

Equidistribution, L -functions and ergodic theory: on some problems of Yu. Linnik

Philippe Michel and Akshay Venkatesh*

Abstract. An old question of Linnik asks about the equidistribution of integral points on a large sphere. This question proved to be very rich: it is intimately linked to modular forms, to subconvex estimates for L -functions, and to dynamics of torus actions on homogeneous spaces. Indeed, Linnik gave a partial answer using ergodic methods, and his question was completely answered by Duke using harmonic analysis and modular forms. We survey the context of these ideas and their developments over the last decades.

Mathematics Subject Classification (2000). Primary 11F66; Secondary: 11F67, 11M41.

Keywords. Automorphic L -functions, ergodic theory, equidistribution, subconvexity.

1. Linnik's problems

Given Q a homogeneous polynomial of degree m in n variables with integral coefficients, a classical problem in number theory is to understand the integral representations of an integer d by the polynomial Q , as $|d| \rightarrow +\infty$. Let $V_{Q,d}(\mathbb{Z}) = \{\mathbf{x} \in \mathbb{Z}^n, Q(\mathbf{x}) = d\}$ denote the set of such representations (possibly modulo some obvious symmetries). If $|V_{Q,d}(\mathbb{Z})| \rightarrow +\infty$ with d , it is natural to investigate the distribution of the discrete set $V_{Q,d}(\mathbb{Z})$ inside the affine variety “of level d ”

$$V_{Q,d}(\mathbb{R}) = \{\mathbf{x} \in \mathbb{R}^n, Q(\mathbf{x}) = d\}.$$

In fact, one may rather consider the distribution, inside the variety of fixed level $V_{Q,\pm 1}(\mathbb{R})$, of the radial projection $|d|^{-1/m} \cdot V_{Q,d}(\mathbb{Z})$ (here \pm is the sign of d) and one would like to show that, as $|d| \rightarrow +\infty$, the set $|d|^{-1/m} \cdot V_{Q,d}(\mathbb{Z})$ becomes equidistributed with respect to some natural measure $\mu_{Q,\pm 1}$ on $V_{Q,\pm 1}(\mathbb{R})$. Here, to take care of the case where $V_{Q,d}(\mathbb{Z})$ and $\mu_{Q,\pm 1}(V_{Q,\pm 1}(\mathbb{R}))$ are infinite, equidistribution w.r.t. $\mu_{Q,\pm 1}$ is defined by the following property: for any two sufficiently nice compact subsets $\Omega_1, \Omega_2 \subset V_{Q,\pm 1}(\mathbb{R})$ one has

$$\frac{| |d|^{-1/m} \cdot V_{Q,d}(\mathbb{Z}) \cap \Omega_1 |}{| |d|^{-1/m} \cdot V_{Q,d}(\mathbb{Z}) \cap \Omega_2 |} \longrightarrow \frac{\mu_{Q,\pm 1}(\Omega_1)}{\mu_{Q,\pm 1}(\Omega_2)} \quad \text{as } |d| \rightarrow +\infty. \quad (1.1)$$

*The research of the first author is partially supported by the Marie Curie RT Network “Arithmetic Algebraic Geometry” and by the “RAP” network of the Région Languedoc-Roussillon. The research of the second author is supported by a Clay Mathematics Research Fellowship and NSF grant DMS-0245606.

The most general approach to this kind of problems is the circle method of Hardy–Littlewood. (Un)fortunately, that method is fundamentally limited to cases where the number of variables n is large compared with the degree m . To go further, one is led to make additional hypotheses on the varieties $V_{Q,d}$. It was anticipated by Linnik in the early 60's, and systematically suggested by Sarnak in the 90s [55], [56], [68], that for varieties which are homogeneous with respect to the action of some algebraic group $G_{\mathbb{Q}}$, one should be able to take advantage of this action. Equidistribution problems on such homogeneous varieties are called (after Sarnak), equidistribution problems of Linnik's type.

By now, this expectation is largely confirmed by the resolution of wide classes of problems of Linnik's type ([10], [25], [30]–[32], [35], [57], [58], [65]); and the methods developed to deal with them rely heavily on powerful techniques from harmonic analysis (Langlands functoriality, quantitative equidistribution of Hecke points and approximations to the Ramanujan–Pettersson conjecture) or from ergodic theory (especially Ratner's classification of measures invariant under unipotent subgroups), complemented by methods from number theory.

In this lecture we will not discuss that much the resolution of these important and general cases (for this we refer to [33], [72]); instead, we wish to focus on three, much older, examples of low dimension and degree ($m = 2$, $n = 3$) which were originally studied in the sixties by Linnik and his school. Our point in highlighting these examples is that the various methods developed to handle them are fairly different from the aforementioned ones which, in fact, may not apply or at least not directly.

The three problems correspond to taking Q to be a ternary quadratic form of signature $(3, 0)$ or $(1, 2)$. They are problems of Linnik's type with respect to the action of the orthogonal group $G = \mathrm{SO}(Q)$ on $V_{Q,d}$.

The first problem is for the definite quadratic form $Q(A, B, C) = A^2 + B^2 + C^2$. For d an integer, $V_{Q,|d|}(\mathbb{Z})$ is the set of representations of $|d|$ as a sum of three squares

$$V_{Q,|d|}(\mathbb{Z}) = \{(a, b, c) \in \mathbb{Z}^3, a^2 + b^2 + c^2 = |d|\}$$

and $V_{Q,1}(\mathbb{R}) = S^2$ is the unit sphere. We denote by

$$\mathfrak{g}_d = |d|^{-1/2} \cdot V_{Q,|d|}(\mathbb{Z})$$

the radial projection of $V_{Q,|d|}(\mathbb{Z})$ on S^2 :

Theorem 1 (Duke [17]). *For $d \rightarrow -\infty$, and $d \not\equiv 0, 1, 4 \pmod{8}$ the set \mathfrak{g}_d is equidistributed on S^2 w.r.t. the Lebesgue measure μ_{S^2} .*

It will be useful to recall the “accidental” isomorphism of $\mathrm{SO}(Q)$ with $G = \mathrm{PG}(\mathbb{B}^{(2,\infty)}) = \mathbb{B}_{2,\infty}^\times / Z(\mathbb{B}_{2,\infty}^\times)$ where $\mathbb{B}^{(2,\infty)}$ is the algebra of the Hamilton quaternions. This arises from the identification of the quadratic space (\mathbb{Q}^3, Q) with the trace-0 Hamilton quaternions endowed with the norm form $N(z) = z \cdot \bar{z}$ via the map $(a, b, c) \rightarrow z = a.i + b.j + c.k$.

The second and third problems are relative to the indefinite quadratic form $Q(A, B, C) = B^2 - 4AC$, which is the discriminant of the binary quadratic forms $q_{A,B,C}(X, Y) = AX^2 + BXY + CY^2$. In that case, there is another “accidental” isomorphism of $SO(Q)$ with PGL_2 via the map

$$(a, b, c) \rightarrow q_{a,b,c}(X, Y) = aX^2 + bXY + cY^2$$

which identifies $V_{Q,d}$ with the set \mathcal{Q}_d of binary quadratic forms of discriminant d ; PGL_2 acts on the latter by linear change of variables, twisted by inverse determinant. As $PGL_2(\mathbb{Z})$ acts on $\mathcal{Q}_d(\mathbb{Z})$, one sees that, if $V_{Q,d}(\mathbb{Z}) = \mathcal{Q}_d(\mathbb{Z})$ is non empty (i.e. if $d \equiv 0, 1 \pmod{4}$), it is infinite; so the proper way to define the equidistribution of $|d|^{-1/2} \cdot V_{Q,d}(\mathbb{Z})$ inside $V_{Q,\pm 1}(\mathbb{R}) = \mathcal{Q}_{\pm 1}(\mathbb{R})$ is via (1.1). However, it is useful to formulate these problems in a slightly different (although equivalent) form which will be suitable for number theoretic applications. Let $\mathbb{H}^\pm = \mathbb{H}^+ \cup \mathbb{H}^- = \mathbb{C} - \mathbb{R} = PGL_2(\mathbb{R})/SO_2(\mathbb{R})$ denote the union of the upper and lower half-planes and $Y_0(1)$ denote the (non-compact) modular surface of full level i.e. $PGL_2(\mathbb{Z}) \backslash \mathbb{H}^\pm \simeq PSL_2(\mathbb{Z}) \backslash \mathbb{H}^+$.

As is well known, the quotient $PSL_2(\mathbb{Z}) \backslash \mathcal{Q}_d(\mathbb{Z})$ is finite, of cardinality some *class number* $h(d)$. For negative discriminants d , one associates to each $PSL_2(\mathbb{Z})$ -orbit $[q] \subset \mathcal{Q}_d(\mathbb{Z})$, the point $z_{[q]}$ in $Y_0(1)$ defined as the $PGL_2(\mathbb{Z})$ -orbit of the unique root of $q(X, 1)$ contained in \mathbb{H}^+ . These points are called *Heegner points of discriminant*¹ d and we set

$$\mathcal{H}_d := \{z_{[q]}, [q] \in PSL_2(\mathbb{Z}) \backslash \mathcal{Q}_d(\mathbb{Z})\} \subset Y_0(1).$$

An equivalent form to (1.1) for $Q(A, B, C) = B^2 - 4AC$ and $d \rightarrow -\infty$ is the following:

Theorem 2 (Duke [17]). *As $d \rightarrow -\infty$, $d \equiv 0, 1 \pmod{4}$, the set \mathcal{H}_d becomes equidistributed on $Y_0(1)$ w.r.t. the Poincaré measure $d\mu_P = \frac{3}{\pi} \frac{dx dy}{y^2}$.*

For positive discriminants d , one associates to each class of integral quadratic form $[q] \in \mathcal{Q}_d(\mathbb{Z})$ the positively oriented geodesic, $\gamma_{[q]}$, in $Y_0(1)$ which is the projection to $Y_0(1)$ of the geodesic line in \mathbb{H}^+ joining the two (real) roots of $q(X, 1)$. This is a closed geodesic – in fact, all closed geodesics on $Y_0(1)$ are of that form – whose length is essentially equal to the logarithm of the fundamental solution to Pell’s equation $x^2 - dy^2 = 4$. We denote by

$$\Gamma_d := \{\gamma_{[q]}, [q] \in PSL_2(\mathbb{Z}) \backslash \mathcal{Q}_d(\mathbb{Z})\}$$

the set of all geodesics of discriminant d .

Theorem 3 (Duke [17]). *As $d \rightarrow +\infty$, $d \equiv 0, 1 \pmod{4}$, d not a perfect square, the set Γ_d becomes equidistributed on the unit tangent bundle of $Y_0(1)$, $S_1^*(Y_0(1))$, w.r.t. the Liouville measure $d\mu_L = \frac{3}{\pi} \frac{dx dy}{y^2} \frac{d\theta}{2\pi}$.*

¹For simplicity, we will ignore non-primitive forms.

These three problems (in their form (1.1)) were first proved by Linnik and by Skubenko by means of Linnik's *ergodic method*; we will return to this method in Section 6. The proof however is subject to an additional assumption which we call *Linnik's condition*, namely:

Theorem 4 (Linnik [54], [55], Skubenko [71]). *Let p be an arbitrary fixed prime, then the equidistribution statements of Theorems 1, 2, 3 hold for the subsequence of d such that p is split in the quadratic extension $K_d = \mathbb{Q}(\sqrt{d})$.*

We will see in Section 6 that Linnik's condition has a natural ergodic interpretation. It can be relaxed to the condition that for each d there is a prime $p = p(d) \leq |d|^{\frac{1}{10^{10} \log \log |d|}}$ which splits in $\mathbb{Q}(\sqrt{d})$. The latter condition is satisfied, for instance, by assuming that the L -functions of quadratic characters satisfying the Generalized Riemann Hypothesis (GRH). In particular, Linnik's condition (resp. the weaker one) is automatically fulfilled for subsequences of d such that K_d is a *fixed* quadratic field (resp. $\text{disc}(K_d) = \exp\left(O\left(\frac{\log |d|}{\log \log |d|}\right)\right)$); however, in these cases, the proof of Theorems 1, 2, 3 is much simpler (see [11] for instance); so, as it is (from our perspective at least) the hardest case, we will limit ourselves to d 's which are fundamental discriminants (i.e. $d = \text{disc}(K_d)$).

Acknowledgements. The first author is scheduled to give a presentation based on this work in the ICM 2006. Since much of it is based on our joint work, we have decided to write this paper jointly. The results of Section 6 are all joint work with M. Einsiedler and E. Lindenstrauss and will also be discussed in their contribution to these proceedings [28].

It is our pleasure to thank Bill Duke, Henryk Iwaniec and Peter Sarnak for both their consistent encouragement and for many beautiful ideas which underlie the whole field. Peter Sarnak and Hee Oh carefully read an early draft and provided many helpful comments and corrections. We also would like to thank our collaborators Manfred Einsiedler and Elon Lindenstrauss, for patiently explaining ergodic ideas and methods to us.

2. Linnik's problems via harmonic analysis

Duke's unconditional solution of Linnik's problems is via harmonic analysis and in a sense, is very direct as it proceeds by verifying Weyl's equidistribution criterion. Let (X, μ) denote any of the probability spaces (S^2, μ_{S^2}) , $(Y_0(1), \mu_P)$, $(S_*^1(Y_0(1)), \mu_L)$. For each case and for appropriate d , let μ_d denote the probability measure formed out of the respective sets \mathcal{G}_d , \mathcal{H}_d or Γ_d : for instance for $X = S^2$,

$$\int_{S^2} \varphi \mu_d = \frac{1}{|\mathcal{G}_d|} \sum_{\substack{(a,b,c) \in \mathbb{Z}^3 \\ a^2+b^2+c^2=|d|}} \varphi\left(\frac{a}{\sqrt{|d|}}, \frac{b}{\sqrt{|d|}}, \frac{c}{\sqrt{|d|}}\right).$$

Showing that μ_d weak- $*$ converges to μ amounts to show that, for any φ ranging over a fixed orthogonal basis (made of continuous functions) of the L^2 -space $L^2_0(X, \mu)$, the Weyl sum

$$W(\varphi, d) := \int_X \varphi \mu_d \quad \text{converges to 0 as } |d| \rightarrow +\infty. \tag{2.1}$$

In the context of Theorem 1 (resp. Theorem 2, resp. Theorem 3) such bases are taken to consist of non-constant harmonic polynomials (resp. Maass forms and Eisenstein series of weight 0, resp. Maass forms and Eisenstein series of non-negative, even, weight).

2.1. Duke’s proof. The decay of the period integral $W(\varphi, d)$ is achieved by realizing it in terms of the d -th Fourier coefficient of a modular form of half-integral weight and level 4; this modular form – call it $\tilde{\varphi}$ – is obtained from φ through a theta correspondance.

In the case of Theorem 1, and when φ is a non-constant harmonic polynomial of degree r , this comes from the well known fact that the theta-series

$$\tilde{\varphi}(z) = \theta_\varphi(z) = \sum_{|d| \geq 1} \left(\sum_{\substack{(a,b,c) \in \mathbb{Z}^3 \\ a^2+b^2+c^2=|d|}} \varphi(a, b, c) \right) e(|d|z)$$

is a modular form of weight $k = 3/2 + r$ for the modular group $\Gamma_0(4)$. This is a special case of a (theta) correspondance of Maass, which itself is now a special case of the theta correspondance for dual pairs; it associates to an automorphic form φ for an orthogonal group $SO_{p,q}$ of signature (p, q) , a Maass form $\tilde{\varphi}$ of weight $(q - p)/2$. Moreover, Maass provided a formula expressing the Fourier coefficients of $\tilde{\varphi}$ in terms of a certain integral of φ .

