



Matthew P. Young

## Weyl-type hybrid subconvexity bounds for twisted $L$ -functions and Heegner points on shrinking sets

Received June 27, 2014

**Abstract.** Let  $q$  be odd and squarefree, and let  $\chi_q$  be the quadratic Dirichlet character of conductor  $q$ . Let  $u_j$  be a Hecke–Maass cusp form on  $\Gamma_0(q)$  with spectral parameter  $t_j$ . By an extension of work of Conrey and Iwaniec, we show  $L(u_j \times \chi_q, 1/2) \ll_{\varepsilon} (q(1 + |t_j|))^{1/3+\varepsilon}$ , uniformly in both  $q$  and  $t_j$ . A similar bound holds for twists of a holomorphic Hecke cusp form of large weight  $k$ . Furthermore, we show that  $|L(1/2 + it, \chi_q)| \ll_{\varepsilon} ((1 + |t|)q)^{1/6+\varepsilon}$ , improving on a result of Heath-Brown.

As a consequence of these new bounds, we obtain explicit estimates for the number of Heegner points of large odd discriminant in shrinking sets.

**Keywords.** Subconvexity,  $L$ -functions, equidistribution, Heegner points

### 1. Introduction

#### 1.1. Cubic moments

Let  $\chi_q$  be a real, primitive character of conductor  $q$  (odd, squarefree). Suppose that  $u_j$  is a Hecke–Maass cuspidal newform of level dividing  $q$ , and that  $f \in B_k(q)$  where  $B_k(q)$  denotes the set of weight  $k$  holomorphic Hecke newforms of level dividing  $q$ .

In a remarkable paper in the analytic theory of  $L$ -functions, Conrey and Iwaniec [CI] showed

$$\sum_{f \in B_k(q)} L(f \times \chi_q, 1/2)^3 \ll_{k,\varepsilon} q^{1+\varepsilon} \quad (1.1)$$

for  $k \geq 12$ , and

$$\sum_{t_j \leq T} L(u_j \times \chi_q, 1/2)^3 \ll_{T,\varepsilon} q^{1+\varepsilon}. \quad (1.2)$$

Consequently,  $L(\pi \times \chi_q, 1/2) \ll q^{1/3+\varepsilon}$  for  $\pi$  associated to  $f$  or  $u_j$ , by the nonnegativity of these central values [W], [KZ], [KS], [Gu]. Since the conductor of the twisted  $L$ -function is  $q^2$ , this amounts to a Weyl-type subconvexity exponent (meaning the exponent is  $1/6$ , compared to the convexity bound which has exponent  $1/4$ ) in the  $q$ -aspect.

M. P. Young: Department of Mathematics, Texas A&M University, College Station, TX 77843-3368, U.S.A.; e-mail: myoung@math.tamu.edu

*Mathematics Subject Classification (2010):* 11M41, 11M06, 11F67

Along with the Maass forms, one naturally also includes the continuous spectrum furnished by the Eisenstein series which leads to bounds for Dirichlet  $L$ -functions, namely  $L(1/2 + it, \chi_q) \ll_t q^{1/6+\varepsilon}$ . This subconvexity bound of Conrey and Iwaniec gave the first improvement on the Burgess bound [Bu] of  $q^{3/16+\varepsilon}$ , for real characters. There is also work of Heath-Brown [H-B1] that improves on the Burgess bound but for moduli that factor in a favorable way.

