1
Introduction

1.1 Presentation

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 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 with p, 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. One may quote among these the vector sieve of Brüdern and Fouvry [**17**], with applications to Lagrange's theorem with almost prime variables; the 'crible étrange' of Fouvry and Michel [**42**], with applications to sign changes of Kloosterman sums, and Poonen's striking sieve procedure for finding smooth hypersurfaces of large degree over finite fields [**105**] (which we describe briefly in Example 4.11).

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. [**46**]). Interesting applications have been found, e.g. Duke's theorem on elliptic curves over $\mathbf{Q}$ with 'maximal' p-torsion fields for all p [**32**]. All these were much of

the same flavour however, and in particular depended only on the character theory of finite abelian groups as far as the underlying harmonic analysis was concerned.

In [**80**], we introduced a new large sieve inequality to study 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 behaviour of the numerator of the zeta function of a smooth projective curve in a family with large monodromy, improving significantly a result of Chavdarov [**22**]. (We will survey and again improve these results in Chapter 8.)

As explained in the preface, while working on devising a general framework of the sieve that can recover both the classical forms or the version in [**80**], a number of new applications emerged. Some of them are in areas of number theory not usually directly linked to sieve methods, and some in decidedly different contexts. Hence the goal of this book is to present the large sieve as a general mathematical *principle* which has potential applications outside number theory. For this reason, we start from scratch, assuming only a knowledge of basic linear algebra and properties of finite-dimensional Hilbert spaces to derive the basic inequality.

Roughly speaking, this inequality states that, given a measure space X with finite measure, and surjective maps from X to a family (X_ℓ) of *finite* sets, the measure of the set of those $x \in X$ which have image in X_ℓ outside some given sets Ω_ℓ, for finitely many ℓ, can be estimated from above by means of two quantities. One involves the 'densities' of the sets Ω_ℓ in X_ℓ, and is independent of X, while the other (the 'large sieve constant') is the norm of a certain bilinear form which depends on X and X_ℓ, but is independent of Ω_ℓ. This form of the sieve statement is similar to Montgomery's inequality, and much stronger than Linnik's original version (see, e.g. [**98**], [**11**], [**67**, 7.4]).

Obtaining this inequality is really straightforward and is done, in Chapter 2, in a few pages – the innovation, for what it's worth, is in working in the generality we consider. This does not by itself prove anything, because the large sieve constant needs to be estimated before applications can be derived, and the estimation may turn out to be impossible, or trivial. However, the problem turns out to be further reducible to the study of certain 'exponential sums' (or integrals) over X, which suggests that strong estimates should exist in many situations, related to the equidistribution of the image of X in X_ℓ. This equidistribution may be expected to be true in many cases, for fixed ℓ at least, but a key issue is

uniformity with respect to ℓ: an explicit form of the error term in the equidistribution is required to proceed. In the classical case, the bilinear form estimate was first considered by Bombieri and given its most general expression by Davenport and Halberstam.

This is the time to discuss a thorny terminological issue: this inequality (in its most refined version) takes the form

$$\sum_r \left| \sum_{M \leqslant n < M+N} a_n e(n\xi_r) \right|^2 \leqslant (N - 1 + \delta^{-1}) \sum_n |a_n|^2 \tag{1.1}$$

for arbitrary complex numbers a_n and 'angles' $\xi_r \in \mathbf{R}/\mathbf{Z}$ which are δ-spaced (i.e., such that $\min_{n\in\mathbf{Z}} |\xi_r - \xi_s - n| \geqslant \delta$ for $r \neq s$). It is often itself called 'the large sieve inequality', although it does not mention any idea of sieve, because of its link with the proof of Montgomery's inequality. Correspondingly, when generalizations of (1.1) were developed for independent reasons (replacing the characters $x \mapsto e(x\xi_r)$ by other functions), they were also called 'large sieve inequalities', even when any link to sieve theory had utterly vanished. And in fact these inequalities, particularly those involving Fourier coefficients of automorphic forms of various types, form an important body of work which has had tremendous applications in analytic number theory, starting with the work of Iwaniec, and Deshouillers–Iwaniec, and later with variants due to Duke, Duke–Kowalski, Venkatesh and others. We will not say anything beyond this, and we refer to [**67**, Section 7.7] for a short survey with some applications.