By the accidental isomorphisms recalled above, this provides a correspondance between automorphic forms either for $B_{2,\infty}^\times$ or for PGL_2 , and modular forms of half-integral weight. Under this correspondance, one has, for d a fundamental discriminant

$$W(\varphi, d) = c_{\varphi,d} \frac{\rho_{\tilde{\varphi}}(d) |d|^{-1/4}}{L(\chi_d, 1)} \tag{2.2}$$

where $c_{\varphi,d}$ is a constant depending on φ and mildly on d (i.e. $|d|^{-\varepsilon} \ll_{\varphi,\varepsilon} c_{\varphi,d} \ll_\varepsilon |d|^\varepsilon$ for any $\varepsilon > 0$), $\rho_{\tilde{\varphi}}(d)$ denotes the suitably normalized d -th Fourier coefficient of $\tilde{\varphi}$ and χ_d is the quadratic character corresponding to K_d .

In particular, by Siegel’s lower bound $L(\chi_d, 1) \gg_\varepsilon |d|^{-\varepsilon}$, (2.1) is a consequence of a bound of the form

$$\rho_{\tilde{\varphi}}(d) \ll |d|^{1/4-\delta} \tag{2.3}$$

for some absolute $\delta > 0$. The bound (2.3) is to be expected; indeed the half-integral weight analog of the Ramanujan–Petersson conjecture predicts that any $\delta < 1/4$ is admissible. This conjecture follows from the GRH, but, unlike its integral weight analogue, does not follow from the Weil conjectures.

The problem of bounding Fourier coefficient of modular forms can be approached through a Petersson–Kuznetsov type formula (due to Proskurin in the half-integral weight case): (un)fortunately the standard bound for the Salié sums occurring in the formula yield the above estimate only for $\delta < 0$. This “barricade” was eventually surmounted by Iwaniec (using an ingenious idea of averaging over the level, and obtaining the value $\delta = 1/28$, [41]) for $\tilde{\varphi}$ a holomorphic form of weight $\geq 5/2$ and by B. Duke for general forms by adapting Iwaniec’s argument, and thus concluding the first fully unconditional proof of Theorems 1, 2, 3.

2.2. Equidistribution and subconvex bounds for L -functions. Shortly after Duke’s proof, another approach emerged which turned out to be very fruitful, namely the connection between the decay of Weyl’s sums (2.1) and the *subconvexity problem* for automorphic L -function (see Section 3).

2.2.1. Weyl’s sums as period integrals: Waldspurger type formulae. It goes back to Gauss that the set of classes of quadratic forms $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathcal{Q}_d(\mathbb{Z})$ has the structure of a finite commutative group (the class group) $\mathrm{Cl}(d)$. In particular for the second problem ($d < 0$), \mathcal{H}_d is a homogeneous space under the action of $\mathrm{Cl}(d)$ and the Weyl sums can be seen as period integrals for this action:

$$W(\varphi, d) = \int_{\mathrm{Cl}(d)} \varphi(\sigma.z_d) d\mu_{\mathrm{Haar}}(\sigma).$$

In a similar way, the Weyl’s sums over \mathcal{G}_d and Γ_d can be realized as orbital integrals for the action of some class group. The connection between such orbital integrals and L -functions follows from a formula basically due to Waldspurger. To describe it in greater detail it is useful and convenient to switch an adelic description of the Weyl’s sums.

Let us recall that in the context of Theorem 1 with $Q(A, B, C) = A^2 + B^2 + C^2$ (resp. Theorems 2 and 3, with $Q(A, B, C) = B^2 - 4AC$) a solution $Q(a, b, c) = d$ gives rise to an embedding of the quadratic \mathbb{Q} -algebra K_d into the \mathbb{Q} -algebra $\mathbb{B}^{(2,\infty)}$ (resp. $M_{2,\mathbb{Q}}$) by sending \sqrt{d} to $a.i + b.j + c.k$ (resp. $\begin{pmatrix} b & -2a \\ 2c & -b \end{pmatrix}$). This yields an embedding of \mathbb{Q} -algebraic groups, $T_d := \mathrm{res}_{K/\mathbb{Q}} \mathbb{G}_m / \mathbb{G}_m \hookrightarrow \mathbf{G}$, where $\mathbf{G} = \mathrm{PG}(\mathbb{B}^{(2,\infty)})$ (resp. $= \mathrm{PGL}_2$).

Let $K_{f,\max}$ be a maximal compact subgroup of $\mathbf{G}(A_f)$ in all three cases. In the context of Theorem 1 (resp. Theorem 2, resp. Theorem 3) take $K_\infty = T_d(\mathbb{R}) \cong \mathrm{SO}_2 \subset \mathbf{G}$ (resp. $K_\infty = T_d(\mathbb{R})$, resp. $K_\infty = \{1\}$) and set $K = K_{f,\max} K_\infty$; the quotient $\mathbf{G}(\mathbb{Q}) \backslash \mathbf{G}(A_\mathbb{Q}) / K$ then equals a quotient of S^2 by a finite group of rotations (resp. $Y_0(1)$, resp. the unit tangent bundle of $Y_0(1)$).

It transpires, with these identifications, the subsets

$$\mathcal{G}_d \subset S^2, \quad \mathcal{H}_d \subset Y_0(1), \quad \Gamma_d \subset S_*^1(Y_0(1))$$

may be uniformly described, after choosing a solution z_d , as a compact orbit of the adelic torus T_d :

$$T_d(\mathbb{Q}) \backslash z_d \cdot T_d(\mathbb{A}_{\mathbb{Q}}) / K_{T_d} \subset G(\mathbb{Q}) \backslash G(\mathbb{A}_{\mathbb{Q}}) / K$$

where $K_{T_d} := T_d(\mathbb{A}_{\mathbb{Q}}) \cap K$. In this notation the Weyl sum is given as a *toric integral*

$$W(\varphi, d) = \int_{T_d(\mathbb{Q}) \backslash T_d(\mathbb{A}_{\mathbb{Q}}) / K_{T_d}} \varphi(z_d \cdot t) dt \tag{2.4}$$

when dt is the Haar measure on the toric quotient. A superficial advantage of this notation is that it allows for a uniform presentation of many equidistribution problems for “cycles” associated with quadratic orders in locally symmetric spaces associated to quaternion algebras. Indeed, as we shall see below, one can consider the above equidistribution problems while changing

- the group G to $G = B^\times / Z(B^\times)$ for B any quaternion algebra over \mathbb{Q} ;
- the compact $K_{f,\max}$ to a compact subgroup $K'_f \subset K_{f,\max}$ (i.e. changing the level structure)
- the subgroup K_{T_d} to a subgroup K'_{T_d} (i.e. considering cycles associated to suborders \mathcal{O} of the maximal order \mathcal{O}_d).
- the base field \mathbb{Q} to a fixed totally real number field F .

When φ is a *new cuspform* (the L^2 normalized *new vector* in some automorphic representation π), Waldspurger’s formula [76] relates $|W(\varphi, d)|^2$ (and correspondingly the square of the d -th Fourier coefficient $|\rho_{\tilde{\varphi}}(d)|^2$) to the central value of an automorphic L -function. In its original form, the formula was given up to some non-zero proportionality constant; as we are interested in the size $W(\varphi, d)$ a more precise expression is needed. Thanks to the work of many people ([12], [36], [37], [47], [50], [66], [78], [80], [81]) notably Gross, Zagier and Zhang such an expression is by now available in considerable generality. Under suitable hypotheses (which in the present cases are satisfied), it has the following form

$$|W(\varphi, d)|^2 = c_{\varphi,d} \frac{L(\pi, 1/2)L(\pi \times \chi_d, 1/2)}{L(\chi_d, 1)^2 \sqrt{|d|}} \tag{2.5}$$

where π' is a GL_2 -automorphic representation corresponding to π by the Jacquet–Langlands correspondance and $c_{\varphi,d} > 0$ is a constant which depends mildly on d .

The Waldspurger formula (2.5) is more powerful than (2.2) as it may be extended to a formula for more general toric integrals. Indeed, let χ be a character of the torus $T_d(\mathbb{Q}) \backslash T_d(\mathbb{A}_{\mathbb{Q}}) = K_d^\times \mathbb{A}_{\mathbb{Q}}^\times \backslash \mathbb{A}_{K_d}^\times$ trivial on K_{T_d} . Under suitable compatibility assumptions between χ and φ and possibly under additional coprimality assumptions between the conductors of π, χ , the relation (2.5) generalizes to

$$|W_\chi(\varphi, d)|^2 = c_{\varphi,d_\chi,\chi_\infty} \frac{L(\pi \times \pi_\chi, 1/2)}{L(\chi_d, 1)^2 \sqrt{|d_\chi|}} \tag{2.6}$$

where $W_\chi(\varphi, d)$ is a *twisted toric integral* of the form

$$W_\chi(\varphi, d_\chi) = \int_{\mathbf{T}_d(\mathbb{Q}) \backslash \mathbf{T}_d(\mathbf{A}_\mathbb{Q})} \chi(t) \varphi(z_{d_\chi} \cdot t) dt,$$

π_χ is the GL_2 -automorphic representation (of conductor d_χ) corresponding to χ by quadratic automorphic induction and $L(\pi \times \pi_\chi, s)$ is the Rankin–Selberg L -function of the pair (π, π_χ) .

2.3. Subconvexity and (sparse) equidistribution. We see, from formula (2.5) and Siegel’s lower bound that (2.1) follows from the bound

$$L(\pi \times \chi_d, 1/2) \ll_\pi |d|^{1/2-\delta} \tag{2.7}$$

for some absolute $\delta > 0$; subject to this bound, one obtains another proof of Linnik’s equidistribution problems. More generally, we see from (2.6) that the twisted Weyl sums are decaying, i.e.

$$W_\chi(\varphi, d_\chi) \rightarrow 0 \quad \text{for } d_\chi \rightarrow +\infty, \tag{2.8}$$

as soon as

$$L(\pi \times \pi_\chi, 1/2) \ll |d_\chi|^{1/2-\delta}. \tag{2.9}$$

Both (2.7) and (2.9) are special cases of subconvex bounds for central values of automorphic L -functions and have been proven (see below).

One should note that the decay of the twisted toric integral is useful if one needs to perform *harmonic analysis along* the toric orbit $\mathbf{T}_d(\mathbb{Q}) \backslash z_d \cdot \mathbf{T}_d(\mathbf{A}_\mathbb{Q}) / \mathbf{K}_{\mathbf{T}_d}$: this is particular the case when one needs equidistribution only for a strictly smaller suborbit of the full orbit, a problem we call a *sparse equidistribution problem*.

For instance one has:

Theorem 5 ([39]). *There is an absolute constant $0 < \eta < 1$ such that: for each fundamental discriminant $d < 0$, choose $z_{0,d} \in \mathcal{H}_d$ a Heegner point and choose G_d a subgroup of $\mathrm{Cl}(d)$ of size $|G_d| \geq |\mathrm{Cl}(d)|^\eta$ then the sequence of suborbits*

$$\mathcal{H}_d^\ell := G_d \cdot z_{0,d} = \{\sigma \cdot z_{0,d}, \sigma \in G_d\}$$

is equidistributed on $Y_0(1)$ w.r.t. μ_P .

One has also similar sparse equidistribution results for sufficiently large suborbits of \mathcal{G}_d on the sphere and for sufficiently large geodesic segments of Γ_d [60], [66]. Note however that the present method has fundamental limitations as one cannot take η too close to 0: even under the GRH, one would prove equidistribution only for $\eta > 1/2$. Nevertheless we would like to formulate the following

Conjecture 1 (Equidistribution of subgroups). Fix any $\eta > 0$ and for each fundamental discriminant $d < 0$, choose $z_{0,d} \in \mathcal{H}_d$ a Heegner point and choose G_d a subgroup of $\text{Cl}(d)$ of size $|G_d| \geq |d|^\eta$. Then as $|d| \rightarrow +\infty$, the sequence of suborbits

$$\mathcal{H}_d^l := G_d \cdot z_{0,d} = \{\sigma \cdot z_{0,d}, \sigma \in G_d\}$$

is equidistributed on $Y_0(1)$ w.r.t. μ_P .

This conjecture is certainly difficult in general; however, we expect that, by ergodic methods like the ones described in Section 6, significant progress might be made, at least for subgroups G_d that satisfy suitable versions of Linnik’s condition for some fixed prime p .

2.4. Equidistribution and non-vanishing of L -functions. Before continuing with the subconvexity problem, we would like to point out another interesting application. It combines subconvexity, equidistribution and the period relation (2.5) and applies them to the non-vanishing of L -functions.

Consider, for simplicity, the context of Theorem 2 (see also [62]): let φ be a Maass–Hecke eigenform of weight 0 and π be its associated automorphic representation. If one averages (2.5) over the characters of $\text{Cl}(d)$, one obtains by orthogonality (here the constants $c_{\varphi,d_\chi,\chi_\infty}$ are equal to an absolute constant $c > 0$)

$$c \frac{\sqrt{d}}{|\text{Cl}(d)|^2} \sum_{\chi \in \widehat{\text{Cl}}(d)} L(\pi \times \pi_\chi, 1/2) = \int_{Y_0(1)} |\varphi|^2 \cdot \mu_d$$

and since by Theorem 2

$$\int_{Y_0(1)} |\varphi|^2 \cdot \mu_d \rightarrow \int_{Y_0(1)} |\varphi(z)|^2 d\mu_P(z) > 0 \quad \text{as } d \rightarrow -\infty$$

this shows that for some χ the central value $L(\pi \times \pi_\chi, 1/2)$ does not vanish. Moreover, by the subconvex bound (2.9), one obtains a quantitative form of non-vanishing

$$|\{\chi \in \widehat{\text{Cl}}(d), L(\pi \times \pi_\chi, 1/2) \neq 0\}| \gg |d|^\eta \tag{2.10}$$

for some absolute $\eta > 0$.

Remark 2.1. When π corresponds to an Eisenstein series, stronger results were obtained before by Duke–Friedlander–Iwaniec [24] and Blomer [5]; although this it appears in a somewhat disguised (and more elaborate) form, the basic principle underlying the proof is the same.

By considering equidistribution relative to definite quaternion algebras, one can obtain similar non-vanishing results for central values $L(\pi \times \pi_\chi, 1/2)$ where π_∞ is in the discrete series and the sign of the functional equation of $L(\pi \times \pi_\chi, s)$ is $+1$. In particular when $\pi = \pi_E$ is the automorphic representation associated to an elliptic

curve E/\mathbb{Q} , such estimates provide a lower bound for the size of the “rank-0” part of the group $E(H_K)$ of points of E which are rational over the Hilbert class field of K as $d \rightarrow -\infty$.

