AN INTRODUCTION TO THE LINNIK PROBLEMS

W. Duke
UCLA

Abstract. This paper is a slightly enlarged version of a series of lectures on the Linnik problems given at the SMS–NATO ASI 2005 Summer School on Equidistribution in Number Theory.

Key words: Linnik problem, half-integral weight, CM points

1. Introduction

In these lectures I will discuss the classical Linnik problems about the distribution of lattice points on a sphere and analogous hyperbolic problems associated to binary quadratic forms. These problems were introduced by Linnik and are discussed in his book *Ergodic Properties of Algebraic Fields*. (Linnik, 1968) Linnik applied an intricate ergodic method to solve them subject to a certain condition. In 1987 Iwaniec (Iwaniec, 1987) made a breakthrough in the theory of modular forms of half-integral weight that allowed the Linnik problems to be solved unconditionally using more traditional modular forms methods (Duke, 1988). These methods have since been much further developed in the more general context of subconvexity estimates for L-functions, where they have far-ranging implications/applications. My main purpose is to give an exposition of the original modular forms approach emphasizing the original ideas, which have an intuitive appeal. I will only introduce briefly the connection with L-functions. Recently there has been striking progress by a number of mathematicians in the analytic theory of L-functions in connection with various equidistribution problems. Hopefully, these lectures will provide some background for these developments, and serve as a rough guide to help those interested in pursuing details. An excellent exposition of many of the topics treated here is (Sarnak, 1990).

2. The Linnik Problems

THE SPHERE

Consider the lattice points $\alpha \in \mathbb{Z}^3$ with $|\alpha|^2 = x_1^2 + x_2^2 + x_3^2$, for $\alpha = (x_1, x_2, x_3)$. The set $\Omega_n = \{x = \alpha/|\alpha|; \alpha \in \mathbb{Z}^3; |\alpha|^2 = n\}$ for $n \in \mathbb{Z}^+$ lies on the unit sphere

S^2. By a classical result of Legendre Ω_n is non-empty iff $n \neq 4^a(8b+7)$ for a and b integers, a non-negative. Linnik asked whether the set Ω_n subject to the condition that $n \equiv 1, 2, 3, 5, 6 \pmod 8$ is uniformly distributed with respect to (normalized) Lebesgue measure $d\sigma$ on S^2; is it the case given a reasonable subset of S^2 that the proportion of points in it from Ω_n approaches the measure of the set as $n \to \infty$? Linnik was able to prove this using his "ergodic method" but subject to the condition required by the method that the Legendre symbol $(n/p) = 1$ for a fixed odd prime p. An advance made by Iwaniec in the estimation of Fourier coefficients of cusp forms of half-integral weight later allowed this condition to be removed. To state this, it is convenient to couch the uniform distribution property in terms of the approximation of the integral of a test function by "Riemann sums." For simplicity I will restrict attention here to the most interesting case where n is square-free.

THEOREM A. *Suppose that $f \in C^\infty(S^2)$. Then, as $n \to \infty$ with n square-free and $n \not\equiv 7 \pmod 8$,*

$$\frac{1}{\#\Omega_n} \sum_{x \in \Omega_n} f(x) \to \int_{S^2} f \, d\sigma.$$

I will spend most of the lectures explaining, modulo many technical details, the proof of this result. It should be pointed out that one may ask the same question about the lattice points on a ellipsoid given by a positive definite integral ternary quadratic form. Then, most of the interest shifts to the question of characterizing by mean of congruences those integers n that are represented by the form. For square-free n, the analytic techniques used to prove Theorem A apply directly, but for general n the issue becomes quite delicate (see e.g. (Duke and Schulze-Pillot, 1990; Duke, 1997)).

CM POINTS

Another problem introduced by Linnik concerns the distribution of roots of integral quadratic equations with a large negative discriminant. Here the appropriate setting is $\pm\Gamma\backslash\mathcal{H}$, where $\mathcal{H}$ is the upper half-plane and $\Gamma = \mathrm{SL}(2, \mathbb{Z})$ is the modular group. The quadratic equations are best introduced via positive definite binary quadratic forms

$$Q = Q(x, y) = ax^2 + bxy + cy^2, \quad d = b^2 - 4ac = \operatorname{disc} Q < 0$$

with $a, b, c \in \mathbb{Z}$, $a > 0$. After Gauss, there are only finitely many Γ-equivalence classes of such forms with a given d (see (Cox, 1989)).

For a given $Q = ax^2 + bxy + cy^2$ with $\operatorname{disc} Q = d$, the root of $ax^2 + bx + c = 0$

$$z_Q = \frac{-b + \sqrt{d}}{2a} \in \mathcal{H}$$

associated to Q is called a CM point. It is readily shown that the orbit γz_Q runs over the roots of the forms equivalent to Q, where $\gamma \in \Gamma$ acts as a linear fractional map. Let $\mathcal{F}$ denote the standard fundamental domain for Γ:

$$\mathcal{F} = \left\{ z \in \mathcal{H}; -\tfrac{1}{2} \leq \operatorname{Re} z \leq 0 \text{ and } |z| \geq 1 \text{ or } 0 < \operatorname{Re} z < \tfrac{1}{2} \text{ and } |z| > 1 \right\}.$$

We shall write $\Lambda_d = \{z_Q \in \mathcal{F}; \operatorname{disc} Q = d\}$. For every $d \equiv 0, 1 (\operatorname{mod} 4)$ we can find a (principal) Q with $\operatorname{disc} Q = d$ and associated $z_Q \in \mathcal{F}$:

$$\begin{aligned} d \equiv 0(4) &: \; x^2 - \frac{d}{4} y^2, \quad z_d = \frac{\sqrt{d}}{2}, \\ d \equiv 1(4) &: \; x^2 + xy - \frac{d-1}{y} y^2, \quad z_d = \frac{-1 + \sqrt{d}}{2}. \end{aligned}$$

It is convenient to define by convention a sum over Λ_d to mean that a summand should be weighted by $\frac{1}{2}$ if $Q = a(x^2+y^2)$ and by $\frac{1}{3}$ if $Q = a(x^2+xy+y^2)$ to account for the automorphs in Γ. In particular, $H(d) = \sum_{\Lambda_d} 1$ is called the Hurwitz class number. Recall that d is said to be fundamental when it equals the discriminant of $\mathbb{Q}(\sqrt{d})$. In this case, when $d < -4$, $H(d)$ equals to the class number $h(d)$ of $\mathbb{Q}(\sqrt{d})$ Generally I will only be concerned with fundamental discriminants.