After presenting and commenting on the basic framework, the rest of the book is devoted to the explanation of a number of instances of sieves and the issues surrounding them. This is done first with the examples of Chapter 4 which present a number of (mostly) classical situations in this context, and describe some of their applications for convenience. We also indicate there the relation with the inclusion-exclusion technique in probability and combinatorics, which shows in particular that the general sieve bound is sharp, and include a first new application: an amusing 'elliptic sieve' which is related to questions surrounding the number of prime divisors of the denominators of rational points on an elliptic curve. In turn, this is linked to the analysis of the prime factorization of elements of the so-called 'elliptic divisibility sequences' first introduced by M. Ward. We find rather easily that 'most' elements have many prime factors, which complements recent heuristics and results of Silverman, Everest, Ward and others concerning the paucity of primes and prime powers in such sequences.

The following chapters are less classical and concern new (or recent) applications of the sieve ideas, which are quite independent of one another.

'Probabilistic' sieves are discussed briefly in Chapter 6, with an application to 'random' finitely presented groups, and sieving in a discrete finitely generated group G is described in much more detail in Chapter 7, where some of the most appealing new results are obtained. Indeed, for symmetric random walks on some finitely generated groups, a very transparent treatment of the large sieve constant is possible, and Property (τ) (or the expanding properties of Cayley graphs of quotients of G) appears as a completely natural tool. When this feature is present, it leads to strong sieve results. Moreover, very interesting applications arise, including surprising ones in geometry or topology.

Finally, in Chapter 8, we review and extend the sieve result of [**80**] concerning the distribution of geometric Frobenius conjugacy classes in finite monodromy groups over finite fields, and derive some new applications. There are links here with the case of arithmetic groups, and comparison of the sieve bounds coming from Property (τ) in the former case and the Riemann Hypothesis over finite fields in the latter is quite interesting.

The final part of the book is a series of appendices which review briefly some of the topics which are probably not known to all readers. This includes a discussion of small sieves, for purpose of comparison and reference, including a sample application; a survey of some techniques that are used to prove density results in matrix groups over finite fields, which are also of independent interest and involve work of Chavdarov [**22**] and non-trivial estimates for exponential sums over finite fields; a survey of representation theory of groups, involving both the classical theory for finite groups, and what is needed to describe Property (T) and Property (τ); some estimates for sums of multiplicative functions; and a short survey of basic topological facts which we use in some of our applications.

Whenever we treat an example, we give at least all definitions required to understand the essential parts of the statements, and precise references for any unproved facts which can not be assumed to be known by every potential reader. It is expected that most readers will at least once think 'Everyone knows *this*!' when reading some part of the notes, but they may not be able to say this of all such basic references.

1.2 Some new applications of the large sieve

Before going further, it seems natural to list here a few applications of the sieve framework we are going to describe. Most of those below are, to the best of our knowledge, new results, although some of them could well have been

proven before. We seek concreteness in this list: the precise results will usually be stronger and more general.

Our first result is in fact obtained from the 'traditional' large sieve in one variable, which we apply in a rather twisted way.

Theorem 1.1 *Let $E/\mathbf{Q}$ be an elliptic curve with rank $r \geqslant 1$ given by a Weierstrass equation*

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6, \quad \text{where } a_i \in \mathbf{Z}.$$

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

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

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

The second statement is an example of the philosophy that random walks on a set give a way of stating properties of random elements of X, even when there is no natural probability measure on X. Here X is the set of integers $\mathbf{Z}$, and we use simple random walks to compensate for the absence of a translation-invariant probability measure on $\mathbf{Z}$.