The bounds (1.1) and (1.2) depend polynomially on  $k$  and  $T$ , respectively, but in an unspecified way (Conrey and Iwaniec state that perhaps  $k^3$  is acceptable). Motivated by some problems related to the equidistribution of Heegner points, it is desirable to obtain bounds as strong as possible in the  $T$ -aspect; see Section 2 below for further discussion of applications. Along these lines, we mention a handful of results that estimate a cubic moment in the spectral/weight aspect with fixed level (often level 1). Ivić [Iv] showed

$$\sum_{T \leq t_j \leq T+1} L(u_j, 1/2)^3 \ll T^{1+\varepsilon}, \quad (1.3)$$

where the Maass forms are of level 1. For reference, Weyl's law for  $\Gamma_0(q) \backslash \mathbb{H}$  gives

$$\sum_{T \leq t_j \leq T+1} 1 \asymp qT,$$

uniformly in both  $q, T \gg 1$  (here the notation  $f(x) \asymp g(x)$  means there exist constants  $c_1, c_2 > 0$  such that  $c_1 f(x) \leq g(x) \leq c_2 g(x)$  for all  $x$  under consideration). Ivić's approach is quite different from that of [CI], and also leads to a Weyl-type subconvexity bound for level 1 in the archimedean (spectral parameter) aspect. The weight  $k$  analog is due to Zhao Peng [Pen] who showed

$$\sum_{f \in B_k(1)} L(f, 1/2)^3 \ll k^{1+\varepsilon}.$$

Again this implies a Weyl-type subconvexity bound in the weight aspect (with fixed level 1). Analogs of these cubic moments were used by Xiaoqing Li [Li] to give the first subconvexity bound for a self-dual  $L$ -function on  $\mathrm{GL}_3$  in the  $t$ -aspect (as well as certain  $\mathrm{GL}_3 \times \mathrm{GL}_2$  Rankin–Selberg twists). Furthermore, Blomer [Bl] obtained subconvexity for twists of a self-dual  $\mathrm{GL}_3$  form by a quadratic Dirichlet character in the  $q$ -aspect. Q. Lu [Lu] has modified Xiaoqing Li's [Li] method to handle variations of (1.3) with  $t_j$  in a larger window ( $T \leq t_j \leq T + T^{3/8+\varepsilon}$ ) and for level  $q$  (however, no dependence on  $q$  is given).

Recently, Petrow [Pet] has extended the Conrey–Iwaniec bound (1.1) to all weights  $k \geq 2$  which then implies corresponding bounds for Fourier coefficients of weight  $(k+1)/2$  cusp forms. Petrow's work is complementary to our results here as we focus on large  $k$  (or  $T$ ). The main idea of Petrow's work is the development of a Motohashi-type spectral summation formula for the cubic moment, which is crucial in establishing (1.1) in the especially interesting case  $k = 2$ .

Our main result is

**Theorem 1.1.** *With notation as above, we have*

$$\sum_{f \in B_k(q)} L(f \times \chi_q, 1/2)^3 \ll_\varepsilon (kq)^{1+\varepsilon} \quad (1.4)$$

for  $\chi_q(-1) = i^k$  (otherwise the central values all vanish), and for  $T \gg 1$ ,

$$\sum_{T \leq t_j \leq T+1}^* L(u_j \times \chi_q, 1/2)^3 \ll_\varepsilon (q(1+T))^{1+\varepsilon},$$

where the star on the sum indicates the sum is restricted to even Maass forms (again, otherwise the central values vanish). Similarly for the Eisenstein series:

$$\int_T^{T+1} |L(1/2 + it, \chi_q)|^6 dt \ll_\varepsilon (q(1+T))^{1+\varepsilon}. \quad (1.5)$$

The estimate (1.4) holds for any even  $k \geq 12$ .

The new feature here compared to the previously mentioned results is that our estimates are completely uniform in  $q$  and  $T$  (or  $k$ ) together. This leads to a Weyl-type subconvexity exponent valid in a wide range in  $(q, T)$  parameter space, namely

$$L(u_j \times \chi_q, 1/2) \ll (q(1+T))^{1/3+\varepsilon},$$

and similarly for the holomorphic forms. For reference, the analytic conductor of  $L(u_j \times \chi_q, 1/2)$  is  $q^2(1+T^2)$ . The previously best-known subconvexity bound for these twisted  $L$ -functions with uniformity in both  $q$  and  $T$  (or  $k$ ) is apparently due to Blomer and Harcos [BH]. Their bound is of Burgess quality, meaning that in the  $q$ -aspect the exponent is  $3/8 + \varepsilon$  instead of  $1/3 + \varepsilon$ ; however, their result is more general in that it allows  $u_j$  (or  $f$ ) to have arbitrary level, and the twisting character does not have to be quadratic.