An interesting problem is to address the case where the sign of the functional equation is -1 . In this case, $L(\pi \times \pi_\chi, 1/2) = 0$ and one considers instead the question of non-vanishing of the first derivative $L'(\pi \times \pi_\chi, 1/2)$. At least when π_∞ is in the holomorphic discrete series and π has trivial central character, the Gross–Zagier formula (and its extensions by Zhang) interprets $L'(\pi \times \pi_\chi, 1/2)$ as the “height” of some Heegner cycle above some modular (or Shimura) curve. This is not quite a period integral; however the height decomposes as a sum of local heights indexed by the places v of \mathbb{Q} . These local heights are either simple or can be interpreted as periods integrals over quadratic cycles associated with K which live over appropriate adelic quotients $\mathbf{G}^{(v)}(\mathbb{Q}) \backslash \mathbf{G}^{(v)}(\mathbf{A}) / \mathbf{K}_v$ where $\mathbf{G}^{(v)}$ is associated to a quaternion algebra $\mathbf{B}^{(v)}$ ramified at v .

It seems then plausible that one can compute the asymptotic of the average $\sum_\chi L'(\pi \times \pi_\chi, 1/2)$ by using the equidistribution property of quadratic cycles on these infinitely many quotients. One consequence of this would then be, for compatible E and K , a lower bound for the rank of $E(H_K)$:

$$\text{rank}_{\mathbb{Z}} E(H_K) \gg |d|^\eta$$

for some $\eta > 0$ as $d \rightarrow -\infty$.

Remark 2.2. A few years ago, Vatsal and Cornut [16], [73], [74] used period relations and equidistribution in a similar way to obtain somewhat stronger non-vanishing results for Rankin–Selberg L -functions but associated to anti-cyclotomic² characters of a *fixed* imaginary quadratic field. Note that one of their main ingredient to obtain equidistribution came from ergodic theory and precisely from Ratner’s theory.

3. The subconvexity problem

Although the subconvexity problem is a venerable topic in number theory – its study begins with Weyl’s estimate $|\zeta(1/2 + it)| \ll_\varepsilon t^{1/6+\varepsilon}$ – there has been a renaissance of interest in it recently. This owes largely to the observation that a resolution of the subconvexity problem for automorphic L -functions on GL has many striking applications, as we have just seen to Linnik’s equidistribution problems or to “Arithmetic Quantum Chaos.” We refer to [44] for a discussion of all these questions in the broader context of the analytic theory of automorphic L -functions.

Let $\Pi = \Pi_\infty \otimes \bigotimes_p' \Pi_p$ some reasonable “automorphic object”: by automorphic object we mean, for instance an automorphic representation or more generally an admissible representation constructed out of automorphic representations via the

²The case of cyclotomic characters was carried out even earlier by Rohrlich, by more direct methods.

formalism of L -groups (for instance the Rankin–Selberg convolution $\pi_1 \times \pi_2$ of two automorphic representations on some linear groups). To Π , one can usually associate a collection of local L -factors

$$L(\Pi_p, s) = \prod_{i=1}^d \left(1 - \frac{\alpha_{\Pi,i}(p)}{p^s}\right)^{-1}, \quad p \text{ prime}, \quad L(\Pi_\infty, s) = \prod_{i=1}^d \Gamma_{\mathbb{R}}(s - \mu_{\Pi,i})$$

where $\Gamma_{\mathbb{R}}(s) = \pi^{-s/2}\Gamma(s/2)$ and $\{\alpha_{\Pi,i}(p)\}, \{\mu_{\Pi,i}\}$ are called the local numerical parameters of Π at p and at infinity; from these local datas one forms a global L -function

$$L(\Pi, s) = \sum_{n \geq 1} \frac{\lambda_\Pi(n)}{n^s} = \prod_p L(\Pi_p, s).$$

In favourable cases, one can show that $L(\Pi, s)$ has analytic continuation to \mathbb{C} and satisfies a functional equation which we normalize into the form

$$q_\Pi^{s/2} L(\Pi_\infty, s) L(\Pi, s) = w_\Pi q_\Pi^{(1-s)/2} \overline{L(\Pi_\infty, 1 - \bar{s}) L(\Pi, 1 - \bar{s})},$$

where $|w_\Pi| = 1$ and $q_\Pi > 0$ is an integer called the conductor of Π . We recall (after Iwaniec–Sarnak [44]) that *the analytic conductor* of Π is the function of the complex variable s given by

$$C(\Pi, s) = q_\Pi \prod_{i=1}^d |s - \mu_{\Pi,i}|.$$

It is expected, and known in many cases, that the following *convexity bound* for the values of $L(\Pi, s)$ holds on the critical line $\Re s = 1/2$: for any $\varepsilon > 0$, one has

$$L(\Pi, s) \ll_{\varepsilon,d} C(\Pi, s)^{1/4+\varepsilon}.$$

This is known, in particular, when Π is an automorphic cuspidal representation of $\mathrm{GL}(n)$ over any number field, [63]. The Lindelöf conjecture, which is a consequence of the GRH, asserts that in fact $L(\Pi, s) \ll_{\varepsilon,d} C(\Pi, s)^\varepsilon$. In many applications, however, it is sufficient to improve the convexity bound.

The subconvexity problem consists in improving the exponent $1/4$ to $1/4 - \delta$ for some positive absolute δ . In fact, for most applications it is sufficient to improve that exponent only with respect to one of the three type of parameters s , q_Π or $\prod_{i=1}^d (1 + |\mu_{\Pi,i}|)$; these variants are called the s -aspect, the q -aspect (or *level*-aspect) and the ∞ -aspect (or *eigenvalue*-aspect) respectively. See [34], [44] for an introduction to the subconvexity problem in this generality. During the last decade, there has been considerable progress on the subconvexity problem for L -functions associated to GL_1 and GL_2 automorphic forms. In this lecture, we mainly discuss the recent progress made on the q -aspect, although the other aspects are very interesting, both for applications and for conceptual reasons (see [6], [42], [46], [70]). In the level aspect, one has

Theorem 6. *Let F be a fixed number field and π_2 be a fixed cuspidal automorphic representation of $\mathrm{GL}_2(\mathbf{A}_F)$. Let χ_1, π_1 denote respectively a $\mathrm{GL}_1(\mathbf{A}_F)$ -automorphic representation (i.e. a Grössencharacter), a $\mathrm{GL}_2(\mathbf{A}_F)$ -automorphic representation and let q_1 denote either the conductor of χ_1 or π_1 and $q_1 = N_{F/\mathbb{Q}}(q_1)$. There exists an absolute constant $\delta > 0$ such that for $\Re s = 1/2$ one has*

$$L(\chi_1, s) \ll_s q_1^{1/4-\delta}, \quad (3.1)$$

$$L(\chi_1 \times \pi_2, s) \ll_{s, \pi_2, \chi_1, \infty} q_1^{1/2-\delta}, \quad (3.2)$$

$$L(\pi_1, s) \ll_{s, \pi_1, \infty} q_1^{1/4-\delta}, \quad (3.3)$$

$$L(\pi_1 \times \pi_2, s) \ll_{s, \pi_2, \pi_1, \infty} q_1^{1/2-\delta}. \quad (3.4)$$

Thus the subconvexity problem is solved in the q_1 -aspect for all these L -functions.

- For $F = \mathbb{Q}$, the bound for Dirichlet L -functions (3.1) is due to Burgess (see also [15]). The bound for twisted L -function (3.2) is basically due to Duke–Friedlander–Iwaniec [19] (see also [7], [9] for the general bound over \mathbb{Q} with a good subconvex exponent). The bound (3.3) is mainly to a series of works by Duke–Friedlander–Iwaniec: [20] for π_1 with trivial central character and [21], [22], [23] for the much harder case of a central character of conductor q_1 ; it has been recently completed for π_1 with arbitrary central character by Blomer, Harcos and the first author in [8]. The bound for Rankin–Selberg L -functions (3.4) for π having trivial central character is due to Kowalski, the first author and Vanderkam ([52]) by generalizing the methods of [20] and to Harcos and the first author for π_1 with an arbitrary central character [39], [60].

- In the case of a number field of higher degree, the first general subconvex result is due to Cogdell–Piatetski-Shapiro–Sarnak [13]: it consists of (3.2) when F is a totally real field and $\pi_{2, \infty}$ is in the holomorphic discrete series (i.e. corresponds to a holomorphic Hilbert modular form). Recently, the second author developed a new method and established, amongst other things, the bounds (3.1), (3.2), (3.3) and (3.4) for F an arbitrary number field, π_2 fixed but arbitrary and π_1 with a trivial central character [75]. Eventually the authors combined their respective methods from [60] and [75] to obtain (3.3) and (3.4) for π_1 with an arbitrary central character.

3.1. Amplification and the shifted convolution problem. Arguably, the most successful approach to subconvexity in the q -aspect is via the method of moments or more precisely via its variant, the *amplification method*. For the sake of completeness we briefly recall the mechanism and refer to [34] and [43] for the philosophy underlying this method.

Given Π_1 and a (well chosen) family of automorphic objects $\mathcal{F} = \{\Pi\}$ containing Π_1 , the amplification method builds on the possibility to obtain a bound for the amplified k -th moment of the $\{L(\Pi, s), \Pi \in \mathcal{F}\}$, $\Re s = 1/2$, of the form

$$\sum_{\Pi \in \mathcal{F}} |L(\Pi, s)|^k \left| \sum_{\ell \leq L} \lambda_{\Pi}(\ell) a_{\ell} \right|^2 \ll_{\varepsilon} |\mathcal{F}|^{1+\varepsilon} \sum_{\ell \leq L} |a_{\ell}|^2 \quad (3.5)$$

for any $\varepsilon > 0$, where the $(a_\ell)_{\ell \leq L}$ are *a priori* arbitrary complex coefficients and where L is some positive power of $|\mathcal{F}|$. Such a bound is expected if L is sufficiently small compared with $|\mathcal{F}|$, since the individual bound $|L(\Pi, s)|^k \ll_\varepsilon |\mathcal{F}|^\varepsilon$ would follow from the GRH and the estimate

$$\sum_{\Pi \in \mathcal{F}} \left| \sum_{\ell \leq L} \lambda_\Pi(\ell) a_\ell \right|^2 = |\mathcal{F}|(1 + o(1)) \sum_{\ell \leq L} |a_\ell|^2$$

should be a manifestation of the *quasi-orthogonality* of the $\{(\lambda_\Pi(\ell))_{\ell \leq L}\}_{\Pi \in \mathcal{F}}$ which is a frequently recurring theme in harmonic analysis. Assuming (3.5), one deduces a subconvex bound for $|L(\Pi_1, s)|^k$ by restricting (3.5) to one term and by choosing the coefficients $a_\ell = a_\ell(\Pi_1)$ appropriately.

All the subconvex bounds presented in Theorem 6 can be obtained by considering for $L(\Pi, s)$ an L -function of *Rankin–Selberg* type, i.e. an L -function either of the form $L(\chi_1 \times \pi_2, s)$ or of the form $L(\pi_1 \times \pi_2, s)$ with π_2 a fixed (not necessarily cuspidal) GL_2 -automorphic representation. The families \mathcal{F} considered are then essentially of the form $\{\chi \times \pi_2, q_\chi = q_1\}$ or $\{\pi \times \pi_2, q_\pi = q_1, \omega_\pi = \omega_{\pi_1}\}$ and the bound (3.5) is achieved for the second moment ($k = 2$). To analyze effectively the left-hand side of (3.5) one needs a manageable expression for $L(\Pi, s)$ for s on the critical line. The traditional method to do so is to apply an *approximate functional equation* technique which expresses $L(\Pi, s)$ essentially as a partial sum of the form

$$\Sigma(\Pi) := \sum_{n \geq 1} \frac{\lambda_\Pi(n)}{n^s} W\left(\frac{n}{\sqrt{q_\Pi}}\right)$$

with W a rapidly decreasing function (which depends on s and on Π_∞). In the context of Theorem 6, the second amplified moment (3.5) are then computed and transformed by spectral methods. These involve, in particular, the orthogonality relations for characters and the Kuznetsov–Petersson formula. These computations reduce the subconvex estimates to the problem of estimating non-trivially sums of the form

$$\Sigma_\pm(\varphi_2, \ell_1, \ell_2, h) := \sum_{\ell_1 m \pm \ell_2 n = h} \overline{\rho_{\varphi_2}(m)} \rho_{\varphi_2}(n) \mathcal{W}\left(\frac{m}{q}, \frac{n}{q}\right), \tag{3.6}$$

the trivial bound being $\ll_{\varphi_2} q^{1+o(1)}$; here $h = O(q)$ is a non-zero integer, $\rho_{\varphi_2}(n)$ denotes the n -th Fourier coefficient of some automorphic form φ_2 in the representation space of π_2 , $\mathcal{W}(x, y)$ is a rapidly decreasing function and $\ell_1, \ell_2 \leq L$ are the parameters occurring as indices of the amplifier $(a_\ell)_{\ell \leq L}$. These sums are classical in analytic number theory and are called *shifted convolution sums*; the problem of estimating them non-trivially for various ranges of h, m, n is called a *shifted convolution problem*.³

³Historically, the shifted convolution problem already occurred in the work of Kloosterman on the number of representations of an integer n by the quadratic form $a_1.x^2 + a_2.y^2 + a_3.z^2 + a_4.t^2$, and also in Ingham’s work on the additive divisor problem. In Kloosterman’s case φ_2 is a theta-series of weight 1, whereas in Ingham’s case φ_2 is the standard non-holomorphic Eisenstein series.

3.2. Shifted convolutions via the circle method. In order to solve a shifted convolution problem, one needs an analytically manageable expression of the linear constraint $\ell_1 m \pm \ell_2 n = h$; one is to suitably decompose the integral

$$\delta_{\ell_1 m \pm \ell_2 n - h = 0} = \int_{\mathbb{R}/\mathbb{Z}} \exp(2\pi i(\ell_1 m \pm \ell_2 n - h)\alpha) d\alpha,$$

and there are several methods to achieve this; the first possibility in this context was Kloosterman’s refinement of the circle method; other possibilities are the Δ -symbol method, used in [19] and [20] to prove some cases of (3.2) and (3.3) or Jutila’s method of overleaping intervals which is particularly flexible [38], [45]. These methods provide an expression of the above integral into weighted sums of Ramanujan type sums of the form

$$\sum_{\substack{a \bmod c \\ (a,c)=1}} e\left(\frac{(\ell_1 m \pm \ell_2 n - h) \cdot a}{c}\right)$$

for c ranging over relatively small moduli. Such decomposition makes it possible to essentially “separate” the variable m from n and to reduce $\Sigma(\varphi_2, \ell_1, \ell_2, h)$ to sums over moduli c on additively twisted sums of Fourier coefficients

$$\sum_c \cdots \sum_{\substack{a \bmod c \\ (a,c)=1}} e\left(\frac{-ha}{c}\right) \left(\sum_m \overline{\rho_{\varphi_2}(m)} e\left(\frac{\ell_1 m a}{c}\right) \mathcal{W}\left(\frac{m}{q}\right)\right) \left(\sum_n \rho_{\varphi_2}(n) e\left(\pm \frac{\ell_2 n a}{c}\right) \mathcal{W}\left(\frac{n}{q}\right)\right).$$

The independent m - and n -sums are then transformed via the Voronoï summation formula with the effect of replacing the test functions $\mathcal{W}(\frac{\cdot}{q})$ by some Bessel transform and the additive shift $e(\pm \frac{\ell_2 a \cdot}{c})$ by $e(-\pm \frac{\bar{\ell}_2 \bar{a} \cdot}{c})$ where \bar{a} denotes the multiplicative inverse of $a \bmod c$. After these transformations and after averaging over $a \bmod c$ the sum $\Sigma_{\pm}(\varphi_2, \ell_1, \ell_2, h)$ takes essentially the following form (possibly up to a main term which occurs if φ_2 is an Eisenstein series [18])

$$MT_{\pm}(\varphi_2, \ell_1, \ell_2, h) + \sum_{c \equiv 0(\ell_1 \ell_2 q \pi_0)} \sum_{h'} \left(\sum_{\mp \ell_1 n - \ell_2 m = h'} \alpha_m \overline{\rho_{\varphi_2}(m)} \beta_n \rho_{\varphi_2}(n) \right) \text{Kl}(-h, h'; c) \mathcal{V}(h, h'; c) \tag{3.7}$$

where $MT_{\pm}(\varphi_2, \ell_1, \ell_2, h)$ is non-zero only if φ_2 is an Eisenstein series (in which case it is a main term of size $\approx_{\ell_1, \ell_2, \varphi_2} q^{1+o(1)}$), α_m, β_n are smooth coefficients, $\text{Kl}(-h, h'; c)$ is a Kloosterman sum and \mathcal{V} is a smooth function. Eventually, Weil’s bound for Kloosterman sums

$$\text{Kl}(-h, h'; c) \ll (h, h', c)^{1/2} c^{1/2+o(1)}$$

gives the formula

$$\Sigma_{\pm}(\varphi_2, \ell_1, \ell_2, h) = MT_{\pm}(\varphi_2, \ell_1, \ell_2, h) + O_{\varphi_2, \varepsilon}((\ell_1 \ell_2)^A q^{3/4+\varepsilon}) \tag{3.8}$$

for some absolute constant A . Finally, from (3.8) one can deduce (3.2), (3.3), (3.4) when π has trivial central character although the derivation may be quite delicate if φ_2 is an Eisenstein series (cf. [20] and see also [51]).