A $\mathrm{PSL}(2, \mathbb{R})$-invariant measure for $\mathcal{H}$ is given by $dx\, dy/y^2$ and

$$\iint_{\mathcal{F}} dx\, dy/y^2 = \pi/3.$$

Let us denote by $d\mu = (3/\pi)\; dx\; dy/y^2$ the normalized invariant measure. The second Linnik problem concerns the distribution of the $z_Q \in \mathcal{F}$ as $d \to -\infty$.

THEOREM B. *Suppose that $f \in C^\infty(\mathcal{H})$ is Γ-invariant and bounded on $\mathcal{H}$. Then, as $d \to -\infty$ with d a fundamental discriminant,*

$$\frac{1}{\#\Lambda_d} \sum_{z \in \Lambda_d} f(z) \to \int_{\Gamma \backslash \mathcal{H}} f\, d\mu.$$

The proof of this result is quite analogous to that of Theorem A but requires more machinery. The main reason for this is the fact that $\Gamma\backslash\mathcal{H}$ is non-compact.

There is a parallel result one can obtain for indefinite forms as $d \to +\infty$, namely the uniform distribution of closed geodesics on $\pm\Gamma\backslash\mathcal{H}$ when grouped by discriminant. In fact, the proof of Theorem B yields this result as well. This problem is in fact a revealing paradigm for more general situations in which infinite unit groups exist (see. e.g. (Cohen, 2005) and references given there).

3. Holomorphic Modular Forms of Half-Integral Weight

This subject is based on the properties of the Jacobi theta series

$$\theta(z) = \sum_{n\in\mathbb{Z}} e(n^2 z),$$

which has a product representation via the Jacobi triple product formula: write $q = e(z)$

$$\theta(z) = \prod_{n=1}^{\infty}(1 - q^{2n})(1 + q^{2n-1})^2.$$

This remarkable function satisfies for $\gamma \in \Gamma_0(4)$, where $\Gamma_0(N) = \{\gamma \in \mathrm{SL}(2,\mathbb{Z}) \colon c \equiv 0(N)\}$, the transformation formula

$$\theta(\gamma z) = j(\gamma, z)\theta(z),$$

where $j(\gamma, z) = (c/d)\varepsilon_d^{-1}(cz + d)^{1/2}$, with (c/d) the (extended) Legendre symbol, $\varepsilon_d = \begin{cases} 1, & d \equiv 1(4) \\ i, & d \equiv 3(4) \end{cases}$ and $z^{1/2} = |z|^{1/2}\exp(\frac{1}{2} i \arg z)$, with $-\pi < \arg z \leq \pi$, (see (Shimura, 1973)). Actually, Jacobi studied $\theta(z/2)$, whose relevant group is conjugate to $\Gamma_0(4)$, namely $\Gamma(2)$.

For $k \in \frac{1}{2}\mathbb{Z}^+$ and $N \equiv 0(4)$ if $2k$ is odd, a holomorphic modular form of weight k for $\Gamma_0(N)$ is a holomorphic function on $\mathcal{H}$ sit for $\gamma \in \Gamma_0(N)$

$$f(\gamma z) = j(\gamma z)^{2k} f(z),$$

together with the condition that f be holomorphic in the cusps of $\Gamma_0(N)$. The usual way to do this is to define the Fourier expansion of f in each cusp and require that no negative terms occur. This is easily done at $i\infty$, where the Fourier expansion must look like

$$f(z) = \sum_{n\geq 0} a(n)e(nz). \tag{1}$$

For other cusps and $2k$ odd this is a little bit trickier and is best done using a cover of $\mathrm{SL}(2,\mathbb{R})$ (see (Shimura, 1973) or (Koblitz, 1984)). For our purposes it is enough to impose the equivalent growth condition on the invariant function $y^{k/2}|f(z)| = F(z)$ that

$$F(z) \ll y^A + y^{-A} \quad \text{for some} \quad A \geq 0 \tag{2}$$

and all $z \in \mathcal{H}$. Let $M_k(N)$ denote the space of all such functions; it is known to be finite dimensional. The subspace of cusp forms $S_k(N)$ consists of these $F \in M_k(N)$ whose zeroth Fourier coefficient in every cusp vanishes. For $k > 0$ this is equivalent to having (2) with $A = 0$.

The proof of Theorem A relies heavily on non-trivial estimates for the Fourier coefficients of cusp forms. This turns out to be rather harder when $2k$ is odd, which is the case needed. Let us recall the trivial bound of Hecke for a cusp form f and any k:

$$|a(n)| \underset{f}{\ll} n^{k/2}. \tag{3}$$

The proof is easy. For any $y > 0$

$$a(n)e^{-2\pi ny} = \int_0^1 e(-nx)f(x+iy)\,dx$$

and so using (2) with $A = 0$ gives

$$\begin{aligned} |a(n)| &\leq e^{2\pi ny}y^{-k/2}\int_0^1 F(x+iy)\,dx \\ &\ll e^{2\pi ny}y^{-k/2}. \end{aligned}$$

Taking $y = 1/n$ gives (3).

Hecke's bound certainly can fail for non-cusp forms; consider the easiest example when weight $k = 4$ of the Eisenstein series

$$E_4(z) = c_4 \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} (cz+d)^{-4} = 1 + 240\sum_{n=1}^{\infty} \sigma_3(n)e(nz), \tag{4}$$

which has $\sigma_3(n) = \sum_{d|n} d^3$ and cannot be bounded by a constant times n^2.

It is an important fact that one can make enough modular forms via Eisenstein series to subtract off the growth of an arbitrary modular form in the cusps, leaving a cusp form. This is harder for $k = \frac{1}{2}, 1, \frac{3}{2}$, and 2 since then the Eisenstein series do not converge absolutely. In fact, in these cases one is stuck dealing with non-holomorphic modular forms. This turns out to be the main difference between Theorems A and B.