One pleasant feature of our proof is that in large part it handles the holomorphic and Maass cases simultaneously, and there is no need to use the intricate asymptotic expansions of Bessel functions uniform in both the index and the argument.

The bound (1.5) implies  $|L(1/2 + it, \chi_q)| \ll (q(1+|t|))^{1/6+\varepsilon}$ , which improves on results of Heath-Brown [H-B1], [H-B2] and Huxley and Watt [HW], for quadratic characters.

## 1.2. Arithmetical applications of the cubic moments

We will see in Section 2 that for applications to equidistribution of Heegner points on thin sets, the cubic moment itself is more useful than the subconvexity bound that it implies. One easy-to-state application is the following

**Proposition 1.2.** *Suppose  $-D < 0$  is a sufficiently large, odd fundamental discriminant.*

- (1) Fix  $4/9 < \eta \leq 1/2$ . There exist solutions to  $b^2 \equiv -D \pmod{4a}$  for some  $a$  and  $b$  with  $b \asymp D^\eta$ .

- (2) Fix  $1/3 < \eta \leq 1/2$ . There exist solutions to  $b^2 \equiv -D \pmod{4a}$  for some  $a$  and  $b$  with  $a \asymp D^\eta$ .

In both cases, the number of solutions is  $\gg h(-D)/D^{1/2-\eta}$ , where  $h(-D)$  is the class number, but the implied constant is ineffective.

These are both special cases of a more general result on the distribution of Heegner points in shrinking sets. See Theorem 2.1 below and following discussion for elaboration. In principle, a subconvexity bound of the form  $L(u_j \times \chi_{-D}, 1/2) \ll D^{1/2-\delta}(1+|t_j|)^B$  (for some  $\delta, B > 0$ ) would lead to a version of Proposition 1.2 with some  $\eta < 1/2$ . One purpose in this paper is to strive for a small numerical value of  $\eta$ .

## 2. Applications

In this section we discuss some arithmetical applications of the new hybrid subconvexity bound. These all can be expressed as certain lattice point estimates.

The Heegner points of (fundamental) discriminant  $-D < 0$  can be identified with the collection of  $\mathrm{SL}_2(\mathbb{Z})$ -orbits of binary quadratic forms  $ax^2 + bxy + cy^2$  of discriminant  $b^2 - 4ac = -D$ . For each such quadratic form, one has the Heegner point  $\tau = (-b + i\sqrt{D})/(2a)$  which of course can be chosen to lie inside the usual fundamental domain  $\mathcal{F}$  for  $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathbb{H}$ . Let  $\Lambda_D$  denote the set of  $\mathrm{SL}_2(\mathbb{Z})$ -classes of Heegner points of discriminant  $-D$ . The cardinality of  $\Lambda_D$  is  $h(-D)$ , the class number.

Duke [D] showed that  $\Lambda_D$  becomes equidistributed in  $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathbb{H}$  as  $D \rightarrow \infty$ , in part by extending work of Iwaniec [Iw1] bounding the Fourier coefficients of half-integral weight cusp forms. Duke used a period formula of Maass to relate the Weyl sums over the Heegner points to these Fourier coefficients.