3.3. Shifted convolutions and spectral theory. In [70], Sarnak, inspired by ideas of Selberg, developed a purely spectral approach to the shifted convolution sums (3.6) (previously some special cases have been treated by others, for instance by A. Good). This method, which at present has been entirely worked out when φ_2 is a classical holomorphic cuspform (say of weight $k \geq 2$ and level q_2), is based on the analytic properties of the Dirichlet series

$$D(\varphi_2, \ell_1, \ell_2, h, s) = \sum_{\substack{m, n \geq 1 \\ \ell_1 m - \ell_2 n = h}} \frac{\overline{\rho_{\varphi_2}(m)} \rho_{\varphi_2}(n)}{(\ell_1 m + \ell_2 n)^s} \left(\frac{\sqrt{\ell_1 \ell_2 m n}}{\ell_1 m + \ell_2 n} \right)^{k-1}.$$

Note that for $h = 0$ this series is essentially a Rankin–Selberg L -function. As in the Rankin–Selberg case, the analytic properties of D follows from an appropriate integral representation in the form of a triple product integral; however, for $h \neq 0$ one needs to replace the Eisenstein series by a Poincaré series. Precisely, one has $D(s) = (2\pi)^{s+k-1} (\ell_1 \ell_2)^{1/2} \Gamma^{-1}(s+k-1) I(s)$ with

$$\begin{aligned} I(\varphi_2, \ell_1, \ell_2, h, s) &:= ((\ell_1 y)^{k/2} \varphi_2(\ell_1 z) \cdot (\ell_2 y)^{k/2} \overline{\varphi_2}(\ell_2 z), P_h(z, s)) \\ &= \int_{\Gamma_0(q_2 \ell_1 \ell_2) \backslash \mathbb{H}} (\ell_1 y)^{k/2} \overline{\varphi_2}(\ell_1 z) \cdot (\ell_2 y)^{k/2} \varphi_2(\ell_2 z) P_h(z, s) \frac{dx dy}{y^2} \end{aligned}$$

where $P_h(z, s)$ is a non-holomorphic Poincaré series of weight 0. The analytic continuation for D follows from that of $P_h(\cdot, s)$; in particular, from its spectral expansion one deduce that the latter is absolutely convergent for $\Re s > 1$ and has holomorphic continuation in the half-plane $\Re s > 1/2 + \theta$ where θ measures the quality of available results towards the Ramanujan–Petersson conjecture:

Hypothesis H_{θ} . For any cuspidal automorphic form π on $\text{GL}_2(\mathbb{Q}) \backslash \text{GL}_2(\mathbb{A}_{\mathbb{Q}})$ with local Hecke parameters $\{\alpha_{\pi, i}(p), i = 1, 2\}$ for $p < \infty$ and $\{\mu_{\pi, i}, i = 1, 2\}$ one has the bounds

$$\begin{aligned} |\alpha_{\pi, i}(p)| &\leq p^{\theta}, \quad i = 1, 2, \\ |\Re \mu_{\pi, i}| &\leq \theta, \quad i = 1, 2, \end{aligned}$$

provided π_p, π_{∞} are unramified, respectively.

Remark 3.1. Hypothesis H_{θ} is known for $\theta > 3/26$ thanks to the works of Kim and Shahidi [48], [49].

A bound for $D(s)$ in a non-trivial domain is deduced from the spectral expansion of the inner product $I(s)$ over an suitable orthonormal basis of Maass forms, $\{\psi\}$ say, and of Eisenstein series of weight 0 and level $\ell_1\ell_2q_0$: one has

$$\sum_{\psi} \langle (\ell_1 y)^{k/2} \varphi_2(\ell_1 z) \cdot (\ell_2 y)^{k/2} \overline{\varphi_2(\ell_2 \bar{z})}, \psi \rangle \langle \psi, P_h(z, s) \rangle + \text{Eisenstein spectrum}. \quad (3.9)$$

For ψ in the cuspidal basis, let it_{ψ} denote the archimedean parameter $\mu_{\pi,1}$ of the representation π containing ψ ; the second inner product $\langle \psi, P_h(z, s) \rangle$ equals the Fourier coefficient of ψ , $\overline{\rho_{\psi}(-h)}$ times a factor bounded by $(1 + |t_{\psi}|)^B e^{\frac{\pi}{2}|t_{\psi}|}$. The Fourier coefficient $\overline{\rho_{\psi}(-h)}$ is bounded by $O(|h|^{\theta+o(1)})$ by Hypothesis H_{θ} at the non-archimedean places. The problem now, as was pointed out by Selberg, is to have a bound for the triple product integral $\langle (\ell_1 y)^{k/2} \varphi_2(\ell_1 z) \cdot (\ell_2 y)^{k/2} \overline{\varphi_2(\ell_2 \bar{z})}, \psi \rangle$ which exhibits an exponential decay for the form $O((1 + |t_{\psi}|)^C e^{-\frac{\pi}{2}|t_{\psi}|})$, so as to compensate the exponential growth of $\langle \psi, P_h(z, s) \rangle$. In this generality, this exponential decay property for triple product was achieved by Sarnak in [69]; later, a representation theoretic version of Sarnak's arguments as well as some improvements were given by Bernstein–Reznikov [2]. The final consequence of these bounds is the following estimate

$$\Sigma_{-}(\varphi_2, \ell_1, \ell_2, h) = O_{\varphi_2, \varepsilon}((\ell_1 \ell_2)^A q^{1/2+\theta+\varepsilon}). \quad (3.10)$$

This approach is important for several reasons:

- It ties more closely the subconvexity problem for GL_2 L -functions – a problem whose origin lies in analytic number theory – to the Ramanujan–Petersson conjecture for $\text{GL}_2(\mathbf{A}_{\mathbb{Q}})$; or, in other words, to the spectral gap property which is a classical problem in the harmonic analysis of groups;
- it gives an hint that automorphic period integrals might be useful in the study of the subconvexity problem: this will be largely confirmed in Section 4.
- This approach is sufficiently smooth that it can be extended to number fields of higher degree: a few years ago, Cogdell–Piatetski-Shapiro–Sarnak used the amplification method in conjunction with this approach to obtain (3.2) when F is totally real and π_{∞} is a holomorphic discrete series (see [13]).

Remark 3.2. The methods of sections 3.2 and 3.3 are closely related. This can be seen already by remarking that Weil's bound for Kloosterman sums yield the saving $q^{3/4+\varepsilon}$ in (3.8) which is precisely the saving following from Hypothesis $H_{1/4}$ in (3.10); moreover $H_{1/4}$ (a.k.a the Selberg–Gelbart–Jacquet bound) can be obtained by applying Weil's bound to the Kloosterman sums. One can push this coincidence further, by applying, in (3.7) the Kuznetsov–Petersson formula *backwards* in order to transform the sums of Kloosterman sums into sums of Fourier coefficients of Maass

forms:

$$(3.7) = \sum_{\psi} \sum_{h'} \left(\sum_{\mp \ell_1 n - \ell_2 m = h'} \alpha_m \overline{\rho_{\varphi_2}(m)} \beta_n \rho_{\varphi_2}(n) \right) \overline{\rho_{\psi}}(-h) \rho_{\psi}(h') \tilde{\mathcal{V}}(h, h', it_{\psi})$$

+ Discrete series Spectrum + Eisenstein spectrum. (3.11)

Thus, we have realized the spectral expansion of the shifted convolution sum $\Sigma_{\pm}(\varphi_2, \ell_1, \ell_2, h)$ in a way similar to that obtained in (3.9); from there, we may use again the full force of spectral theory. This may look like a rather circuitous path to obtain the spectral expansion; this method however has some technical advantage over the method discussed in Section 3.3: it works even if φ_2 is a Maass form, without the need to find appropriate test vector or to obtain exponential decay for triple product integrals! The spectral decomposition (3.11) will be very useful in the next section.

3.4. The case of a varying central character. The methods discussed so far are sufficient to establish (3.2), (3.3), (3.4) as long as the conductor of the central character, ω_1 say, of π_1 is significantly smaller than q_1 . The case of a varying central character reveals new interesting features which we discuss here. To simplify, we consider the extremal (in a sense hardest) case where both conductors are equal $q_{\omega_1} = q_1 =: q$.

3.4.1. Subconvexity via bilinear Kloosterman fractions. As usual for the subconvexity problem, the first result is due to Duke–Friedlander–Iwaniec for the case (3.3) [22], [23]. As pointed out above, the problem of bounding $L(\pi_1, s)$ for $\Re s = 1/2$ may be formulated as the problem of bounding a Rankin–Selberg L -function

$$L(\pi_1 \times \pi_2, s) = L(\pi_1, s)^2$$

where $\pi_2 = 1 \boxplus 1$ is the representation corresponding to the the fully unramified Eisenstein series. Eventually, another approach was considered in [22], [23], which comes from the identity

$$|L(\pi_1, s)|^2 = L(\pi_1 \times \bar{\chi}, 1/2) L(\tilde{\pi}_1 \times \chi, 1/2)$$

where $\tilde{\pi}_1$ is the contragredient and $\chi = \omega_1 | \cdot |^{-it}$, $t = \Im s$. The amplification method applied to the family $\{L(\pi \times \bar{\chi}, 1/2) L(\tilde{\pi} \times \chi, 1/2), q_{\pi} = q_1 := q, \omega_{\pi} = \omega_1\}$ yields in practice to shifted convolution sums of the form ([22], [23])

$$\sum_{\ell_1 ad - \ell_2 bc = h} \bar{\chi}(a) \chi(c) \mathcal{W}\left(\frac{a}{q^{1/2}}, \frac{b}{q^{1/2}}, \frac{c}{q^{1/2}}, \frac{d}{q^{1/2}}\right),$$

with $h \approx q$, $h \equiv 0(q)$. The later is essentially a truncated version of the shifted convolution sums associated to the Eisenstein series $E(1, \chi)$ of the representation $1 \boxplus \chi$; the new feature by comparison with the previous shifted convolution problems

is that the coefficients $\rho_{E(1,\chi)}(n) = \sum_{bc=n} \chi(c)$ vary with q , which is essentially the range of the variables $m = ad$ and $n = bc$. Since χ has conductor q and a, c vary in ranges of size $\approx q^{1/2}$ one cannot really use the arithmetical structure of the weights $\bar{\chi}(a), \chi(c)$ so this shifted convolution problem is basically reduced to the non-trivial evaluation of a quite general sum:

$$\sum_{\ell_1 ad - \ell_2 bc = h} \alpha_a \gamma_c \mathcal{W}\left(\frac{a}{q^{1/2}}, \frac{b}{q^{1/2}}, \frac{c}{q^{1/2}}, \frac{d}{q^{1/2}}\right) = MT((\alpha_a), (\gamma_c), \ell_1, \ell_2, h) + O((\ell_1 \ell_2)^A q^{1-\delta}) \tag{3.12}$$

for some $\delta > 0$ absolute and where $MT((\alpha_a), (\gamma_c), \ell_1, \ell_2, h)$ denotes a natural main term and with $(\alpha_a)_{a \sim q^{1/2}}, (\gamma_c)_{c \sim q^{1/2}}$ arbitrary complex numbers of modulus bounded by 1. Since the b variable is smooth, the condition $\ell_1 ad - \ell_2 bc = h$ is essentially equivalent to the congruence condition $\ell_1 ad \equiv h \pmod{c \ell_2}$. One can then analyze this congruence by Poisson summation applied on the remaining smooth variable d which yields sums of Kloosterman fractions of the shape

$$\sum_{\substack{a \sim A, c \sim C \\ (a,c)=1}} \alpha_a \gamma_c e\left(h \frac{\bar{a}}{c}\right), \quad \text{for } h \neq 0$$

and where the values of a, c, h and α_a, γ_c may be different from the previous ones. In [21] such sums are bounded non-trivially for any ranges A, C (the most crucial one being $A = C$).

A remarkable feature of this proof is that the bound is obtained from an application of the amplification method in a very unexpected direction, namely by amplifying the trivial (!) multiplicative characters $\chi_{0,a}$ of modulus a in the family of sums

$$\left\{ \sum_{\substack{c \sim C \\ (a,c)=1}} \gamma_c \chi(c) e\left(h \frac{\bar{a}}{c}\right), \chi \pmod{a}, a \sim A \right\}.$$

Remark 3.3. Note that (3.12) is more general than needed for (3.3) and may be used in other contexts (e.g. Bombieri–Vinogradov type results). On the other hand, in the subconvexity context, this method uses the special shape of Eisenstein series and does not seem to generalize to Rankin–Selberg L -functions.

3.4.2. Subconvexity of Rankin–Selberg L -functions via subconvexity for twisted L -functions. The case of Rankin–Selberg L -functions over $\mathbb{Q}, L(\pi_1 \times \pi_2, s)$ when π_2 is essentially fixed and π_1 has a central character ω_1 of large conductor was treated in [39], [60]. In the case of a varying central characters, subconvexity comes from an estimate for an average of shifted convolution sums of h of the form:

$$\sum_{0 < |h| \ll q} \bar{\omega}(h) \Sigma_{\pm}(\varphi_2, \ell_1, \ell_2, h) \ll_{\varphi_2} (\ell_1 \ell_2)^A q^{3/2-\delta} \tag{3.13}$$

for some $A, \delta > 0$ absolute. Observe however that this is stronger than just the shifted convolution problem on average over h . In particular even under the Ramanujan–Petersson conjecture (H_0), the individual bound (3.10) is “just” not sufficient: this means that one has to account for the averaging over the h variable.