4. Theta Series With Harmonic Polynomials

The relevance of modular forms to the Linnik problems is through the concept of a *Weyl sum*. Recall that for a finite set of points X_n on $S^1 = \mathbb{R}/\mathbb{Z}$, the Weyl criterion for equidistribution of X_n with respect to Lebesgue measure as $n \to \infty$ is that for each $m \in \mathbb{Z}$, $m \neq 0$,

$$\frac{1}{\#X_n}\sum_{\theta\in X_n} e(n\theta) \to 0$$

as $n \to \infty$. The situation for our points Ω_n on S^2 is very similar. Observe that

$$\left(\frac{x+iy}{|x+iy|}\right)^m = e(m\theta)$$

and

$$\left(\frac{x-iy}{|x-iy|}\right)^m = e(-m\theta)$$

if $\theta = \arg(x+iy)/2\pi$, for $m > 0$. Now $(x+iy)^m$ and $(x-iy)^m$ are homogeneous harmonic polynomials on $\mathbb{R}^2$. This example generalizes beautifully to $\mathbb{R}^n$. In particular for $\mathbb{R}^3$ it can be shown that any $f \in C^2(S^2)$ can be uniformly approximated by a finite sum of homogeneous harmonic polynomials in $\mathbb{R}^3$ restricted to S^2 (for a proof see Stein (Stein and Weiss, 1971, Corollary 2.3, p. 141)). Thus the Weyl criteria for uniform distribution of the lattice points on S^2 requires that we prove that for any homogeneous harmonic polynomial $P(x)$ on $\mathbb{R}^3$ of degree $\ell > 0$

$$\frac{1}{\#\Omega_n} \sum_{X \in \Omega_n} P(X) \to 0 \quad \text{as} \quad n \to \infty,$$

as in Theorem A. Equivalently, we require

$$\sum_{\substack{\alpha \in \mathbb{Z}^3 \\ |\alpha|^2 = n}} P\left(\frac{\alpha}{|\alpha|}\right) = o(r_3(n))$$

where $r_3(n) = \#\{\alpha \in \mathbb{Z}^3 \colon |\alpha|^2 = n\}$.

PROPOSITION 4.1. *The theta series*

$$\theta(z, P) = \sum_{\alpha \in \mathbb{Z}^3} P(\alpha) e(|\alpha|^2 z) = \sum_n r(n; P) e(nz)$$

is a holomorphic modular form of weight $\frac{3}{2} + \ell$ *for* $\Gamma_0(4)$, *which is a cusp form if* $\ell > 0$. *Also,* $\theta(z, P) = 0$ *unless* ℓ *is even.*

Proof. See (Shimura, 1973). □

When $\ell = \deg P = 0$ we have

$$\theta(z, 1) = \theta^3(z) = \sum_{n \geq 0} r_3(n) e(nz).$$

To prove Theorem A, we need two ingredients:

(L) $r_3(n) \gg_\varepsilon n^{1/2-\varepsilon}$ for n as in Theorem A and all $\varepsilon > 0$,

(U) $|r(n, P)| \ll n^{k/2-1/4-\delta}$ for n square-free and some fixed $\delta > 0$, when $\ell > 0$.

To see this, note $\sum_{|\alpha|^2=n} P(\alpha/|\alpha|) = n^{-\ell/2} r(n;P)$ and $k/2 - \frac{1}{4} = \ell/2 + \frac{1}{2}$, so (U) says equivalently $\sum_{|\alpha|^2=n} P(\alpha/|\alpha|) \ll n^{1/2-\delta} = o(r_3(n))$. As we shall review below, (L) follows from classical results of Gauss and Siegel, but with an ineffective constant.

5. Linnik Problem for Squares and the Shimura Lift

At this point we see how far from (U) Hecke's exponent $k/2$ is. Before turning to this problem in earnest, let us treat a related problem that leads to integral weights, namely the distribution of rational points on S^2. These points are in one-one correspondence with the primitive $(\alpha_1, \alpha_2, \alpha_3) \in \mathbb{Z}^3$ with $|\alpha|^2 = m^2$ via $\alpha \mapsto (1/m)\alpha$, $m > 0$. Here m is the height of the point. This easily leads us to consider the Linnik problem on S^2 for $n = m^2$.

Building an earlier results of Stieltjes, Hurwitz showed that

$$\sum_{n=1}^{\infty} r_3(n^2) n^{-s} = 6(1 - 2^{1-s}) \frac{\zeta(s)\zeta(s-1)}{L(s, \chi_{-4})},$$

where $\chi_{-4}(\cdot) = (\frac{-4}{\cdot})$ is the Kronecker symbol. One easily derives from this that for odd n

$$r_3(n^2) \gg n, \tag{5}$$

which is even better than (L). This phenomenon was generalized by Shimura and is called the Shimura lift. In our case we can infer for $\ell > 0$ that there is a cusp form $F(z) = \Sigma a(n) e(nz)$ of weight $2\ell + 2$ for $\Gamma_0(2)$ such that

$$\sum_{n=1}^{\infty} r(n^2, P) n^{-s} = \frac{\sum_1^{\infty} a(n) n^{-s}}{L(s - \ell, \chi_{-4})}.$$

(see (Niwa, 1975)).
Thus

$$r(n^2, P) = \sum_{d|n} a(d) \mu(\tfrac{n}{d}) \chi_{-4}(\tfrac{n}{d}) (\tfrac{n}{d})^{\ell}$$

and so in place of (U) we need a bound for $a(n)$ of the form

$$|a(n)| \ll n^{\ell+1-\delta} \quad \text{as} \quad k - \tfrac{1}{2} = \ell + 1,$$

in order to beat the lower bound $r_3(n^2) \gg n$. Thus any non-trivial bound *for weight* $2\ell + 2$ cusp forms gives Theorem A for squares. It is then an easy matter to restrict to primitive points and derive the uniform distribution of rational points of a given height on S^2 as the height tends to infinity.