Following the method of Harcos and Michel [HM], here we directly relate the Weyl sums to central  $L$ -values, which is provided by a formula of Waldspurger/Zhang [W], [Z]. Suppose that  $u_j$  is a Hecke–Maass cusp form, orthonormalized with  $dx dy/y^2$  (not probability measure), and define the Weyl sum

$$W_{D,u_j} = \sum_{\tau \in \Lambda_D} u_j(\tau).$$

Then (e.g., see [LMY, (5.1)] for this particular formulation entirely in terms of  $L$ -functions)

$$|W_{D,u_j}|^2 = \frac{\sqrt{D} L(u_j \times \chi_{-D}, 1/2) L(u_j, 1/2)}{2L(\mathrm{sym}^2 u_j, 1)}. \quad (2.1)$$

A similar formula holds for Eisenstein series (see [IK, (22.45)] or [LMY, (5.8)]), namely if we define  $W_{D,t} = \sum_{\tau \in \Lambda_D} E(\tau, 1/2 + it)$ , then

$$|W_{D,t}| = c(D) D^{1/4} \frac{|L(1/2 + it, \chi_{-D}) \zeta(1/2 + it)|}{|\zeta(1 + 2it)|}$$

for some function  $0 < c(D) \leq 10$  (in fact  $c(D)$  only depends on the number of units of  $\mathbb{Q}(\sqrt{-D})$ ).

Next we will set up equidistribution of Heegner points on thin sets, somewhat similarly to [Y] (which treated QUE for thin sets). In [LMY] we studied Heegner points with varying level, which is another notion of “thin”: the number of Heegner points is independent of the level  $q$  (if say  $q$  is prime), yet the volume of  $\Gamma_0(q)\backslash\mathbb{H}$  increases with  $q$ . The difference here is that we are fixing the level to be 1, and varying the archimedean aspect. For a given  $V \geq 1$ , choose a smooth and compactly-supported function  $\phi : \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H} \rightarrow \mathbb{R}$  satisfying  $|\Delta^n \phi(x + iy)| \leq C(n)V^{2n}$  for all  $n = 0, 1, 2, \dots$ . We view the list of  $C(n)$  as fixed, and  $V$  may vary with the discriminant  $-D$ .

**Theorem 2.1.** *Let  $\phi$  be as above, and let  $-D < 0$  be an odd fundamental discriminant. Then*

$$\sum_{\tau \in \Lambda_D} \phi(\tau) = h(-D) \int_{\mathcal{F}} \phi(z) \frac{3}{\pi} \frac{dx dy}{y^2} + O(\|\phi\|_2 D^{5/12+\varepsilon} V^{1+\varepsilon} + (DV)^{-100}). \tag{2.2}$$

The implied constant depends only on  $\varepsilon > 0$  and the  $C(n)$ . If one assumes the Lindelöf hypothesis for  $L(u_j \times \chi_{-D}, s)$  and  $L(s, \chi_{-D})$ , then (2.2) holds with  $D^{5/12}$  replaced by  $D^{1/4}$ .

The point is that we are able to explicitly give the dependence of the error term on  $\phi$  (previous works on equidistribution such as [D] typically treated  $\phi$  as fixed). The implied constant is in principle effective (but quite difficult to compute); the ineffectivity in Proposition 1.2 arises from ensuring the main term is larger than the error term, which in turn relies on Siegel’s lower bound on the class number.

*Proof of Theorem 2.1.* By a spectral decomposition of  $\phi$ , we have

$$\sum_{\tau \in \Lambda_D} \phi(\tau) = h(-D)\langle \phi, 3/\pi \rangle + \sum_j \langle \phi, u_j \rangle W_{D,u_j} + (\text{Eisenstein}).$$

We bound the spectral coefficients by

$$(1/4 + t_j^2)^N \langle \phi, u_j \rangle = \langle \phi, \Delta^N u_j \rangle = \langle \Delta^N \phi, u_j \rangle \ll V^{2N}.$$

Thus if  $t_j \gg V(DV)^\varepsilon$ , then  $\langle \phi, u_j \rangle$  is very small. Therefore,

$$\begin{aligned} & \left| \sum_{\tau \in \Lambda_D} \phi(\tau) - h(-D)\langle \phi, 3/\pi \rangle \right| \\ & \leq \sum_{t_j \ll V(DV)^\varepsilon} |\langle \phi, u_j \rangle W_{D,u_j}| + |(\text{Eisenstein})| + O((DV)^{-100}). \end{aligned} \tag{2.3}$$