This bound is achieved through the spectral decomposition of the shifted convolution sums (3.11): plugging this formula into the left-hand side of (3.13) one obtains a sum over the orthonormal basis $\{\psi\}$ of sums of the form

$$\sum_{0 < |h| \ll q} \overline{\omega}(h) \rho_{\psi}(-h)$$

if ψ belong to the space V_{τ} of some automorphic representation, the later sums are partial sums associated to the twisted L -function $L(\tau \times \overline{\omega}, s)$. In that case, the subconvexity bound for twisted L -functions (3.2) is exactly sufficient to give (3.13).

Remark 3.4. Hence the subconvexity bound for an L -functions of degree 4 has been reduced to a collection of subconvex bounds for L -functions of automorphic forms of small level twisted by the original central character ω ! This surprising phenomenon is better explained via the approach described in the next section.

4. Subconvexity via periods of automorphic forms

4.1. The various perspectives on an L -function. From the perspective of analytic number theory, the definition of L -function might be “an analytic function sharing the key features of $\zeta(s)$: analytic continuation, functional equation, Euler product.”

However, there are various “incarnations” of L -functions attached to automorphic forms; although equivalent, different features become apparent in different incarnations. For instance, one can define and study L -functions via constant terms of Eisenstein series (the Langlands-Shahidi method), via periods of automorphic forms (the theory of integral representations, which begins with the work of Hecke, or indeed already with Riemann), or via a Dirichlet series (which is often taken as their defining property).

Thus far in this article, we have discussed the subconvexity from the perspective of Dirichlet series. In particular, we have studied periods (e.g. (2.4)) by relating them to L -functions (via (2.2)) and then proving subconvexity for the latter. Relatively recently, the subconvexity question has also been successfully approached via the “period” perspective by reversing this usual process: namely by deducing subconvexity from a geometric study of the periods. The first such result (in the eigenvalue aspect) was given by Bernstein-Reznikov [3], [4], and a little later a result in the level aspect was given by Venkatesh [75]. The two methods seem to be quite distinct. We shall discuss these briefly, and then discuss in more detail the joint work [61] of the authors, which also uses the period perspective.

These approaches are closely related to existing work, but in many cases the period perspective allows certain conceptual simplifications and it brings together harmonic analysis and ideas from dynamics. Such conceptual simplifications are particularly of value in passing from \mathbb{Q} to a general number field; so far, with the exception of the result of Cogdell–Piatetski-Shapiro–Sarnak, all the results in Theorem 6 in the case $F \neq \mathbb{Q}$ are proven via the period approach.

On the other hand, it might be noted that a slight drawback to the period approach to subconvexity is that, especially for automorphic representations with complicated ramification, one must face the difficulty of choosing appropriate test vectors.

4.2. Triple product period and triple product L -function. At present, all known results towards the subconvexity of triple product L -functions $L(\pi_1 \times \pi_2 \times \pi_3, 1/2)$ arise from the “period” perspective.

The period of interest is

$$\int_{\mathrm{PGL}_2(\mathbb{Q}) \backslash \mathrm{PGL}_2(A)} \varphi_1(g) \varphi_2(g) \varphi_3(g) dg$$

where $\varphi_i \in \pi_i$, and each π_i is an automorphic cuspidal representation of GL_2 . It is expected that this period, and the variants when GL_2 is replaced by the multiplicative group of a quaternion algebra, is related to the central value of the triple product L -function $L(\pi_1 \times \pi_2 \times \pi_3, 1/2)$, see [40]. A precise relationship has been computed for the case of Maass forms at full level in [77]; indeed, the following formula is established:

$$\left| \int_{\mathrm{SL}_2(\mathbb{Z}) \backslash \mathbb{H}} \varphi_1 \varphi_2 \varphi_3 d\mu \right|^2 = \frac{\Lambda(\varphi_1 \times \varphi_2 \times \varphi_3, 1/2)}{\Lambda(\wedge^2 \varphi_1, 1) \Lambda(\wedge^2 \varphi_2, 1) \Lambda(\wedge^2 \varphi_3, 1)} \quad (4.1)$$

where Λ denotes the completed L -function and $d\mu$ is a suitable multiple of $\frac{dx dy}{y^2}$.

4.2.1. The eigenvalue aspect: the method of Bernstein–Reznikov. Let Γ be a (discrete) cocompact subgroup of $\mathrm{SL}_2(\mathbb{R})$, let \mathbb{H} be the upper half-plane, let φ_1, φ_2 be fixed eigenfunctions of the Laplacian on $\Gamma \backslash \mathbb{H}$ and φ_λ an eigenfunction with eigenvalue $\lambda := 1/4 + r^2$. In the paper [4], Bernstein and Reznikov establish the following bound:

$$r^2 e^{\pi r/2} \left| \int_{\Gamma \backslash \mathbb{H}} \varphi_1(z) \varphi_2(z) \varphi_\lambda(z) d\mu_z \right|^2 \ll_\varepsilon r^{5/3+\varepsilon}. \quad (4.2)$$

In fact the bound (4.2) remains valid if $\Gamma \backslash \mathbb{H}$ has only finite volume and $\varphi_1, \varphi_2, \varphi_\lambda$ are cusp forms. In particular, when $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ and $\varphi_1, \varphi_2, \varphi_\lambda$ are Hecke–Maass forms associated to automorphic representations $\pi_1, \pi_2, \pi_\lambda$ respectively, the bound (4.2) translates, via (4.1), to the subconvex bound

$$L(\pi_1 \times \pi_2 \times \pi_\lambda, 1/2) \ll_{\varepsilon, \pi_1, \pi_2} r^{5/3+\varepsilon} \quad (4.3)$$

while the convexity bound for the left-hand side is $r^{2+\varepsilon}$.

Their method is based on the properties of the (local) real group $G = \mathrm{PGL}_2(\mathbb{R})$ and, in particular, on the fact that the space of G -invariant functionals on $\pi_1 \otimes \pi_2 \otimes \pi_\lambda$ is *at most* one dimensional. Hence the proof is purely local and by contrast to the method of [75], does not use either Hecke operators or the spectral gap.

4.2.2. The level aspect: the method of Venkatesh. Let F be a number field. Let π_2, π_3 be fixed automorphic cuspidal representations on $\mathrm{PGL}_2(A_F)$ – say with co-prime conductor – and let π_1 be a third automorphic cuspidal representation with conductor \mathfrak{q} , a prime ideal of F . In [75] it is established that

$$L(\pi_1 \times \pi_2 \times \pi_3, 1/2) \ll_{\pi_1, \infty, \pi_2, \pi_3} N(\mathfrak{q})^{1-\frac{1}{13}} \tag{4.4}$$

contingent on a suitable version of (4.1) when the level of one factor varies.⁴ The convexity bound for the left-hand side is $N(\mathfrak{q})^{1+\varepsilon}$.

Remark 4.1. In [75], a form of (4.4) is proved when π_2 and/or π_3 are Eisenstein series: in that case, (4.1) corresponds to simple computations in the Rankin–Selberg method and so is unconditional. In particular, this yields the bounds (3.3) and (3.4) for π_1 with trivial central character.

For reasons of space, we do not explain the details of the proof; in any case, this can also be approached by the method outlined in Section 4.3. It uses, in particular, quantitative results and ideas from ergodic theory, and the bound H_θ with $\theta < 1/4$, in the notation of Section 3.3.

4.3. Central character. In this section, we return to Section 3.4 and explain, via periods, the bound (3.4). In particular, this sheds light on the “reason” for the reduction to a lower rank subconvexity problem that was encountered in that section. The content of this section is carried out in detail in [61].

Let π_1, π_2 be automorphic cuspidal representations of $\mathrm{GL}_2(A_F)$. Let ω be the central character of π_1 . For simplicity, we restrict ourselves to the case where π_2 , the “fixed” form, has level 1 and trivial central character; and where “all the ramification of π_1 comes from the central character,” i.e. π_1 and ω have the same conductor \mathfrak{q} .

Let us first give a very approximate “philosophical” overview of the proof. There is an identity between mean values of L -functions of the following type:

$$\sum_{\pi_1} L(\pi_1 \times \pi_2, 1/2) \longleftrightarrow \sum_{\tau \text{ level } 1} L(\tau, 1/2)L(\tau \times \omega, 1/2) \tag{4.5}$$

where the left-hand summation is over π_1 of central character ω and conductor \mathfrak{q} , whereas the right-hand summation is over automorphic representations τ of trivial

⁴This has not appeared in the literature to our knowledge, except in the case where one of the π_j are Eisenstein; however, it should amount to a routine though very involved computation of p -adic integrals.

central character and level 1. It includes the trivial (one-dimensional) automorphic representation, which is in fact the dominant term and actually needs to be handled by regularization.⁵

By means of a suitable amplifier, one can restrict the left-hand summation to pick out a given π_1 . When one does this, the necessary bounds on the right-hand side follow from two different inputs:

1. Subconvexity for $L(\tau \times \omega, 1/2)$ (in the aspect where ω varies), to handle the nontrivial τ .
2. A bound showing decay of matrix coefficients of p -adic groups, to handle the contribution of τ the trivial representation

4.3.1. The source of (4.5) via periods. Writing $Y_A = \mathrm{PGL}_2(F) \backslash \mathrm{PGL}_2(A_F)$, we note that the Rankin–Selberg L -function may be expressed as a period integral:

$$L(s, \pi_1 \times \pi_2) \sim \int_{Y_A} \varphi_1(g)\varphi_2(g)E_s(g) dg$$

where $\varphi_i \in \pi_i$ are the respective newforms, and E_s is the Eisenstein series corresponding to the new vector of the automorphic representation $|\cdot|^s \boxplus \omega^{-1}|\cdot|^{-s}$. Here \sim means that there is a suitable constant of proportionality, depending on the archimedean types of the representations.

Let $\mathcal{B}_{\omega, \mathfrak{q}}$ be an orthogonal basis for the space of forms on $\mathrm{GL}_2(F) \backslash \mathrm{GL}_2(A_F)$ of level \mathfrak{q} and central character ω ; let $\mathcal{B}_{1,1}$ be an orthogonal basis for the space of forms on Y_A of full level and trivial central character. By spectral expansion, we have the following identity:

$$\begin{aligned} \sum_{\varphi_1 \in \mathcal{B}_{\omega, \mathfrak{q}}} \left| \int_{Y_A} \varphi_1 \varphi_2 E_s \right|^2 &= \int_{Y_A} |\varphi_2 \cdot E_s|^2 \\ &= \int_{Y_A} |\varphi_2|^2 |E_s|^2 = \sum_{\psi \in \mathcal{B}_{1,1}} \langle |E_s|^2, \psi \rangle \overline{\langle |\varphi_2|^2, \psi \rangle}. \end{aligned} \tag{4.6}$$

Here the ψ -summation is, *a priori*, over an orthonormal basis for $L^2(Y_A)$; however, the summand $\langle |\varphi_2|^2, \psi \rangle$ vanishes unless ψ is of level 1 and trivial central character. Note that the ψ -summation should, strictly, include a continuous contribution for the Eisenstein series, which also needs to be suitably regularized. This is not a trivial matter and occupies a good deal of [61]; we shall suppress it for now.

In any case, if ψ belongs to the space of an automorphic representation τ , then the Rankin–Selberg method shows that $\langle |E_s|^2, \psi \rangle$ is a multiple of $L(\tau, 2s - 1/2)L(\tau \times \omega, 1/2)$. Thus (4.6) basically yields (4.5)!

⁵Note that “morally”, when τ is trivial, the L -function $L(\tau, s)L(\tau \times \omega, s) = \zeta(s+1/2)\zeta(s-1/2)L(\omega, s-1/2)L(\omega, s+1/2)$. Thus we obtain a pole at $s = 1/2$.

4.3.2. Amplification and the decay of matrix coefficients. We restrict to $s = 1/2$ for concreteness, although the method works for any s . The identity (4.6) does not suffice to obtain a nontrivial bound on $\int_{Y_A} \varphi_1 \varphi_2 E_{1/2}$, for the left-hand summation is too large. To localize it, one introduces an amplifier. We phrase it adelicly, but it should be made clear this is still the amplifier of Friedlander–Iwaniec.

For any function f on $\mathrm{GL}_2(F) \backslash \mathrm{GL}_2(\mathbf{A}_F)$ and any $g_0 \in \mathrm{GL}_2(\mathbf{A}_F)$, we write $f^{g_0}(g) = f(gg_0^{-1})$. Then one has the following tiny variant of (4.6), for $g_1, g_2 \in \mathrm{GL}_2(\mathbf{A}_F)$:

$$\begin{aligned} \sum_{\varphi_1 \in \mathcal{B}_{\omega, \mathfrak{q}}} \left(\int_{Y_A} \varphi_1^{g_1^{-1}} \varphi_2 E_{1/2} \right) \overline{\left(\int_{Y_A} \varphi_1^{g_2^{-1}} \varphi_2 E_{1/2} \right)} \\ = \sum_{\psi \in L^2(Y_A)} \langle E_{1/2}^{g_1} E_{1/2}^{g_2}, \psi \rangle \overline{\langle \varphi_2^{g_1} \varphi_2^{g_2}, \psi \rangle}. \end{aligned} \tag{4.7}$$

This is again an identity of this shape (4.5), but with slightly more freedom due to the insertion of g_1, g_2 . The left-hand (resp. right-hand) side is still proportional, by the Rankin–Selberg method, to $L(\pi_1 \times \pi_2, 1/2)$ (resp. $L(\tau, 1/2)L(\tau \times \omega, 1/2)$, if $\psi \in \tau$) but the constants of proportionality depend – in a precisely controllable way – on g_1, g_2 . In effect, this allows one to introduce a “test function” $h(\pi_1)$ into the identity (4.5), thereby shortening the effective range of summation. It should be noted that in (4.7), by contrast with (4.5), the right hand ψ -summation is no longer over ψ of level 1; however, it involves only those ψ which are invariant by $\mathrm{PGL}_2(\hat{\mathbb{Z}}) \cap g_1^{-1} \mathrm{PGL}_2(\hat{\mathbb{Z}}) g_1 \cap g_2^{-1} \mathrm{PGL}_2(\hat{\mathbb{Z}}) g_2$, and in particular their level is bounded in a way that depends predictably on g_1, g_2 .

A subconvex bound for $L(\pi_1 \times \pi_2, 1/2)$ follows from any method to get nontrivial bounds on the right-hand side of (4.7) for general g_1, g_2 .

1. To deal with the case when ψ is perpendicular to the constants, we note that the terms $\langle E_{1/2}^{g_1} E_{1/2}^{g_2}, \psi \rangle$ are, by Rankin–Selberg, certain multiples of $L(\tau, 1/2)L(\tau \times \omega, 1/2)$ whenever $\psi \in \tau$, the space of an automorphic representation. We then apply subconvex bounds for $L(\tau \times \omega, 1/2)$, in the aspect when ω varies.
2. To deal with the case $\psi = \mathrm{Const}$, we note that the term $\langle \varphi_2^{g_1} \overline{\varphi_2^{g_2}}, \psi \rangle$ is, in that case, simply a multiple of the matrix coefficient $\langle \varphi_2^{g_1 g_2^{-1}}, \varphi_2 \rangle$. Thus it is bounded by bounds on the decay of matrix coefficients.

Obviously this description is dishonest, for the term $\langle |E_{1/2}|^2, \psi \rangle$ is not even convergent for $\psi = \mathrm{Const}$! However, this is a technical and not a conceptual difficulty: the only two analytic ingredients required are the two above.