6. Nontrivial Estimates for Fourier Coefficients

At first look, the methods we shall apply to establish non-trivial estimates for the Fourier coefficients of cusp forms of integral and half-integral weights appear to be the same. However, there is a striking difference. Roughly speaking, one must overcome the bound given by Weil's bound for Kloosterman sums in the half-integral weight case. In fact, this bound is more appropriately called "trivial," as we will see.

The story about obtaining non-trivial bounds in the integral weight case has a complex plot. Here I will describe the Kloosterman sums approach. It will be observed that the role of Hecke operators has been ignored so far. Such an omission becomes a serious liability in the integral weight case but, since their role in the half-integral weight case is less central, at least for our purposes here, we will continue to not emphasize them.

Historically speaking, the first approach to obtaining non-trivial estimates for Fourier coefficients was via the circle method. Kloosterman produced his sums in this context and by non-trivially estimating them solved an important problem on the representations of integers by positive definite integral quadratic forms in four variables. Later it was found by Petersson and Selberg that one could take direct advantage of automorphy by constructing Poincaré series.

Consider for $\Gamma = \Gamma_0(N)$ and $m \geq 0$ the function

$$P_m(z,k) = \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} j(\gamma,z)^{-2k} e(m\gamma z),$$

which converges absolutely and uniformly on compact subsets of $\mathcal{H}$, provided that $k > 2$. It is not hard to show that $P_m \in S_k(N)$ for $m > 0$ and in fact they span S_k. Consider that for $f \in S_k(N)$ with $f(z) = \sum_1^\infty a(n)e(nz)$

$$\begin{aligned}\langle P_m, f\rangle &= \int_{\Gamma\backslash\mathcal{H}} P_m(z)\overline{f}(z)y^k \frac{dx\,dy}{y^2} \\ &= \int_{\Gamma\backslash\mathcal{H}} \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} (j(\gamma,z))^{-2k} e(m\gamma z)\overline{f}(z)y^{k-2}\,dx\,dy \\ &= \int_0^\infty \int_0^1 e(mz)\overline{f}(z)y^{k-2}\,dx\,dy \\ &= \overline{a}(m)(4\pi m)^{-k+1}\Gamma(k-1). \end{aligned} \tag{6}$$

Thus $\langle P_m, f\rangle = 0$ for all $m \geq 1 \Rightarrow a(m) = 0 \Rightarrow f \equiv 0$. It follows that it is enough for us to estimate the Fourier coefficients of P_m.

A nice calculation (see (Sarnak, 1990) or (Iwaniec, 1997)) shows that if we write

$$P_m(z,k) = \sum_{n>0} \hat{P}_m(n)e(nz)$$

then

$$\hat{P}_m(n) = 2(n/m)^{(k-1)/2} \left\{ \delta_{m,n} + 2\pi i^{-k} \sum_{\substack{c\equiv 0(N)\\ c>0}} J_{k-1}\!\left(\frac{4\pi\sqrt{mn}}{c}\right) K(m,n;c)c^{-1} \right\} \tag{7}$$

where

$$J_{k-1}(z) = \sum_{\ell\geq 0} \frac{(-1)^\ell}{\ell!\Gamma(\ell+k)} \left(\frac{z}{2}\right)^{k-1+2\ell}$$

is the J-Bessel function and

$$K(m,n;c) = \sum_{\substack{d (\mathrm{mod}\, c)\\ (d,c)=1}} \left(\frac{c}{d}\right)^{2k} \overline{\varepsilon}_d^{2k} e\left(\frac{m\overline{d}+nd}{c}\right)$$

is a Kloosterman sum of weight k, a kind of finite analogue of J_{k-1} (via an integral representation).

SUPPOSE $K > 2$ IS EVEN

This is enough to handle the Linnik problem for squares above, since the cusp form there was weight $2\ell + 2$. We shall ignore the case of odd integral k, even though in general this is very interesting.

For even k the Kloosterman sum is

$$K(m,n;c) = \sum_{\substack{d(c)\\ (d,c)=1}} e\left(\frac{m\overline{d}+nd}{c}\right)$$

and satisfies the famous Weil bound

$$|K(m,n;c)| \underset{\varepsilon}{\ll} c^{1/2+\varepsilon}.$$

The J-Bessel function satisfies for $x > 0$

$$J_{k-1}(x) \ll \min\left\{x^{k-1}, \frac{1}{\sqrt{x}}\right\}$$

so we may conclude from (7) that

$$\begin{aligned}
\hat{P}_m(n) &\ll_\varepsilon n^{(k-1)/2}\left(\sum_{c\geq 4\pi\sqrt{mn}} \left(\frac{\sqrt{n}}{c}\right)^{k-1} c^{-1/2+\varepsilon}\right) + n^{(k-1)/2}\left(\sum_{c<4\pi\sqrt{mn}} \left(\frac{c}{\sqrt{n}}\right)^{1/2} c^{-1/2+\varepsilon}\right)\\
&\ll n^{k-1}\sum_{c\geq 4\pi\sqrt{mn}} c^{1/2-k+\varepsilon} + n^{k/2-3/4}\sum_{c<4\pi\sqrt{mn}} c^{\varepsilon}\\
&\ll n^{k/2-1/4+\varepsilon}.
\end{aligned}$$

This gives the following result, hence the Linnik problem A *for squares*.

PROPOSITION 6.1. *For $k > 2$ even and $f \in S_k(N)$ with $f = \Sigma a(n)e(nz)$ we have*

$$a(n) \underset{\varepsilon}{\ll} n^{k/2-1/4+\varepsilon}.$$

Remarks.

1. Any non-trivial bound for $K(m,n;c)$ yields a non-trivial bound for $a(n)$. This is what Kloosterman accomplished.
2. Another way to obtain a non-trivial estimate is the Rankin–Selberg method. This works well for integral weights but falls short of Weil's bound.