By Cauchy’s and Bessel’s inequalities, we see that (2.3) is

$$\ll \|\phi\|_2 \left( \sum_{t_j \ll V(DV)^\varepsilon} |W_{D,u_j}|^2 \right)^{1/2} + |(\text{Eisenstein})| + (DV)^{-100}.$$

Next we use (2.1) and Hölder’s inequality (with exponents 3, 3, 3), showing that (2.3) is

$$\ll \|\phi\|_2 D^{1/4+\varepsilon} V^{1/3+\varepsilon} \times \left( \sum_{t_j \ll V(DV)^\varepsilon} L(u_j \times \chi_{-D}, 1/2)^3 \right)^{1/6} \left( \sum_{t_j \ll V(DV)^\varepsilon} L(u_j, 1/2)^3 \right)^{1/6} + \dots,$$

where the dots indicate the continuous spectrum and the error term. By Theorem 1.1, this is

$$\ll \|\phi\|_2 D^{5/12+\varepsilon} V^{1+\varepsilon} + (DV)^{-100}.$$

Note that the sums over  $t_j$  are over level 1 Maass forms, which when applying Theorem 1.1 are included into all Maass forms of level dividing  $D$ . This remark explains how the Lindelöf hypothesis replaces  $D^{5/12}$  by  $D^{1/4}$  in (2.2).

Although we have not explicitly written the contribution of the Eisenstein series, needless to say that the same bound holds for this part as for the Maass form contribution.  $\square$

Next we interpret this bound arithmetically. There are a variety of choices one can make. One particularly simple option is to choose  $\phi$  to approximate the region  $1/(2V) \leq |x| \leq 1/V, 1 \leq y \leq 2$ . One can choose  $\phi$  so that  $\|\phi\|_1 \asymp V^{-1}$  and  $\|\phi\|_2 \asymp V^{-1/2}$ . Theorem 2.1 implies that there exists a  $\tau \in \Lambda_D$  inside the support of  $\phi$  provided that  $h(-D)/V \gg D^{5/12+\varepsilon} V^{1/2}$ , which is valid for  $V \ll D^{1/18-\varepsilon}$  by Siegel’s ineffective lower bound on the class number. Therefore, there exists a solution (actually,  $\gg h(-D)/V$  solutions) to  $b^2 \equiv -D \pmod{4a}$  with  $\sqrt{D}/a \asymp 1$  and  $|b/a| \asymp V^{-1}$ . That is,  $a \asymp \sqrt{D}$  and  $b \asymp \sqrt{D}/V$ , so setting  $V = D^{1/2-\eta}$  we see that any  $4/9 < \eta \leq 1/2$  is allowable. This gives part (1) of Proposition 1.2.

Part (2) could be proved along similar lines as part (1), giving the same value  $\eta > 4/9$ . Inspired by a comment of W. Duke<sup>1</sup>, we can obtain the numerical improvement as follows. For  $z \in \mathcal{F}$ , let  $\phi(x + iy) = g(y/V)$  where  $g$  is nonnegative, has support on  $[1, 2]$ , and satisfies  $g(y) = 1$  for  $1.25 \leq y \leq 1.75$ . Then  $\langle \phi, u_j \rangle = 0$  by a direct calculation using the Fourier expansion. Similarly, the projection of  $\phi$  onto Eisenstein series picks up only the constant term of the Fourier expansion, giving

$$\langle \phi, E(\cdot, 1/2 + it) \rangle = \int_0^\infty \left( y^{1/2+it} + \frac{\zeta^*(1-2it)}{\zeta^*(1+2it)} y^{1/2-it} \right) g(y/V) \frac{dy}{y^2},$$

which is  $\ll V^{-1/2} |\tilde{g}(-1/2 + it)|$ . Therefore, (2.3) simplifies in this case to

$$\left| \sum_{\tau \in \Lambda_D} \phi(\tau) - h(-D) \langle \phi, 3/\pi \rangle \right| \ll \frac{D^{1/4}}{V^{1/2}} \int_{-\infty}^\infty |\tilde{g}(-1/2 + it)| \frac{|L(1/2 + it, \chi_{-D}) \zeta(1/2 + it)|}{|\zeta(1 + 2it)|} dt.$$

---

<sup>1</sup> Talk, “The distribution of modular closed geodesics revisited” at the Analysis, Spectra, and Number Theory Conference, December 2014.