Theorem 1.2 *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 $\mathbf{P}(X_k = \pm 1) = 1/2$ 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

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

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

This is proved in Chapter 6. It may be expected that results of this type can be recovered from their 'deterministic' analogues using the Central Limit Theorem. However, this is not likely to be feasible (or wise) when considering similar questions about random unimodular matrices. In Chapter 7, we prove the following result using Property (τ), which confirms that generic elements of $SL(n, \mathbf{Z})$ have 'arithmetically generic' characteristic polynomials:

Theorem 1.3 *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 G, i.e., a sequence of G-valued random variables such that $X_0 = 1$ and*

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

where (ξ_k) is a sequence of S-valued independent random variables with

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

Then, almost surely, there are only finitely many k for which the characteristic polynomial $\det(T - X_k) \in \mathbf{Z}[T]$ does not have the full symmetric group $\mathfrak{S}_n$ as Galois group, or in other words, the set of matrices in $SL(n, \mathbf{Z})$ with characteristic polynomials having small Galois group is transient for the random walk. In particular, so is the set of those having reducible characteristic polynomial.

In fact (see Theorem 7.4), we will derive this by showing that the probability that $\det(T - X_k)$ be reducible decays exponentially fast with k (in the case $n \geqslant 3$ at least). The following is a consequence of a similar statement for symplectic groups, and it answers a question of Maher [**96**, Question 1.3] (see Proposition 7.17).

Theorem 1.4 *Let $g \geqslant 1$ be an integer, let G be the mapping class group of a closed surface Σ_g of genus g. Then the set of non-pseudo-Anosov elements in G is transient for any symmetric random walk on G where the steps are chosen among a fixed finite symmetric generating set of G.*

These two examples of sieves in discrete groups correspond to properties which are invariant under conjugation. The next result does not have this property, showing that the sieve is not limited to this situation. For the sake of diversity, we state the result somewhat differently in the language of products of N matrices chosen among the generating set.

Theorem 1.5 *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*

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

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.

Finally, here is a sample of what the sieve for Frobenius can do, as described in Chapter 8. Except for a slightly weaker exponent γ, it could have been proved easily with the techniques of [**80**].

Theorem 1.6 *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

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

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

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

2
The principle of the large sieve

2.1 Notation and terminology

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. Many readers, especially analytic number theorists, may find it excessively formal, but the framework we describe has so many different incarnations that it seems preferable to be very precise in this book, and to give a name to the objects involved to refer to them later on. Concrete applications will be able to eschew reproducing all this, by using self-contained statements such as those included in Section 3.5, which involve none of the newfangled terminology.

Hence, let's start. First of all, 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 of as a family of colourings 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 $|B|$ for the measure $\mu(B)$ of a measurable set $B \subset X$.

The last pieces of data are a finite subset $\mathcal{L}^*$ of Λ, called the *prime sieve support*, and a family $\Omega = (\Omega_\ell)$ of *sieving sets*,[1] $\Omega_\ell \subset Y_\ell$, defined for $\ell \in \mathcal{L}^*$.

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

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 have in mind 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 Chapter 8 for a conspicuous example where F is not the identity, Chapter 6 for interesting situations where the measure space (X, μ) is a probability space, and F a random variable, and Chapter 7 for another example.

2.2 The large sieve inequality

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,

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

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 $m \subset \Lambda$. In order to simplify notation, 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, 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 $n, m \in S(\Lambda)$). Also we sometimes do not explicitly distinguish between $\ell \in \Lambda$ and $\{\ell\} \in S(\Lambda)$.

A *sieve support* $\mathcal{L}$ associated to a prime sieve support $\mathcal{L}^*$ is any (finite) family of subsets of $\mathcal{L}$. (In general, $\mathcal{L}$ will have additional properties, in particular it will be such that $\{\ell\} \in \mathcal{L}$ for any $\ell \in \mathcal{L}^*$, but it is not necessary to assume this.)

If Λ is a set of primes, $\mathcal{L}$ 'is' a set of squarefree integers only divisible by primes in $\mathcal{L}^*$ (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' colourings of Y obtained by looking at all possible finite tuples of colourings.) If $m = \emptyset$, Y_m is a set with a single element, and ρ_m is a constant map. Note that ρ_m is not surjective in general.

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 by 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, with minor changes, but the stronger assumption is no problem for applications), and that ν is a probability density, i.e., we have

$$\sum_{y \in Y_\ell} \nu_\ell(y) = 1. \tag{2.1}$$

We denote by $L^2(Y_\ell, \nu_\ell)$, or simply $L^2(Y_\ell)$, the Hilbert space of functions on Y_ℓ with this inner product.