4.3.3. Mysterious identities between families of L -functions. In a sense, the period identity (4.6) (or, approximately equivalent, the identity (4.5)) is the key point of

the above discussion; it explains immediately why one has the “reduction of degree” discussed in Section 3.4. This is another example of the phenomenon discussed in Section 4.1: the identity (4.5) is obvious from the “period” perspective, but not at all clear from the viewpoint of L -functions considered as Dirichlet series.

Another example of such a phenomenon – identities between *a priori* different families of L -functions – is Motohashi’s beautiful formula [64] for the 4th moment of ζ . Roughly speaking, it relates integrals of $|\zeta(1/2 + it)|^4$ to sums of $L(\varphi, 1/2)^3$, where φ varies over Maass forms. If φ is a suitably normalized Maass form on $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathbb{H}$, its completed L -function is given by the Hecke period $\Lambda(\varphi, 1/2 + it) = \int_0^\infty \varphi(iy) y^{it} d^\times y$. Applying Plancherel’s formula shows that $\frac{1}{2\pi} \int_{-\infty}^\infty |\Lambda(\varphi, 1/2 + it)|^2 dt = \int_{y=0}^\infty |\varphi(iy)|^2 d^\times y$. Again, one can spectrally expand $|\varphi|^2$, yielding:

$$\begin{aligned} \frac{1}{2\pi} \int_{-\infty}^\infty |\Lambda(\varphi, 1/2 + it)|^2 dt &= \sum_{\psi} \frac{\langle |\varphi|^2, \psi \rangle}{\langle \psi, \psi \rangle} \int_0^\infty \psi(iy) d^\times y \\ &= \sum_{\psi} \frac{\langle |\varphi|^2, \psi \rangle}{\langle \psi, \psi \rangle} \Lambda(\psi, 1/2) \end{aligned} \tag{4.8}$$

where, again, the ψ -sum is over an orthogonal basis, suitably normalized, for $L^2(Y_0(1))$, and, again, we suppress the continuous spectrum; hence (4.8) expresses a relation between mean values of $L(\varphi, 1/2 + it)$, where t varies, and the family of L -functions $L(\psi, 1/2)$, where ψ varies. Specializing (4.8) to the case of φ the Eisenstein series at the center of symmetry yields a formula “of Motohashi type.” We emphasize that this argument has not been carried out rigorously to our knowledge and it would likely involve considerable technical difficulty (for the integrals diverge in the Eisenstein case). Nevertheless, this approach may have value insofar as it offers some insight into the origin of such formulae. A. Reznikov has given a very general and elegant formalism [67] that encapsulates such identities as (4.5) and (4.8); one hopes that further analytic applications will stem from his formalism.

5. Applications

5.1. Subconvexity and functoriality. Via the functoriality principle of Langlands, it is now understood that the same L -function may be attached to automorphic forms on different groups. This gives rise to the possibility of studying the same L -function in different ways.

A recent instance where this kind of idea played a decisive role was the attempt to solve the subconvexity problem for the L -functions of the class group characters of a quadratic field K of large discriminant. In [24], the problem was solved but only under the assumption that K has sufficiently many small *split* primes (this would follow from GRH, but so far, has been established unconditionally only for special discriminants). This assumption, which was also encountered by the second author

in [75] in the context of periods and is closely related to Linnik’s condition, is a fundamental and major unsolved issue that arises in many contexts, e.g. in work on the André–Oort conjecture [79]. A key observation of [23], is that, by functoriality, (in that case due to Hecke and Maass) a class group character L -function is the L -function of a Maass form of weight 0 or 1, with Laplace eigenvalue $1/4$. For these, as we have just seen, the subconvexity problem can be solved independently of any assumption.

In view of this example, we find it useful to spell out explicitly some direct consequences of the subconvex bounds of Theorem 6 and of functoriality.

Corollary 5.1. *Let F be a fixed number field and $\rho : \text{Gal}(\bar{F}/F) \rightarrow \text{GL}_2(\mathbb{C})$ be a modular Galois representation (for instance, if the image of ρ in $\text{PGL}_2(\mathbb{C})$ is soluble). Let \mathfrak{q}_ρ be the Artin conductor of ρ and let $L(\rho, s)$ be its Artin L -function, then for $\Re s = 1/2$*

$$L(\rho, s) \ll_{F,s} N_{F/\mathbb{Q}}(\mathfrak{q}_\rho)^{1/4-\delta}$$

for $\delta > 0$ some absolute constant.

Corollary 5.2. *Let F be a fixed number field and K be an extension of F of absolute discriminant $\text{disc}(K/\mathbb{Q}) =: \Delta_K$ and let $\zeta_K(s)$ be the Dedekind zeta function of K ; then, if K/F is abelian or cubic, one has for $\Re s = 1/2$*

$$\zeta_K(s) \ll_{F,s} |\Delta_K|^{1/4-\delta}$$

for $\delta > 0$ some absolute constant.

Corollary 5.3. *Let F be a fixed number field, π be a fixed $\text{GL}_2(\mathcal{A}_F)$ -automorphic cuspidal representation and let K be an extension of F of absolute discriminant $\text{disc}(K/\mathbb{Q}) =: \Delta_K$. If K/F is abelian or cubic, we denote by π_K the base change lift of π from F to K (which exist by the works of Saito–Shintani–Langlands and Jacquet–Piatetski-Shapiro–Shalika). For $\Re s = 1/2$, one has*

$$L(\pi_K, s) \ll_{F,\pi,s} |\Delta_K|^{1/2-\delta}$$

for $\delta > 0$ some absolute constant.

5.2. Equidistribution on quaternionic varieties. We define a quaternionic variety as the locally homogeneous space given as an adelic quotient of the following form: for F a totally real number field, B a quaternion algebra over F , let G be the \mathbb{Q} -algebraic group $\text{res}_{F/\mathbb{Q}} B^\times / Z(B^\times)$; one has

$$G(\mathbb{R}) \simeq \text{PGL}_2(\mathbb{R})^{f'} \times \text{SO}(3, \mathbb{R})^{f-f'}$$

where $f = \text{deg } F$ and f' is the number of real place of F for which B splits. Let K_∞ be a compact subgroup of $G(\mathbb{R})$ of the form

$$\text{SO}(2, \mathbb{R})^{f'} \times \prod_{v=1}^{f-f'} K_v$$

with $K_v =$ either $\mathrm{SO}_2(\mathbb{R})$ or $\mathrm{SO}_3(\mathbb{R})$ and let X denote the quotient $\mathbf{G}(\mathbb{R})/K_\infty$; finally let K_f be an open compact subgroup of $\mathbf{G}(A_f)$ and $K := K_\infty \cdot K_f$.

The quaternionic variety $V_K(\mathbf{G}, X)$ is defined as the quotient

$$V_K(\mathbf{G}, X) := \mathbf{G}(\mathbb{Q}) \backslash X \times \mathbf{G}(A_f)/K_f = \mathbf{G}(\mathbb{Q}) \backslash \mathbf{G}(A_\mathbb{Q})/K.$$

It has the structure of a Riemannian manifold whose connected components are quotients, by a discrete subgroup of $\mathbf{G}(\mathbb{R})$, of the product of $(\mathbb{H}^\pm)^{f'} \times (S^2)^{f''}$ for $f'' \leq f - f'$. The case of the sphere and of the modular surface correspond to the case $F = \mathbb{Q}$, \mathbf{B} the algebra of 2×2 matrices $M_2(\mathbb{R})$ or the Hamilton quaternions $\mathbf{B}^{(2, \infty)}$.

Let K/F be a quadratic extension with an embedding into \mathbf{B} , and let \mathbf{T} denote the \mathbb{Q} -torus “ $\mathrm{res}_{F/\mathbb{Q}} K^\times / F^\times$ ”. As was pointed out in Section 2.2.1, there exists, in great generality, a precise relationship between

1. central values of some Rankin–Selberg L -function $L(\pi_\chi \times \pi_2, s)$ (for which the sign of the functional equation $w(\pi_\chi \times \pi_2)$ is $+1$); and
2. (the square of) twisted Weyl sums

$$\int_{\mathbf{T}(\mathbb{Q}) \backslash \mathbf{T}(A_\mathbb{Q})} \chi(t) \varphi_2(z.t) dt.$$

These Weyl sums describe the distribution properties of toric orbits, $\mathbf{T}(\mathbb{Q}) \backslash z.\mathbf{T}(A_\mathbb{Q})$ of cycles associated to (orders of) K inside $V_K(\mathbf{G}, X)$.

The general scheme is that, in cases where these formula have been written out explicitly, the subconvex bound (3.2) (along possibly with hypothesis H_θ for some $\theta < 1/2$) yields at once the equidistribution of the *full* orbit and the subconvex bounds (3.4) yield the equidistribution of *big enough* suborbits of the toric orbit. We present below some sample results on these lines:

5.2.1. Hilbert’s eleventh problem. When B is totally definite, $K_\infty = \mathrm{SO}_2(\mathbb{R})^f$, $X = (S^2)^f$ is a product of spheres. In this case, the equidistribution of toric orbits (relative to a totally imaginary quadratic field) above can be interpreted in terms of the integral representations of a totally positive integer $d \in \mathcal{O}_F$ by a totally positive definite quadratic form q (more precisely $-q$ “is” the norm form $N_{B/F}(\mathbf{x})$ on the space of quaternions of trace 0). The following theorem of Cogdell–Piatetski-Shapiro–Sarnak combines the formula of [1] with (3.2) for π_2 holomorphic.

Theorem 7. *Let F be a totally real number field and q be an integral positive definite quadratic form over F ; there is an absolute (ineffective) constant $N_{F,q} > 0$ such that if d is a squarefree totally positive integer with $N_{F/\mathbb{Q}}(d) > N_{F,q}$ then d is integrally represented by q iff d is everywhere locally integrally represented. Moreover, in the later case, the number, $r_q(d)$, of all such integral representation satisfies*

$$r_q(d) \gg_{q,F} N_{F/\mathbb{Q}}(d)^{1/2+o(1)} \quad \text{as } N_{F/\mathbb{Q}}(d) \rightarrow +\infty.$$

Remark 5.1. The question of the integral representability of d by some form in the genus of q was completely settled a long time ago by Siegel, in a quantitative way, through the Siegel mass formula. The present theorem (in a slightly more precise form) can then be interpreted by saying that the various representations d are *equidistributed* amongst the various genus classes of q ; moreover it can be strengthened to an “equidistribution on ellipsoids” statement, cf. [26] for $F = \mathbb{Q}$.

5.2.2. CM points on quaternionic Shimura varieties. When B is indefinite at some real place and $K_\infty = \mathrm{SO}_2(\mathbb{R})^{f'} \times \mathrm{SO}_3(\mathbb{R})^{f-f'}$ the quaternionic variety $V_K(\mathbf{G}, X)$ is a Shimura variety, $\mathrm{Sh}_{K_f}(\mathbf{G}, X)$ (a Hilbert modular variety of complex dimension f'). It has the structure of the complex points of an algebraic variety defined over some reflex field E/F .

In this setting, the generalization of the set of Heegner point is the so called set of “CM” points, $\mathcal{H}_\mathfrak{d}$, which is associated to a quadratic order $\mathcal{O}_\mathfrak{d}$ (say of discriminant \mathfrak{d}) of a (not necessarily fixed) totally imaginary K/F . In that case and under some natural local condition, the equidistribution of

$$\mathcal{H}_\mathfrak{d} = T(\mathbb{Q}) \backslash_{z_\mathfrak{d}} \cdot T(A_\mathbb{Q}) / T(\widehat{\mathcal{O}}_\mathfrak{d})$$

on $\mathrm{Sh}_{K_f}(\mathbf{G}, X)$ as $|N_{F/\mathbb{Q}}(\mathfrak{d})| \rightarrow +\infty$ was established independently by Clozel–Ullmo, Cohen and Zhang [12], [14], [82] by using the subconvex bound (3.2) of the second author. For instance, one has

Theorem 8. *Suppose $K_f = K_{f, \max}$ is a maximal compact subgroup of $\mathbf{G}(A_F)$, then for $|N_{F/\mathbb{Q}}(\mathfrak{d})| \rightarrow +\infty$ and \mathfrak{d} coprime with $\mathrm{disc}(F)$, the set $\mathcal{H}_\mathfrak{d}$ becomes equidistributed on $\mathrm{Sh}_{K_f}(\mathbf{G}, X)$ w.r.t. the hyperbolic measure.*

Similarly, as in Theorem 5, the bound (3.4) allows one to show the equidistribution of strict suborbits of $z_\mathfrak{d}$:

Theorem 9. *With the notations as above, there is an absolute constant $0 < \eta < 1$ such that, for any subtoric orbit $\mathcal{H}'_\mathfrak{d} \subset \mathcal{H}_\mathfrak{d}$ of size satisfying $|\mathcal{H}'_\mathfrak{d}| \geq |\mathcal{H}_\mathfrak{d}|^\eta$, then $\mathcal{H}'_\mathfrak{d}$ is equidistributed on $\mathrm{Sh}_{K_f}(\mathbf{G}, X)$ as $|N(\mathfrak{d})| \rightarrow +\infty$.*

As was pointed out by Zhang [82], the possibility of considering strict suborbits of the full toric orbit has a nice arithmetic interpretation; the Galois orbits on CM points correspond to “subtoric orbits” of the type considered in Theorem 9.

6. Linnik’s ergodic method: a modern perspective

As discussed, Linnik achieved partial results towards Theorems 1–3 by using some ingenious ideas which he collectively referred to as “the ergodic method.” As Linnik pointed out (see, e.g., [57, Chapter XI, comments on Chapters IV–VI]) despite this name, this method remained rather *ad hoc* and did not fit into ergodic theory as it is

normally understood: that is to say, dynamics of a measure-preserving transformation. The joint work of the authors with M. Einsiedler and E. Lindenstrauss [29], remedies this, both putting Linnik's original work into a more standard ergodic context, and giving the first higher rank generalizations.

6.1. The source of dynamics. Although the relevance of dynamics to integral points on the sphere is not immediately apparent, it is not difficult to see from an adelic perspective. We have already mentioned in Section 2.2.1 that all three theorems (Theorems 1–3) may be considered as questions about the distribution of an orbit of an adelic torus $z_d \cdot T_d(\mathbf{A})$ inside $\mathbf{G}(\mathbb{Q}) \backslash \mathbf{G}(\mathbf{A})$.

One can, therefore, hope to use results about the dynamics of a local torus $T_d(\mathbb{Q}_v)$ acting on $\mathbf{G}(\mathbb{Q}) \backslash \mathbf{G}(\mathbf{A})$ for some fixed place v . However, this is possible only if $T_d(\mathbb{Q}_v)$ is noncompact; for otherwise there is no dynamics of interest. This leads to Linnik's condition (cf. Theorem 4), because $T_d(\mathbb{Q}_p)$ is noncompact precisely when $\mathbb{Q}(\sqrt{d})$ is split at p .

6.2. Linnik's method in the light of modern ergodic theory. Much of this joint work is based on the recent work of Einsiedler and Lindenstrauss on classification of invariant measures for toric actions, which is discussed in their contribution to these proceedings [28].

A central concept here is that of *entropy*; we briefly reprise the definition. We recall that if \mathcal{P} is a partition of the probability space (X, ν) , the entropy of \mathcal{P} is defined as $h_\nu(\mathcal{P}) := \sum_{S \in \mathcal{P}} -\nu(S) \log \nu(S)$. If T is a measure-preserving transformation of (X, ν) , then the measure entropy of T is defined as

$$h(T) = \sup_{\mathcal{P}} \lim_{n \rightarrow \infty} \frac{h_\nu(\mathcal{P} \vee T^{-1}\mathcal{P} \vee \dots \vee T^{-(n-1)}\mathcal{P})}{n} \quad (6.1)$$

where the supremum is taken over all finite partitions of X .