For general $k > 2$ and $f_j = \Sigma a_j(n)e(nz)$ an ortho-normal basis for $S_k(N)$, (6) gives

$$\begin{aligned} P_m(z,k) &= \sum_{j=1}^{J} \langle P_m, f_j \rangle f_j \\ &= \frac{\Gamma(k-1)}{(4\pi m)^{k-1}} \sum_{j=1}^{J} \overline{a}_j(m) f_j(z) \end{aligned}$$

and so

$$\hat{P}_m(n,k) = \frac{\Gamma(k-1)}{(4\pi m)^{k-1}} \sum_{j=1}^{J} \overline{a}_j(m) a_j(n).$$

Writing (7) for $\hat{P}_m(n,k)$ yields the Petersson formula. It is especially useful for estimations when $n = m$:

$$\frac{\Gamma(k-1)}{(4\pi n)^{k-1}} \sum_{j=1}^{J} |a_j(n)|^2 = 1 + 2\pi i^{-k} \sum_{c \equiv 0(N)} c^{-1} J_{k-1}\left(\frac{4\pi n}{c}\right) K(n,n;c). \tag{8}$$

It is easily checked that this yields the estimate of Proposition 6.1 again.

For integral k this method reaches its limit here. One must introduce Hecke operators and interpret the Fourier coefficients of Hecke eigenforms algebraically. This led to Deligne's proof of the Ramanujan conjecture (Eichler proved the case $k = 2$).

THEOREM 6.2 (Deligne). *For $k \in \mathbb{Z}^+$ and $f \in S_k(N)$ we have*

$$a(n) \ll n^{(k-1)/2+\varepsilon}.$$

7. Salié Sums

When $2k$ is odd the Kloosterman sum still satisfies

$$|K(m,n;c)| \underset{\varepsilon}{\ll} c^{1/2+\varepsilon}$$

and Proposition 6.1 still holds, but now it is insufficient to get the Linnik problem since we needed to obtain for $\delta > 0$ and $k \geq 5/2$ the bound

$$|r(n,P)| \ll n^{k/2-1/4-\delta}. \tag{U}$$

Perhaps it is appropriate that the exponent $k/2 - \frac{1}{4}$ is in fact "trivial" in the sense that Weil's bound in this case is entirely elementary. This is due to the fact that the Kloosterman sum can be evaluated, a fact observed in special cases by Salié. This evaluation is one of the keys behind Iwaniec's result. In this section we give a recent proof of Salié's result found by Árpád Tóth via Gauss sums (Tóth, 2005).

For the case $k = \frac{3}{2} + \ell$, ℓ even, the Kloosterman sum is

$$K(m,n;c) = \sum_{d(c)} \varepsilon_d \left(\frac{c}{d}\right) e\left(\frac{md + n\overline{d}}{c}\right).$$

This sum can be evaluated in a simpler form. By (8) we only need the case $m = n$.

An application of the Chinese reminder theorem and quadratic reciprocity puts the main behaviour on the Salié sum for $q > 0$ odd (factor out the even part)

$$S(m,n;q) = \sum_{a(\operatorname{mod} q)} \left(\frac{a}{q}\right) e\left(\frac{ma + n\overline{a}}{q}\right).$$

The Jacobi symbol makes the Salié sum a finite analogue of J_{k-1} for $2k$ odd, which is elementary; for instance,

$$J_{1/2}(z) = \sqrt{\frac{2}{\pi z}} \sin z. \tag{9}$$

By changing variables when $(n,q) = 1$ we have

$$S(n,n,q) = \left(\frac{n}{q}\right) S(n^2,1,q).$$

The analogue of (9) is

PROPOSITION 7.1.

$$S(n^2,1,q) = \varepsilon_q \sqrt{q} \sum_{x^2 \equiv 1(q)} e\left(\frac{2xn}{q}\right).$$

Tóth's proof. We use the Gauss sum

$$G(a,b;q)=\sum_{x(q)} e\left(\frac{ax^2+bx}{q}\right)$$

with evaluation

$$G(a,0;q)=\varepsilon_q\sqrt{q}\left(\frac{a}{q}\right).$$

Now let $A=\sum_{x^2\equiv n^2(q)} e(2x/q)$ so that we must show

$$S(n^2,1;q)=\varepsilon_q\sqrt{q}A.$$

Now

$$\begin{aligned} A &= \frac{1}{q}\sum_{x(q)} e\left(\frac{2x}{q}\right)\sum_{a(q)} e\left(\frac{a(x^2-n^2)}{q}\right)\\ &= \frac{1}{q}\sum_{a(q)} G(a,2;q)\, e\left(\frac{-an^2}{q}\right)\\ &= \frac{1}{q}\sum_{(a,q)=1} G(a,2;q)\, e\left(\frac{-an^2}{q}\right)\end{aligned}$$

since $G(a,b;q)=0$ if $(a,q)\nmid 2$ and q is odd (exercise). But for $(a,q)=1$

$$\begin{aligned} G(a,2;q) &= e\left(\frac{-\bar{a}}{q}\right)G(a,0;q)\\ &= e\left(\frac{-\bar{a}}{q}\right)\left(\frac{a}{c}\right)\varepsilon_q\sqrt{q}\end{aligned}$$

so

$$A=\frac{\varepsilon_q}{\sqrt{q}}\sum\left(\frac{a}{c}\right)e\left(\frac{-an^2-\bar{a}}{q}\right),$$

which gives the result since $A=\bar{A}$. □

8. An estimate of Iwaniec

In 1987 Iwaniec (Iwaniec, 1987) proved

THEOREM 8.1. *Let $f\in S_k(N)$ with $2k\geq 5$ odd. Then, for n square-free,*

$$|a(n)| \ll_{\varepsilon} n^{k/2-1/4-1/28+\varepsilon}.$$

Remark. By the Shimura lift this holds for all n. It also holds for forms with $k = \frac{1}{2}, \frac{3}{2}$ but now the square-free condition is needed.