Since  $g$  is fixed,  $\tilde{g}(-1/2 + it)$  has rapid decay, and so by the original bound of Conrey–Iwaniec we have

$$\sum_{\tau \in \Lambda_D} \phi(\tau) \gg \frac{h(-D)}{V} + O(V^{-1/2} D^{1/4+1/6+\varepsilon}).$$

The main term is larger than the error term provided  $V \ll D^{1/6-\varepsilon}$ . A Heegner point in the support of this  $\phi$  has  $a \asymp V^{-1} D^{1/2}$ , which can be made as small as  $D^{1/2-1/6+\varepsilon}$ , as claimed. These Heegner points are approaching the cusp  $\infty$  fairly quickly. The Lindelöf hypothesis would imply that  $a \gg D^\varepsilon$  is allowable.

Finally, we mention one other variation where we only ask for an upper bound on the number of Heegner points in a small box.

**Corollary 2.2.** *Fix  $x_0 + iy_0 \in \mathcal{F}$ , and let  $V \gg 1$ . Then the number of Heegner points  $\tau \in \mathcal{F}$  of discriminant  $-D$  such that  $|\tau - (x_0 + iy_0)| \ll V^{-1}$  is*

$$\ll h(-D)/V^2 + D^{5/12+\varepsilon} V^\varepsilon.$$

*The implied constant depends on  $x_0 + iy_0$  and  $\varepsilon$ .*

One pleasing feature of this upper bound is that it is  $o(D^{1/2})$  for a wide range of values of  $V$ . To prove Corollary 2.2, we apply Theorem 2.1 with a nonnegative function  $\phi$  that equals 1 on  $|\tau - (x_0 + iy_0)| \ll V^{-1}$ , such that  $\|\phi\|_1 \asymp V^{-2}$  and  $\|\phi\|_2 \asymp V^{-1}$ .

### 3. High-level sketch of the method

Here we give a brief overview of the proof focusing on the essential details. By an approximate functional equation, it suffices to show

$$\sum_{n_1, n_2, n_3 \ll (qT)^{1+\varepsilon}} \sum_{T \leq t_j \leq T+\Delta}^* w_j^* \frac{\chi_q(n_1 n_2 n_3) \lambda_j(n_1) \lambda_j(n_2) \lambda_j(n_3)}{\sqrt{n_1 n_2 n_3}} \ll (qT)^{1+\varepsilon}, \quad (3.1)$$

where  $w_j^*$  are weights arising in the Kuznetsov formula, so that  $\sum_{T \leq t_j \leq T+1} w_j^* \asymp qT$  (so the weights are  $\asymp 1$  on average, by Weyl’s law), and where  $\Delta$  is an arbitrarily small power of  $T$ . The Kuznetsov formula converts the sum over  $t_j$  into a sum of Kloosterman sums, transforming the left hand side of (3.1) to the form

$$q \sum_{n_1, n_2, n_3 \ll (qT)^{1+\varepsilon}} \sum_{c \equiv 0 \pmod{q}} \frac{1}{c} \frac{\chi_q(n_1 n_2 n_3) S(n_1 n_2, n_3; c)}{\sqrt{n_1 n_2 n_3}} B\left(\frac{4\pi \sqrt{n_1 n_2 n_3}}{c}\right), \quad (3.2)$$

where  $B$  is given as a certain integral transform. This weight function takes the rough shape

$$B(x) \approx \frac{\Delta T}{\sqrt{x}} \cos\left(x - 2\frac{T^2}{x} + \dots\right),$$

and is very small for  $x \ll \Delta T^{1-\varepsilon}$ , meaning that we may truncate the sum over  $c$  at  $\sqrt{n_1 n_2 n_3} / (\Delta T)$ . Actually, we need to treat two different types of  $B$  because we need to restrict to a sum over the even part of the spectrum, but for Theorem 1.1 both cases are fairly similar. Suppose that  $n_i \asymp N_i$  for each  $i$ , where  $N_i \ll (qT)^{1+\varepsilon}$ .