Here are two results that illustrate the importance of this concept (we denote by Haar the G -invariant probability measure on a quotient space $\Gamma \backslash G$).

The first one is a specialization of the fact that on the unit tangent bundle of a surface of constant negative curvature, the Liouville measure is the unique measure of maximal entropy w.r.t. the action of the geodesic flow:

Fact 1. *Let μ on $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathrm{SL}_2(\mathbb{R})$ be invariant by the diagonal subgroup, and let a be a nontrivial diagonal matrix. Then $h_\mu(a) \leq h_{\mathrm{Haar}}(a)$, with equality if and only if $\mu = \mathrm{Haar}$.*

The second fact lies much deeper and is a result of Einsiedler, Katok and Lindenstrauss [27] which illustrate the phenomenon of *measure rigidity* for the action of tori of rank ≥ 2 :

Fact 2. *Let μ be a probability measure on $\mathrm{SL}_3(\mathbb{Z}) \backslash \mathrm{SL}_3(\mathbb{R})$ invariant by the diagonal subgroup A and let $a \in A$ be nontrivial. If $h_\mu(a) > 0$ and μ is ergodic (w.r.t. A), then $\mu = \mathrm{Haar}$.*

The scheme of [29, I, II and III] is to treat Linnik problems by combining results of the above type – towards the classification of measures with positive entropy – with diophantine ideas that establish positive entropy. In the subsequent sections we discuss some applications of this general scheme; we have aimed for concreteness, but these methods are much more generally applicable.

6.3. Entropy and the “Linnik principle”. In [29, II] we give a purely dynamical proof of Theorem 3. This proof is still based heavily on Linnik’s ideas but it introduces considerable conceptual simplification using the notion of entropy discussed in the previous section, and uses in particular *Fact 1*.

We insist that our proof requires *no* splitting condition at some fixed prime p : the reason is that in the context of Theorem 3, the place $v = \infty$ splits in the real quadratic field $\mathbb{Q}(\sqrt{d})$ and so Linnik’s condition is *satisfied*! Curiously this was apparently never remarked by Linnik and Skubenko who only used the action of a p -adic split torus.

Let $d > 0$ be a fundamental discriminant. The unit tangent bundle of $Y_0(1)$ is identified with $\mathrm{PGL}_2(\mathbb{Z}) \backslash \mathrm{PGL}_2(\mathbb{R})$, and so the subset Γ_d described in Theorem 3 may be regarded as a subset $\Gamma_d \subset \mathrm{PGL}_2(\mathbb{Z}) \backslash \mathrm{PGL}_2(\mathbb{R})$. Considered in this way, Γ_d is invariant by the subgroup

$$A = \left\{ a(t) = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}, t \in \mathbb{R} \right\}$$

of diagonal matrices with positive entries. It supports a natural A -invariant probability μ_d (the one which assigns the same mass to each connected component) and Theorem 3 asserts precisely that μ_d converge weakly to the $\mathrm{PGL}_2(\mathbb{R})$ -invariant probability measure on $\mathrm{PGL}_2(\mathbb{Z}) \backslash \mathrm{PGL}_2(\mathbb{R})$.

The dynamical proof uses Fact 1 together with a Diophantine computation to show that any weak limit of the μ_d has maximal entropy w.r.t. the action of $a(1)$. The Diophantine computation is a version of “Linnik’s basic Lemma,” [57, Theorem III.2.1] which in turn may be deduced from *Siegel’s mass formula*.

6.4. A rank 3 version of Duke’s theorem. A natural “rank 2” version of Theorem 3 is to consider the distribution properties of appropriate collections of compact *flats* inside the Riemannian manifold $\mathrm{PGL}_3(\mathbb{Z}) \backslash \mathrm{PGL}_3(\mathbb{R}) / \mathrm{PO}_3(\mathbb{R})$.

More generally, let D be a \mathbb{R} -split central simple algebra of rank 3 over \mathbb{Q} , i.e. $\dim_{\mathbb{Q}} D = 9$, so that $D \otimes_{\mathbb{Q}} \mathbb{R} = M_3(\mathbb{R})$. Let \mathcal{O}_D be a fixed maximal order in D . Let G be the algebraic group $\mathrm{PG}(D) = D^{\times} / Z(D)^{\times}$; we fix a maximal split torus $A = (\mathbb{R}^{\times})^2$ inside $G(\mathbb{R})$. Let U be the standard maximal compact subgroup $\prod_p \mathrm{PG}(\mathcal{O}_{D,p})$ of $G(\mathbb{A}_f)$. We will assume, for simplicity, that the class number of \mathcal{O}_D is 1, i.e. that $G(\mathbb{A}_f) = G(\mathbb{Q}) \cdot U$.

Let $K \subset D$ be a totally real cubic field, together with an isomorphism $\theta: K \otimes \mathbb{R} \rightarrow \mathbb{R}^3$. We assume for simplicity that $K \cap \mathcal{O}_D$ is the maximal order \mathcal{O}_K of K . This yields, in particular, an embedding of the torus $T_K = \mathrm{res}_{K/\mathbb{Q}} K^{\times} / \mathbb{Q}^{\times}$

into the algebraic group $\mathrm{PG}(\mathbb{D})$. The choice of θ determines a unique $g_\theta \in \mathbf{G}(\mathbb{R})$ so that $g_\theta A g_\theta^{-1} = \mathbf{T}_K(\mathbb{R})$. Setting $U_T = \mathbf{T}_K(\mathbf{A}_f) \cap U$, we consider

$$\Gamma_K := (\mathbf{T}_K(\mathbb{Q}) \backslash \mathbf{T}_K(\mathbf{A}) / U_T) g_\theta \subset \mathbf{G}(\mathbb{Q}) \backslash \mathbf{G}(\mathbf{A}) / U \cong \mathcal{O}_D^\times \backslash \mathrm{PGL}_3(\mathbb{R}).$$

This is a collection of compact A -orbits on $\mathcal{O}_D^\times \backslash \mathrm{PGL}_3(\mathbb{R})$, which are indexed by $\mathbf{T}_K(\mathbb{Q}) \backslash \mathbf{T}_K(\mathbf{A}_f) / U_T$ and the latter quotient is precisely the class group $\mathrm{Cl}(\mathcal{O}_K)$. Consequently, to any subset $S \subset \mathrm{Cl}(\mathcal{O}_K)$ of the class group, we may associate a collection $\Gamma_{K,S}$ of $|S|$ closed A -orbits on $\mathcal{O}_D^\times \backslash \mathrm{PGL}_3(\mathbb{R})$. This set supports a natural A -invariant probability measure (which assigns the same mass to each of the constituent orbits); call this measure $\mu_{K,S}$. For $S = \mathrm{Cl}(\mathcal{O}_K)$ we write simply Γ_K and μ_K .

In [29, I and III] we investigate weak limits of such measures as $\mathrm{disc}(K) \rightarrow +\infty$. In the split case, we obtain equidistribution for the full packet of compact orbits which represents the 3-dimensional analog of Theorem 3:

Theorem 10. *Suppose D is split (i.e. $D = M_3(\mathbb{Q})$). As $\mathrm{disc}(K) \rightarrow \infty$, Γ_K becomes equidistributed on $\mathrm{PGL}_3(\mathbb{Z}) \backslash \mathrm{PGL}_3(\mathbb{R})$ with respect to the Haar measure.*

In the non-split case, we obtain a weak form of equidistribution but valid for rather small packets of compact orbits:

Theorem 11. *Suppose D is not \mathbb{Q} -split. Fix $\delta < 1/2$. There is a constant $c = c(\delta) > 0$ such that, if each set $S \subset \mathrm{Cl}(\mathcal{O}_K)$ satisfies $\frac{|S|}{|\mathrm{Cl}(\mathcal{O}_K)|} \geq \mathrm{disc}(K)^{-\delta}$, then any weak limit of $\mu_{K,S}$ contains a Haar component of size $\geq c(\delta)$.*

The proof of these results follows the general strategy outlined at the end of Section 6.2 and uses Fact 2 mentioned above as a key ingredient.

In the context of Theorem 10, the first point to verify (since $\mathrm{PGL}_3(\mathbb{Z}) \backslash \mathrm{PGL}_3(\mathbb{R})$ is not compact) is that any weak limit of the μ_K , μ say, is a probability measure i.e. that there is no “escape of mass” to ∞ . To circumvent this difficulty, we use a version of the harmonic analytic approach described in Section 2: from Mahler’s compactness criterion, we build test functions which dominate the characteristic function of small neighborhoods of the cusp and we control escape of mass by bounding the corresponding Weyl sums. The test functions are in fact Eisenstein series (associated with the minimal parabolic of PGL_3) and the resulting Weyl sums can be expressed in terms of the Dedekind zeta function of K , ζ_K to which we apply Corollary 5.2 in the case $F = \mathbb{Q}$. More generally, a variant of this construction together with Corollary 5.2 enable us to bound from above the μ -mass of small neighborhoods of *any* point in $\mathrm{PGL}_3(\mathbb{Z}) \backslash \mathrm{PGL}_3(\mathbb{R})$. This is sufficient to imply that *every* ergodic component of μ has positive entropy for some non-trivial $a \in A$. From this we conclude that $\mu = \mu_{\mathrm{Haar}}$ using Fact 2. It should be remarked that the very existence of such test functions is a special feature of the split case.

In the non-split case, there is no issue about possible escape of mass since $\mathcal{O}_D^\times \backslash \mathrm{PGL}_3(\mathbb{R})$ is compact; on the other hand, simple test functions for which the Weyl sums could be evaluated and from which positive entropy could be deduced do

not seem to be available in the cocompact case; instead, we rely on a weaker version of Linnik’s basic lemma – *Linnik’s principle* – from which we deduce, at least, that a *positive proportion* of the ergodic components of μ have positive entropy; again we conclude by applying Fact 2 to these components. Although it does not achieve equidistribution, Theorem 11 nevertheless illustrates a major advantage of the ergodic approaches of [29] over harmonic-analysis methods: ergodic methods allow for nontrivial results even for *very small* torus orbits (“supersparse equidistribution”): indeed, any exponent $\delta < 1/2$ is admissible and since the size of the class group of \mathcal{O}_K is at most $\text{disc}(K)^{1/2+\varepsilon}$, this is as strong as could be hoped for. Distribution problems for *small* torus orbits arise naturally in several arithmetic questions: for instance we expect that measure rigidity results for actions of p -adic tori should allow for partial progress towards Zhang’s measure-theoretic refinement of the André/Oort conjecture [82].

6.5. An application to Minkowski’s Theorem. We first recall

Theorem (Minkowski). *Let K be a number field of degree d and maximal order \mathcal{O}_K ; any ideal class for \mathcal{O}_K possesses an integral representative $J \subset \mathcal{O}_K$ of norm $N(J) = O(\sqrt{\text{disc}(K)})$ where the implicit constant depends only on d .*

We conjecture that this is not sharp for totally real number fields of degree $d \geq 3$:

Conjecture 2. Suppose $d \geq 3$ is fixed. Then any ideal class in a totally real number fields of degree d has an integral representative of norm $o(\sqrt{\text{disc}(K)})$.

Let $m(K)$ denote the maximum, over ideal classes of \mathcal{O}_K , of the minimal norm of a representative. Conjecture 2 asserts that

$$\lim_{\text{disc}(K) \rightarrow \infty} \frac{m(K)}{\text{disc}(K)^{1/2}} = 0,$$

for K varying through totally real fields of fixed degree $d \geq 3$. It may be shown that for any $d \geq 2$ there exists a $c' > 0$ such that there is an infinite set of totally real fields of degree d for which $m(K) \geq c' \cdot \text{disc}(K)^{1/2} (\log \text{disc}(K))^{1-2d}$. Thus Minkowski’s Theorem is rather close to sharp and in fact Conjecture 2 is unlikely to be true for $d = 2$. For $d \geq 3$ this conjecture can be seen as a result of the extra freedom that arises from having a group of units of rank $d - 1 \geq 2$, and is actually a consequence of a stronger conjecture formulated by Margulis [59].

We will call an ideal class of a field K δ -bad if it does not admit a representative of norm $< \delta \cdot \text{disc}(K)^{1/2}$. Let $h_\delta(K)$ be the number of δ -bad ideal classes and let R_K denote the regulator of the field K . In [29, I] it is shown that:

Theorem 12. *Let $d \geq 3$, and let K denote a totally real number field of degree d . For all $\varepsilon, \delta > 0$ we have*

$$\sum_{\text{disc}(K) < X} R_K h_\delta(K) \ll X^\varepsilon. \tag{6.2}$$

In particular:

1. “Conjecture 2 is true for almost all fields”: the number of fields K with discriminant $\leq X$ for which $m(K) \geq \delta \cdot \text{disc}(K)^{1/2}$ is $O_{\varepsilon, \delta}(X^\varepsilon)$, for any $\varepsilon, \delta > 0$.
2. “Conjecture 2 is true for all fields with large regulator”: If $(K_i)_i$ is any sequence of fields for which $\liminf_i \frac{\log R_{K_i}}{\log \text{disc}(K_i)} > 0$, then $m(K_i) = o(\text{disc}(K_i)^{1/2})$.

This is connected to the considerations of Section 6.4 in the following way. Consider the case $d = 3$; to a real cubic field K and suitable additional data we have associated a collection of compact A -orbits $\Gamma_K \subset \text{PGL}_3(\mathbb{Z}) \backslash \text{PGL}_3(\mathbb{R})$, indexed by the class group of K . The key point is the following: the question of the minimal norm of a representative for a given ideal class is closely related to the question of how far the associated A -orbit penetrates into the “cusp” of the noncompact space $\text{PGL}_3(\mathbb{Z}) \backslash \text{PGL}_3(\mathbb{R})$. This allows a geometric reformulation of Theorem 12 that is amenable to analysis by the methods of Section 6.4.

7. Ergodic theory vs. harmonic analysis

In this concluding section, we briefly compare dynamical methods and harmonic analysis.

Fundamentally, the most general type of problem we are considering is the following: let $H \subset G$ be a subgroup of a semisimple \mathbb{Q} -group G ; understand the “distribution” of $H(\mathbb{Q}) \backslash H(A)$ inside $G(\mathbb{Q}) \backslash G(A)$. Indeed such problems arise naturally in a large number of arithmetic questions. Two possible approaches to these questions are the following:

1. Ergodic. Here we choose a suitable finite set of places S and apply results constraining $H(\mathbb{Q}_S)$ -invariant measures on $G(\mathbb{Q}) \backslash G(A)$.
2. Harmonic-analytic. Here we choose a suitable basis φ_i for functions on $G(\mathbb{Q}) \backslash G(A)$ and compute the “periods”

$$\int_{H(\mathbb{Q}) \backslash H(A)} \varphi_i, \quad (7.1)$$

the main goal being to have “good” quantitative upper bound for (7.1)

Moreover, we note that there is considerable potential for interaction between the two approaches: in [29, I], harmonic analysis is used to control escape of mass issues, while in [75] quantitative ideas from ergodic theory are used to give estimates on periods like (7.1).