Iwaniec's estimate makes use of an equivalent form of Proposition 7.1, namely that for q odd and $(n, q) = 1$

$$S(n, n; q) = \left(\frac{n}{q}\right)\varepsilon_q \sqrt{q} \sum_{\substack{ab=q \\ (a,b)=1}} e\left(2n\left(\frac{\overline{a}}{b} - \frac{\overline{b}}{a}\right)\right). \tag{10}$$

He uses a lovely embedding idea in conjunction with the Petersson formula; cusp forms for $\Gamma_0(N)$ are also cusp forms for $\Gamma_0(M)$, if $N|M$. This, together with positivity leads to

$$(\log P)^{-1}\frac{|a(n)|^2}{n^{k-1}} \ll \frac{P}{\log P} + \sum_{P\leq p\leq 2p}\left|\sum_{c\equiv(\operatorname{mod} pN)} \frac{K(n, n; c)}{c} J_{k-1}\left(\frac{4\pi n}{c}\right)\right|. \tag{11}$$

By exploiting the bilinear form of (10) he was able to give eventually the bound

PROPOSITION 8.2.

$$\sum_{P\leq p\leq 2P} |K_{N_P}(x)| \underset{\varepsilon}{\ll} [xP^{-1/2} + xn^{-1/2} + (x + n)^{5/8}(x^{1/4}P^{3/8} + n^{1/8}x^{1/8}P^{1/4})](xnp)^{\varepsilon},$$

where

$$K_Q(x) = \sum_{\substack{c\leq x \\ c\equiv 0(Q)}} c^{-1/2}K(n, n; c)e\left(\frac{2nv}{c}\right) \quad \textit{for } v = 0, 1, -1.$$

When combined with (11), this eventually leads to Theorem 8.1. Of course, this brief description hardly does justice to Iwaniec's argument. Sarnak has given an excellent treatment of the essential ideas in (Sarnak, 1990) and for full details the best reference is Iwaniec's original paper. Iwaniec later gave a different and in some ways simpler proof of theorem (with a weaker exponent) in (Iwaniec, 1997).

9. Theorems of Gauss and Siegel

In order to complete the proof of Theorem A we must now prove (L), since (U) follows from Iwaniec's estimate with any $\delta < 1/28$. In the Disquisitiones, Gauss proved that $r_3(n)$ is related to a class number. Suppose n is square-free. Then for $d = \operatorname{disc}\mathbb{Q}(\sqrt{-n})$ Gauss's formula can be put in the simple form

$$r_3(n) = 12H(d)\left(1 - \left(\frac{d}{2}\right)\right),$$

where $(d/2)$ is the Kronecker symbol. But Siegel proved (see (Iwaniec and Kowalski, 2004)) that

$$H(d) \underset{\varepsilon}{\gg} |d|^{1/2-\varepsilon}$$

for any $\varepsilon > 0$, but with an inneffective constant. Nonetheless, this gives (L), but it should be observed that we are forced to obtain (U) with a power savings—nothing less suffices. On the other hand, *any* $\delta > 0$ is enough.

The proof of Siegel's Theorem is based on the class number formula of Dirichlet. Consider the Eisenstein series for $\Gamma = \mathrm{SL}(2, \mathbb{Z})$

$$E(z, s) = \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} (\mathrm{Im}\, \gamma z)^s, \quad \mathrm{Re} s > 1,$$

for which $\zeta(2s)E(z, s)$ has an analytic continuation with a simple pole at $s = 1$. Now

$$\zeta(2s) \sum_{z_Q \in \Lambda_d} E(z_Q, s) = \left(\frac{|d|}{4}\right)^{s/2} L(s, \chi_d)\zeta(s), \tag{12}$$

and taking residues at $s = 1$ gives the class number formula

$$H(d) = c|d|^{1/2} L(1, \chi_d).$$

Siegel's Theorem is based on properties of $L(1, \chi_d)$.

10. The Nonholomorphic Case (Duke, 1988)

The proof of Theorem B follows along similar lines as the proof of Theorem A, but now we are forced to consider non-holomorphic modular forms. The first step, the identification of the Weyl sums, is accomplished via the spectral decomposition of the hyperbolic Laplacian on $\pm\Gamma\backslash\mathcal{H}$ for $\Gamma = \mathrm{SL}(2, \mathbb{Z})$.

We are lead naturally to consider the following two types of sums

(E) $\sum_{z_Q \in \Lambda_d} E(z_Q, s)$ for $\mathrm{Re} s = \frac{1}{2}$.

(C) $\sum_{z_Q \in \Lambda_d} \varphi(z_Q)$ for φ a Maass cusp form with $\Delta\varphi = \lambda\varphi$, $\Delta = -y^2(\partial_x^2 + \partial_y^2)$.

We just saw that (E) at $s = 1$ leads to the class number formula which indeed gives the same lower bound via Siegel's Theorem that we must overcome. On $\mathrm{Re} s = \frac{1}{2}$ the problem becomes by (12) to estimate in terms of $|d|$ for some $\delta > 0$

$$L(\tfrac{1}{2} + it, \chi_d) \ll |d|^{1/4-\delta}. \tag{13}$$

This is precisely what Burgess (Burgess, 1963) accomplished in 1963, when he applied the RH for curves to get any $\delta < \frac{1}{16}$. Note that we also use the estimate $|\zeta(1 + 2it)| \gg \log(|t| + 2)^{-1}$ of de la Vallée Poussin.

To treat (C), we must generalize the theta function construction of Theorem A. This entails using a theta series for indefinite ternary forms, originally constructed by Siegel. A “theta lift” found in this context by Maass allows one to write (C) in terms of the dth Fourier coefficient of a Maass cusp form of weight $\frac{1}{2}$. An important refinement of the Maass construction was given by Katok and Sarnak (Katok and Sarnak, 1993) that identifies explicitly the eigenvalue dependence.

Although this is technically quite involved, conceptually it is not much different than the holomorphic case. One must replace the Petersson formula with a Kuznetsov formula that relates sums of Kloorterman sums to the whole (weight $\frac{1}{2}$) spectrum. This leads, with an appropriate choice of test functions, to the needed general version of Iwaniec’s estimate (Proposition 8.2).

The Linnik problem for closed geodesics on $\pm\Gamma\backslash\mathcal{H}$ mentioned before is proven at the same time since the needed Weyl integrals occur as the dth coefficients of the same half-integral weight form, where now $d > 0$. One starts as before by considering the role of the Eisenstein series in the Dirichlet class number formula for real quadratic fields.

