Random walks on the mapping class group*, preprint (2006), `arXiv:math.GT/060443`.

[Mi] J. Milne: *Étale cohomology*, Princeton Mathematical Series 33, Princeton Univ. Press, 1980.

[Mo] H.L. Montgomery: *The analytic principle of the large sieve*, Bull. A.M.S 84 (1978), 547–567.

[P] R. Pink: *The Mumford-Tate conjecture for Drinfeld-modules*, Publ. Res. Inst. Math. Sci. 33 (1997), no. 3, 393–425.

[Po] B. Poonen: *Bertini theorems over finite fields*, Ann. of Math. (2) 160 (2004), no. 3, 1099–1127.

[Ri] K. Ribet: *On ℓ-adic representations attached to modular forms*, Invent. math. 28 (1975), 245–275.

[R] I. Rivin: *Counting Reducible Matrices, Polynomials, and Surface and Free Group Automorphisms*, preprint (2006), `arXiv:math.NT/0604489`

[SC] L. Saloff-Coste: *Random walks on finite groups*, in "Probability on discrete structures", 263–346, Encyclopaedia Math. Sci., 110, Springer, Berlin, 2004.

[Sa1] P. Sarnak: email to J. Bourgain, A. Gamburd (May 2006).

[Sa2] P. Sarnak: *Some applications of modular forms*, Cambridge Tracts in Math. 99, Cambridge Univ. Press 1990.

[Sa3] P. Sarnak: slides for Rademacher Lectures, Philadelphia, September 2006, available at `http://www.math.princeton.edu/sarnak/rademacher1.pdf`

[S2] J-P. Serre: *Linear representations of finite groups*, Grad. Texts in Math. 42, Springer Verlag (1977).

[S2] J-P. Serre: *Zeta and L-functions*, in Arithmetical Algebraic Geometry, Harper and Row, New York (1965), 82–92 (=Oeuvres, vol. II, no 64).

[S3] J-P. Serre: *Spécialisation des éléments de* $\mathrm{Br}_2(\mathbf{Q}(T_1,\cdots,T_n))$, C. R. Acad. Sci. Paris Sér. I Math. 311 (1990), no. 7, 397–402 (=Oeuvres, vol. III, no 94).

[S4] J-P. Serre: *Propriétés galoisiennes des points d'ordre fini des courbes elliptiques*, Invent. math. 15 (1972), 259–331 (=Oeuvres, vol. IV, no 150).

[Sha] Y. Shalom: *Bounded generation and Kazhdan's property (T)*, Publ. Math I.H.É.S 90 (1999), 145–168.

[Shi] G. Shimura: *Introduction to the arithmetic theory of automorphic functions*, Princeton Univ. Press, 1971.

[Si1] J. Silverman: *Wieferich's criterion and the abc-Conjecture*, J. of Number Theory 30 (1988), 226–237.

[Si2] J. Silverman: *p-adic properties of division polynomials and elliptic divisibility sequences*, Math. Ann. 332 (2005), 443–471.

[Si3] J. Silverman: *The arithmetic of elliptic curves*, Grad. Texts in Math. 106, 1986.

[Sp] T.A. Springer: *Linear algebraic groups*, 2nd edition, Progress in Math. 9, Birkhaüser (1998).

[St] R. Steinberg: *The representations of $GL(3,q)$, $GL(4,q)$, $PGL(3,q)$ and $PGL(4,q)$*, Canad. J. Math. 3 (1951), 225–235.

[V] C.R. Vinroot: *Twisted Frobenius-Schur indicators of finite symplectic groups*, J. Algebra 293 (2005), 279–311.

[W] M. Ward: *Memoir on elliptic divisibility sequences*, Amer. J. Math. 70 (1948), 31–74.

[Z] D. Zywina: work in progress.

Université Bordeaux I - A2X, 351, cours de la Libération, 33405 Talence Cedex, France

E-mail address: `emmanuel.kowalski@math.u-bordeaux1.fr`