Following [CI], we apply Poisson summation to each sum over  $n_i$  modulo  $c$ , transforming (3.2) into an expression of the form

$$\sum_{m_1, m_2, m_3} \sum_{c \equiv 0 \pmod{q}} \frac{q}{c} G(m_1, m_2, m_3; c) K(m_1, m_2, m_3, c),$$

where with  $e_c(x) = e(x/c)$ , we define

$$\begin{aligned} G(m_1, m_2, m_3; c) &= c^{-3} \sum_{a_1, a_2, a_3 \pmod{c}} \chi_q(a_1 a_2 a_3) S(a_1 a_2, a_3; c) e_c(a_1 m_1 + a_2 m_2 + a_3 m_3) \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} K(m_1, m_2, m_3, c) &= (N_1 N_2 N_3)^{-1/2} \int_{\mathbb{R}^3} B\left(\frac{4\pi \sqrt{t_1 t_2 t_3}}{c}\right) e_c(-m_1 t_1 - m_2 t_2 - m_3 t_3) dt_1 dt_2 dt_3. \end{aligned}$$

Conrey and Iwaniec evaluated  $G$ , showing that if some mild coprimality restrictions are in place then with  $c = qr$ , we have

$$G(m_1, m_2, m_3; c) = \frac{\chi_q(-1) e_c(m_1 m_2 m_3)}{q^2 r} H(\bar{r} m_1 m_2 m_3; q),$$

where  $H$  is a certain two-variable complete character sum (see (9.3) below for its definition or (3.5) for a close variant). As for  $K$ , by an elaborate stationary phase analysis,

$$K(m_1, m_2, m_3, c) \approx \frac{c^2 \Delta T}{(N_1 N_2 N_3)^{1/2}} e_c(-m_1 m_2 m_3) e\left(\alpha \frac{T^2 c}{m_1 m_2 m_3} + \dots\right), \tag{3.4}$$

where  $\alpha \neq 0$  is a fixed constant. Furthermore,  $K$  is very small unless  $m_i \asymp M_i = \sqrt{N_1 N_2 N_3} / N_i$ ; this is already a square-root saving in each variable due to a reduction in length compared to  $N_i$ . Note the quite remarkable cancellation in the primary phase  $e_c(m_1 m_2 m_3)$ , as well as the common feature that  $G$  and  $K$  essentially only depend on  $m_1 m_2 m_3 / r$  and  $q$ . This feature allows for a particularly efficient separation of variables. Our work departs from [CI] in the analysis of  $K$ . When  $T$  is large, then  $K$  multiplied by  $e_c(m_1 m_2 m_3)$  is oscillatory, which makes the separation of variables nontrivial.

Next we discuss how the variables are separated, first arithmetically and then analytically. It suffices to consider a variant of  $H$  where the variables inside the summation have a coprimality condition, that is,

$$H^*(w; q) = \sum_{\substack{u, v \pmod{q} \\ (uv-1, q)=1}} \chi_q(uv(u+1)(v+1)) e_q((uv-1)w), \tag{3.5}$$



in which case from [CI, (11.9)],

$$H^*(w; q) = \frac{1}{\phi(q)} \sum_{\psi \pmod{q}} \tau(\bar{\psi}) g(\chi, \psi) \psi(w).$$