11. Transition to Subconvexity Bounds for *L*-Functions

The appearance of Burgess’s bound (13) strongly hints that the problem of estimating non-trivially the Fourier coefficients of $\frac{1}{2}$-integral weight forms can be converted to the problem of bounding L-functions on the critical line. This is the case, with the paradigm being provided by Waldspurger’s Theorem. It turns out that in order to obtain non-trivial estimates in this way one must go beyond the convexity estimate of the Phragmen–Lindelöf Theorem, hence the name subconvexity bounds (see (Iwaniec and Sarnak, 2000)). This has led to a number of recent developments in the analytic theory of L-functions, which is currently an extremely active area.

After a series of papers by D. Friedlander and Iwaniec on GL(2) L-functions (see (Duke et al., 2002) for references), various convolution L-functions have been considered with associated equidistribution problems. For subconvexity estimates other important new contributions have been made by, among others, Bernstein, Blomer, Conrey, Harcos, Kowalski, Liu, Michel, Reznikov, Sarnak, Vanderkam, Venkatesh, Ye, (see e.g. (Michel, 2004)) for some recent references). The mixture of ergodic methods with topics around subconvexity is an exciting new direction being pursued by Lindenstrauss and Venkatesh.

12. An Application to Traces of Singular Moduli

I will end by describing a recent application of Theorem B to the asymptotics of traces of singular moduli (Duke, 2006).

Recall the classical j-function on $\mathcal{H}$

$$j(z) = \frac{(1 + 240 \sum_{n=1}^{\infty} \sigma_3(n) q^n)^3}{q \, \mathrm{II}_{n=1}^{\infty} (1 - q^n)^{24}} = q^{-1} + 744 + 196884q + \dots$$

where $q = e(z) = e^{2niz}$. Now $j(\gamma z) = j(z)$ for $\gamma \in \Gamma$ and $j(z) = j(E)$ is the j-invariant of the elliptic curve $E/\mathbb{C}$ determinant by $\mathbb{C}/L$, where $L = \{m + nz; m, n \in \mathbb{Z}\}$. For a negative discriminant d a point $z_Q \in \Lambda_d$ is called a CM point since $j(z_Q)$ is the j-invariant of the elliptic curve E which has CM by the order $\mathbb{Z}[z_d]$. In fact, all such curves occur this way. The values $j(z_Q)$ are called singular moduli and are known to be conjugate algebraic integers for $z_Q \in \Lambda_d$. Let $K = Q(\sqrt{d})$ have discriminant $-D$. The field: $K(j(z_d))$ is Abelian over K and unramified outside of (m) where $d = -Dm^2$, called a ring class field. If $d = -D$ is fundamental then $K(j(z_d))$ is the Hilbert class field of K, that is the maximal unramified Abelian extension of K whose degree is the class number $h(d)$ of K (see (Cox, 1989)). Let us restrict to the case of fundamental d. Here is a table of the first few values of $j(z_d)$ (see Table I).

Consider $\mathrm{Tr}\,(j(z_d)) = \sum_{\Lambda_d} j(z_Q)$, which for $d < -4$ fundamental is the sum of the conjugates of $j(z_d)$. Clearly $\mathrm{Tr}\,(j(z_d)) \in \mathbb{Z}$. We shall apply Theorem B to get a precise asymptotic for $\mathrm{Tr}\,(j(z_d))$. A crude asymptotic is

$$\mathrm{Tr}\, j(z_d) = (-1)^d e^{\pi\sqrt{|d|}} + O(e^{\alpha\pi\sqrt{|d|}})$$

for any fixed $\alpha > \frac{1}{2}$. This comes from an easy examination of the height of the other z_Q's in the sum and an estimation of their number. To state a much more refined result, consider the exponential sum for $c > 0$ (later alias–Salié sum)

$$S_d(c) = \sum_{x^2 \equiv d(c)} e(2x/c).$$

Note that $\frac{1}{2} S_d(4) = (-1)^d$. The refinement is

TABLE I.

d	$j(z_d)$
-3	0
-4	12^3
-7	-15^3
-8	20^3
-11	-32^3
-15	$\frac{1}{2}(-191025 - 85995\sqrt{5})$, the first irrational value

Corollary. As $d \to -\infty$ through fundamental discriminants

$$\frac{1}{h(d)}\left(\operatorname{Tr} j(z_d) - \tfrac{1}{2}\sum_{\substack{0<c<2\sqrt{d}\\ c\equiv 0(4)}} S_d(c)e^{4\pi\sqrt{|d|}/c}\right) \to 720.$$

An equivalent form of this result was conjectured recently by Bruinier, Jenkins and Ono. It is remarkable that the constant 720 is an integer!

To see that this result is a consequence of Theorem B, fix $\varepsilon > 0$ and consider for a smooth (C^∞) test function $\psi\colon \mathbb{R}^+ \to [0,1]$ that is 0 on $[0,1]$ and 1 on $[1+\varepsilon,\infty)$, the Γ-invariant Poincaré series

$$h_\varepsilon(z) = \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \psi(\operatorname{Im}\gamma z)e(-\gamma z).$$

Here Γ_∞ consists of those $\gamma\in\Gamma$ that act as translations. Clearly for $\operatorname{Im} z > 1+\varepsilon$ we have

$$h_\varepsilon(z) = e(-z)$$

and so $f(z) = j(z) - h_\varepsilon(z)$ is C^∞, Γ-invariant and bounded on $\mathcal{H}$. By Theorem B we have that as $d\to-\infty$

$$\frac{1}{h(d)}\left(\sum_{z\in\Lambda_d} j(z) - \sum_{z\in\Lambda_d} h_\varepsilon(z)\right) \to \int_{\mathcal{F}} j(z) - h_\varepsilon(z)\, d\mu.$$

Now,

$$\sum_{z\in\Lambda_d} h_\varepsilon(z) = \sum_{\operatorname{Im} z_Q>1} e(-z_Q) + O(\varepsilon h(-d)),$$

after applying again Theorem B to a suitable function.