In any case, the following general principles tend to apply:

1. If H is “a large enough subgroup” of G (say if H acts with an open orbit on the flag variety of G), the periods (7.1) will often have “arithmetic significance”, i.e. are often interpretable in terms of quantities of arithmetic interest such as L -functions and one can at least *hope* for complete, quantitative results via harmonic analysis. Note that, in addition to “standard harmonic analysis,” one should keep in mind the possibility of using an extra important trick: namely, of using *equalities between periods on different groups*. That is to say: often there will be another pair $(H' \subset G')$ with the property that, for each φ_i as above, one may associate functions φ'_i on $G'(\mathbb{Q}) \backslash G'(A)$ so that

$$\int_{H(\mathbb{Q}) \backslash H(A)} \varphi_j = \int_{H'(\mathbb{Q}) \backslash H'(A)} \varphi'_j.$$

The correspondence $\varphi \leftrightarrow \varphi'$ is usually related to functoriality. Thereby one can study the H -periods on G by switching to G' .

2. If H is not a torus, one often apply profitably Ratner’s theorem in the ergodic approach and get strong, *although non-quantitative* results. We have not discussed any examples of this in the present article; a nice instance is [30].
3. If H is a torus, the emerging theory of measure rigidity for torus actions (see in particular [27], [28]) may offer a substitute for Ratner theory. This requires an extra input, positive entropy, and has two further disadvantages (compared to “Ratner theory”) that might be noted:
 - (a) At present there is no good general way to control, either escape of mass when $G(\mathbb{Q}) \backslash G(A)$ is noncompact, or the related phenomenon of concentration on embedded subgroups.
 - (b) One needs to have “Linnik’s condition,” i.e. a fixed set of places S such that $H(\mathbb{Q}_v)$ is noncompact for $v \in S$.

Eventually, as a rough rule, the strength of ergodic theory is that it can handle orbits of very “small” subgroups – which at present seem far beyond the reach of traditional harmonic analysis– and its weakness is that it is not (yet) quantitative. On the other hand, the strength of harmonic analysis is that it imports all the rich internal structure of automorphic forms.

For instance, “why” is it that harmonic-analytic approaches to Theorem 1 have been able to avoid a Linnik-type condition? Our perspective to this question is that the Waldspurger formula (2.5) expresses a period over a non-split torus T_d in terms of the L -function $L(\pi, 1/2) \times L(\pi \times \chi_d, 1/2)$. But, by Hecke theory, the second L -function is expressible as a (χ_d -twisted) period of a form in π over a *split* torus in $\mathrm{GL}_2(\mathbb{Q}) \backslash \mathrm{GL}_2(A_{\mathbb{Q}})$! Thereby one has an *equality* between a period over a *nonsplit* torus T_d and a twisted period over a *split* torus T_{split} . This equality, which is an instance of functoriality, is part of the reason that one is able to sidestep the problem of small split primes that plagues any direct analysis of T_d .

References

- [1] Baruch, E. M., Mao, Z., Central value of automorphic L -functions. Preprint, 2003.
- [2] Bernstein, J., Reznikov, A., Analytic continuation of representations and estimates of automorphic forms. *Ann. of Math. (2)* **150** (1) (1999), 329–352.
- [3] —, Estimates of automorphic functions. *Mosc. Math. J.* (4) (1) (2004), 19–37, 310
- [4] —, Periods, subconvexity and representation theory. *J. Differential. Geometry* **70** (1) (2005), 129–141.
- [5] Blomer, V., Non-vanishing of class group L -functions at the central point. *Ann. Inst. Fourier (Grenoble)* **54** (4) (2004), 831–847.
- [6] —, Shifted convolution sums and subconvexity bounds for automorphic L -functions. *Internat. Math. Res. Notices* (2004), (73), 3905–3926.
- [7] Blomer, V., Harcos, G., Michel, Ph., A Burgess-like subconvex bound for twisted L -functions. *Forum Math.* (2006), to appear.
- [8] —, Bounds for automorphic L -functions. Preprint, 2006.
- [9] Bykovskii, V. A., A trace formula for the scalar product of Hecke series and its applications. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **226**, (1996), 14–36, 235–236; English transl. *J. Math. Sci. (New York)* **89** (1) (1998), 915–932.
- [10] Clozel, L., Oh, H., Ullmo, E., Hecke operators and equidistribution of Hecke points. *Invent. Math.* **144** (2) (2001), 327–351.
- [11] Clozel, L., Ullmo, E., Équidistribution des points de Hecke. In *Contributions to automorphic forms, geometry, and number theory*, Johns Hopkins University Press, Baltimore, MD, 2004, 193–254.
- [12] —, Équidistribution de mesures algébriques, *Compositio Math.* **141** (5) (2005), 1255–1309.
- [13] Cogdell, J. W., On sums of three squares. *J. Théor. Nombres Bordeaux* **15** (1) (2003), 33–44.
- [14] Cohen, P. B., Hyperbolic equidistribution problems on Siegel 3-folds and Hilbert modular varieties. *Duke Math. J.* **129** (1) (2005), 87–127.
- [15] Conrey, J. B., Iwaniec, H., The cubic moment of central values of automorphic L -functions. *Ann. of Math. (2)* **151** (3) (2000), 1175–1216.
- [16] Cornut, C., Vatsal, V., Nontriviality of Rankin-Selberg L -functions and CM points. In *Proceedings of the LMS symposium: L -functions and Galois representations* (2004), to appear.
- [17] Duke, W., Hyperbolic distribution problems and half-integral weight Maass forms. *Invent. Math.* **92** (1) (1988), 73–90.
- [18] Duke, W., Friedlander, J. B., Iwaniec, H., A quadratic divisor problem. *Invent. Math.* **115** (2) (1994), 209–217.
- [19] —, Bounds for automorphic L -functions. *Invent. Math.* **112** (1) (1993), 1–8.
- [20] —, Bounds for automorphic L -functions. II. *Invent. Math.* **115** (2) (1994), 219–239.
- [21] —, Bilinear forms with Kloosterman fractions. *Invent. Math.* **128** (1) (1997), 23–43.
- [22] —, Bounds for automorphic L -functions. III. *Invent. Math.* **143** (2) (2001), 221–248.
- [23] —, The subconvexity problem for Artin L -functions. *Invent. Math.* **149** (3) (2002), 489–577.
- [24] —, Class group L -functions. *Duke Math. J.* **79** (1) (1995), 1–56.

- [25] Duke, W., Rudnick, Z., Sarnak, P., Density of integer points on affine homogeneous varieties, *Duke Math. J.* **71** (1) (1993), 143–179.
- [26] Duke, W., Schulze-Pillot, R., Representation of integers by positive ternary quadratic forms and equidistribution of lattice points on ellipsoids. *Invent. Math.* **99** (1) (1990), 49–57.
- [27] Einsiedler, M., Katok, A., Lindenstrauss, E., Invariant measures and the set of exceptions to Littlewoods conjecture. *Ann. of Math.* (2006), to appear.
- [28] Einsiedler, M., Lindenstrauss, E., Diagonalizable flows on locally homogeneous spaces and number theory. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 1731–1759.
- [29] Einsiedler, M., Lindenstrauss, E., Michel, Ph., Venkatesh, A., Distribution properties of compact orbits on homogeneous spaces I, II & III. In preparation.
- [30] Eskin, A., Oh, H., Representations of integers by an invariant polynomial and unipotent flows. Preprint, 2003.
- [31] Eskin, A., McMullen, C., Mixing, counting, and equidistribution in Lie groups. *Duke Math. J.* **71** (1) (1993), 181–209.
- [32] Eskin, A., Mozes, Sh., Shah, N., Unipotent flows and counting lattice points on homogeneous varieties. *Ann. of Math.* (2) **143** (2) (1996), 253–299.
- [33] Eskin, A., Counting problems and semisimple groups. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 539–552.
- [34] Friedlander, J. B., Bounds for L -functions. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 1, Birkhäuser, Basel 1995, 363–373.
- [35] Gan, W. T., Oh, H., Equidistribution of integer points on a family of homogeneous varieties: a problem of Linnik. *Compositio Math.* **136** (3) (2003), 323–352.
- [36] Gross, B., Heights and the special values of L -series. In *Number theory* (Montreal, Que., 1985), CMS Conf. Proc. 7, Amer. Math. Soc., Providence, RI, (1987, 115–187.
- [37] Gross, B., Zagier, D., Heegner points and derivatives of L -series. *Invent. Math.* **84** (2) (1986), 225–320.
- [38] Harcos, G., An additive problem in the Fourier coefficients of cusp forms. *Math. Ann.* **326** (2) (2003), 347–365.
- [39] Harcos, G., Michel, Ph., The subconvexity problem for Rankin-Selberg L -functions and equidistribution of Heegner points II. *Invent. Math.* **163** (3) (2006), 581–655.
- [40] Harris, M., Kudla, S. S., The central critical value of a triple product L -function. *Ann. of Math.* (2) **133** (3) (1991) 605–672.
- [41] Iwaniec, H., Fourier coefficients of modular forms of half-integral weight. *Invent. Math.* **87** (2) (1987), 385–401.
- [42] —, The spectral growth of automorphic L -functions. *J. Reine Angew. Math.* **428** (1992), 139–159.
- [43] —, Harmonic analysis in number theory. In *Prospects in mathematics* (Princeton, NJ, 1996), Amer. Math. Soc., Providence, RI, 1999, 51–68.
- [44] Iwaniec, H., Sarnak, P., Perspectives on the analytic theory of L -functions. *Geom. Funct. Anal.* (2000) Special Volume, 705–741.

- [45] Jutila, M., Convolutions of Fourier coefficients of cusp forms. *Publ. Inst. Math. (Beograd) (N.S.)* **65** (79) (1999), 31–51.
- [46] Jutila, M., Motohashi, Y., Uniform bounds for Hecke L -functions. *Acta Math.* **195** (2005), 61–115.
- [47] Katok, S., Sarnak, P., Heegner points, cycles and Maass forms. *Israel J. Math.* **84** (1–2) (1993), 193–227.
- [48] Kim, H., Functoriality for the exterior square of GL_4 and the symmetric fourth of GL_2 . *J. Amer. Math. Soc.* **16** (1) (2003), 139–183.
- [49] Kim, Henry H., Shahidi, Freydoon, Functorial products for $GL_2 \times GL_3$ and the symmetric cube for GL_2 . *Ann. of Math. (2)* **155** (3) (2002), 837–893.
- [50] Kohnen, W., Zagier, D., Values of L -series of modular forms at the center of the critical strip. *Invent. Math.* **64** (2) (1981), 175–198.
- [51] Kowalski, E., Michel, Ph., VanderKam, J., Mollification of the fourth moment of automorphic L -functions and arithmetic applications. *Invent. Math.* **142** (1) (2000), 95–151.
- [52] —, Rankin-Selberg L -functions in the level aspect. *Duke Math. J.* **114** (1) (2002), 123–191.
- [53] Lindenstrauss, E., Invariant measures and arithmetic quantum unique ergodicity. *Ann. of Math. (2)* **163** (1) (2006), 165–219.
- [54] Linnik, Yu. V., The asymptotic distribution of reduced binary quadratic forms in relation to the geometries of Lobačevskiĭ. III. *Vestnik Leningrad. Univ.* **10** (8) (1955), 15–27.
- [55] —, Asymptotic-geometric and ergodic properties of sets of lattice points on a sphere. *Amer. Math. Soc. Transl. (2)* **13** (1960), 9–27.
- [56] —, Additive problems and eigenvalues of the modular operators. In *Proceedings of the International Congress of Mathematicians* (Stockholm, 1962), Inst. Mittag-Leffler, Djursholm 1963, 270–284.
- [57] —, *Ergodic properties of algebraic fields*. *Ergeb. Math. Grenzgeb.* 45, Springer-Verlag, New York 1968.
- [58] Linnik, Yu. V., Skubenko, B. F., Asymptotic distribution of integral matrices of third order. *Vestnik Leningrad. Univ. Ser. Mat. Meh. Astronom.* **19** (3) (1964), 25–36.
- [59] Margulis, G., Problems and conjectures in rigidity theory. In *Mathematics: frontiers and perspectives*, Amer. Math. Soc., Providence, RI, 2000, 161–174.
- [60] Michel, Ph., The subconvexity problem for Rankin-Selberg L -functions and equidistribution of Heegner points. *Ann. of Math. (2)*, **160** (1) (2004), 185–236.
- [61] Michel, Ph., Venkatesh, A., Periods, subconvexity and equidistribution. In preparation.
- [62] —, Heegner points and nonvanishing of Rankin-Selberg L -functions. Preprint, 2006.
- [63] Molteni, G., Upper and lower bounds at $s = 1$ for certain Dirichlet series with Euler product. *Duke Math. J.* **111** (1) (2002), 133–158.
- [64] Motohashi, Y., *Spectral theory of the Riemann zeta-function*. Cambridge Tracts in Math., Cambridge University Press, Cambridge 1997.
- [65] Oh, Hee, Hardy-Littlewood system and representations of integers by an invariant polynomial. *Geom. Funct. Anal.* **14** (4) (2004), 791–809.
- [66] Popa, A., Central values of Rankin L -series over real quadratic fields. *Compositio Math.* (2006), to appear.

- [67] Reznikov, A., Rankin-Selberg without unfolding. Preprint, 2005.
- [68] Sarnak, P., Diophantine problems and linear groups. In *Proceedings of the International Congress of Mathematicians* (Kyoto, 1990), Vol. I, The Mathematical Society of Japan, Tokyo, Springer-Verlag, Tokyo, 1991, 459–471.
- [69] —, Integrals of products of eigenfunctions. *Internat. Math. Res. Notices* **1994** (6) (1994), 251–260.
- [70] —, Estimates for Rankin-Selberg L -functions and quantum unique ergodicity. *J. Funct. Anal.* **184** (2) (2001)n 419–453.
- [71] Skubenko, B. F., The asymptotic distribution of integers on a hyperboloid of one sheet and ergodic theorems. *Izv. Akad. Nauk SSSR Ser. Mat.* **26** (1962), 721–752.
- [72] Ullmo, E., Théorie ergodique et géométrie arithmétique, In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 197–206.
- [73] Vatsal, V., Uniform distribution of Heegner points. *Invent. Math.* **148** (1) (2002), 1–46.
- [74] —, Special values of anticyclotomic L -functions. *Duke Math. J.* **116** (2) (2003), 219–261.
- [75] Venkatesh, A., Sparse equidistribution problems, period bounds, and subconvexity. Preprint, 2005.
- [76] Waldspurger, J.-L., Sur les valeurs de certaines fonctions L automorphes en leur centre de symétrie. *Compositio Math.* **54** (2) (1985), 173–242.
- [77] Watson, T., Rankin triple products and quantum chaos, *Ann. of Math.* (2006), to appear.
- [78] Xue, H., Central values of Rankin L -functions. Preprint, 2005.
- [79] Yafaev, A., A conjecture of Yves André’s. *Duke Math. J.* **132** (3) (2006), 393–407.
- [80] Zhang, S., Heights of Heegner points on Shimura curves. *Ann. of Math.* (2) **153** (1) (2001), 27–147.
- [81] —, Gross-Zagier formula for GL_2 . *Asian J. Math.* **5** (2) (2001) 183–290.
- [82] —, Equidistribution of CM-points on quaternion Shimura varieties. Preprint, 2004.

Département de Mathématiques, Université Montpellier II, Place E. Bataillon,
34095 Montpellier cedex, France

Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540, U.S.A.