Also,

$$\sum_{\operatorname{Im} z_Q>1} e(-z_Q) = \tfrac{1}{2}\sum_{\substack{0<c<2\sqrt{|d|}\\ c\equiv 0(4)}} S_d(c)e^{4\pi\sqrt{|d|}/c}$$

comes from the well known Gauss parametrization of roots of $x^2\equiv d(c)$.

Next we need to evaluate

$$\int_{\Gamma\backslash\mathcal{H}} j(z) - h_\varepsilon(z)\, d\mu = \lim_{y\to\infty}\int_{\mathcal{F}_Y} j(z)\, d\mu$$

where $\mathcal{F}_Y = \{z\in\mathcal{F};\operatorname{Im} z < Y\}$ since $\int_{\mathcal{F}_Y} h_\varepsilon(z)\, d\mu = 0$. Lerche, Schellenkens, and Warner showed how to evaluate such an integral using Stokes's Theorem (see (Borcherds, 1998)). One uses the Eisenstein series of weight 2: $E_2(z) = 1 - 24\sum_1^\infty \sigma_1(n)q^n$ and its non-holomorphic modular version

$$\widetilde{E}_2(z) = E_2(z) - \frac{3}{\pi}y^{-1}.$$

Since

$$\partial\widetilde{E}_2/\partial\overline{z} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)\left(\frac{-3}{\pi y}\right) = \frac{3i}{2\pi y^2}$$

and $d\overline{z}\,dz = 2i\;dx\;dy$ we get by Stokes's Theorem

$$\begin{aligned}\frac{3}{\pi}\int_{\mathcal{F}_Y} j(z)\frac{dx\,dy}{y^2} &= \int_{-1/2+iY}^{1/2+iY} j(x+iY)\widetilde{E}_2(x+iY)\,dx \\ &= \text{constant term of } j\widetilde{E}_2(x+iY) \\ &= 744 - 2Y - \frac{3}{\pi}Y^{-1} \to 720, \text{ as } Y\to\infty.\end{aligned}$$

To see that $\int_{\mathcal{F}_Y} h_\varepsilon(z)\,d\mu = 0$, simply integrate the cut-off Poincaré series

$$h_{\varepsilon,Y}(z) = \sum_{\gamma\in\Gamma_\infty\backslash\Gamma}\psi_Y(\operatorname{Im}\gamma z)e(-\gamma z)$$

where $\psi_Y(y) = \begin{cases}\psi(y), & y\le Y\\ 0, & y>Y\end{cases}$, which coincides with h_ε on $\mathcal{F}_Y$. Thus,

$$\int_{\mathcal{F}_Y} h_\varepsilon\;d\mu = \int_{\mathcal{F}} h_{\varepsilon,Y}\;d\mu = 0.$$

Acknowledgements

I wish to thank Zeév Rudnick and the referee for several corrections and suggestions. Research supported in part by NSF Grant DMS–0355564.

References

Borcherds, R. E. (1998) Automorphic forms with singularities on Grassmannians, *Invent. Math.* **132**, 491–562.

Burgess, D. A. (1963) On character sums and L-series. II, *Proc. London Math. Soc. (3)* **13**, 524–536.

Cohen, P. B. (2005) Hyperbolic equidistribution problems on Siegel 3-folds and Hilbert modular varieties, *Duke Math. J.* **129**, 87–127.

Cox, D. A. (1989) *Primes of the form* $x^2 + ny^2$, A Wiley-Interscience Publication, New York, John Wiley & Sons Inc.

Duke, W. (1988) Hyperbolic distribution problems and half-integral weight Maass forms, *Invent. Math.* **92**, 73–90.

Duke, W. (1997) Some old problems and new results about quadratic forms, *Notices Amer. Math. Soc.* **44**, 190–196.

Duke, W. (2006) Modular functions and the uniform distribution of CM points, *Math. Ann.* **334**, 241–252.

Duke, W., Friedlander, J. B., and Iwaniec, H. (2002) The subconvexity problem for Artin L-functions, *Invent. Math.* **149**, 489–577.

Duke, W. and Schulze-Pillot, R. (1990) Representation of integers by positive ternary quadratic forms and equidistribution of lattice points on ellipsoids, *Invent. Math.* **99**, 49–57.

Iwaniec, H. (1987) Fourier coefficients of modular forms of half-integral weight, *Invent. Math.* **87**, 385–401.

Iwaniec, H. (1997) *Topics in classical automorphic forms*, Vol. 17 of *Graduate Studies in Mathematics*, Providence, RI, American Mathematical Society.

Iwaniec, H. and Kowalski, E. (2004) *Analytic number theory*, Vol. 53 of *American Mathematical Society Colloquium Publications*, Providence, RI, American Mathematical Society.

Iwaniec, H. and Sarnak, P. (2000) Perspectives on the analytic theory of L-functions, *Geom. Funct. Anal.* pp. 705–741.

Katok, S. and Sarnak, P. (1993) Heegner points, cycles and Maass forms, *Israel J. Math.* **84**, 193–227.

Koblitz, N. (1984) *Introduction to elliptic curves and modular forms*, Vol. 97 of *Graduate Texts in Mathematics*, New York, Springer-Verlag.

Linnik, Y. V. (1968) *Ergodic properties of algebraic fields*, Translated from the Russian by M. S. Keane. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 45, Springer-Verlag New York Inc., New York.

Michel, P. (2004) The subconvexity problem for Rankin-Selberg L-functions and equidistribution of Heegner points, *Ann. of Math. (2)* **160**, 185–236.

Niwa, S. (1975) Modular forms of half integral weight and the integral of certain theta-functions, *Nagoya Math. J.* **56**, 147–161.

Sarnak, P. (1990) *Some applications of modular forms*, Vol. 99 of *Cambridge Tracts in Mathematics*, Cambridge, Cambridge University Press.

Shimura, G. (1973) On modular forms of half integral weight, *Ann. of Math. (2)* **97**, 440–481.

Stein, E. M. and Weiss, G. (1971) *Introduction to Fourier analysis on Euclidean spaces*, Princeton, N.J., Princeton University Press.

Tóth, Á. (2005) On the evaluation of Salié sums, *Proc. Amer. Math. Soc.* **133**, 643–645 (electronic).