Here  $g(\chi, \psi)$  is  $O(q^{1+\varepsilon})$  by Deligne’s bound, but otherwise we do not need any properties of  $g$ . A similar separation of variables applies to  $K$  by the Mellin transform, giving

$$K(m_1, m_2, m_3, c) \approx \frac{c^{3/2} \Delta}{(N_1 N_2 N_3)^{1/4}} e_c(-m_1 m_2 m_3) L(m_1, m_2, m_3, c),$$

where

$$L(m_1, m_2, m_3, c) = \int_{|u| \ll U} \ell(u) \left( \frac{m_1 m_2 m_3}{c} \right)^{iu} du, \quad U = \frac{T^2 c}{(N_1 N_2 N_3)^{1/2}} \ll \frac{T}{\Delta},$$

and  $\ell(u) \ll 1$  slightly depends on the variables  $m_1, m_2, m_3, c$ , but for this sketch we pretend that it does not.

Therefore, in all, we have transformed (3.1) into an expression of the form

$$\sum_{\substack{m_1, m_2, m_3 \\ m_i \asymp M_i}} \sum_r \frac{C^{3/2} \Delta}{(N_1 N_2 N_3)^{1/4}} \frac{1}{\phi(q)} \sum_{\psi \pmod{q}} \frac{\tau(\bar{\psi}) g(\chi, \psi)}{(qr)^2} \psi(\bar{r} m_1 m_2 m_3) \times \int_{|u| \ll U} \ell(u) \left( \frac{m_1 m_2 m_3}{qr} \right)^{iu} du,$$

where we have further restricted to  $c = qr \asymp C$  (with  $C \ll \sqrt{N_1 N_2 N_3} / (\Delta T)$ ). At this point the expression can be pleasantly arranged into a bilinear structure as

$$\sum_{\psi \pmod{q}} \int_{|u| \ll U} \frac{\Delta q^{-1/2+\varepsilon} |\ell(u) g(\chi, \psi)|}{C^{1/2} (N_1 N_2 N_3)^{1/4}} \left| \sum_{m_1, m_2} \psi(m_1 m_2) (m_1 m_2)^{iu} \right| \times \left| \sum_{m_3, r} \psi(\bar{r} m_3) \left( \frac{m_3}{r} \right)^{iu} \right| du.$$

The hybrid large sieve inequality of Gallagher [Ga] bounds this by

$$\frac{\Delta q^{1/2+\varepsilon}}{C^{1/2} (N_1 N_2 N_3)^{1/4}} (qU + M_1 M_2)^{1/2} \left( qU + \frac{M_3 C}{q} \right)^{1/2} \left( \frac{M_1 M_2 M_3 C}{q} \right)^{1/2}.$$

An easy calculation shows this is bounded by

$$\Delta (qT/\Delta)^{1/2} (qT)^{1/2+\varepsilon} = \Delta^{1/2} (qT)^{1+\varepsilon},$$

consistent with our claim (3.1) (taking  $\Delta = T^\varepsilon$ ).

#### 4. Initial setup

Many of the early structural steps are similar to [CI] (and could now be considered standard), so in places we will refer to [CI] for the details.

Let  $T \geq 100$  and  $q$  be an odd squarefree integer, and let  $2 \leq \Delta < T/2$ . Towards (1.4), we shall obtain the bound

$$\sum_{T \leq k \leq T+\Delta}^* \sum_{f \in B_k(q)} w_f^* L(f \times \chi_q, 1/2)^3 \ll \Delta T q (Tq)^\varepsilon, \quad (4.1)$$

where the star indicates that  $k$  (necessarily even) satisfies  $\chi_q(-1) = i^k$ , which in turn simply fixes  $k \pmod{4}$  (assuming  $q$  is chosen). For the Maass forms, we will show

$$\sum_{T \leq t_j \leq T+\Delta}^* w_j^* L(u_j \times \chi_q, 1/2)^3 \ll \Delta T q (Tq)^\varepsilon, \quad (4.2)$$

where  $w_f^*$  and  $w_j^*$  are certain weights arising in the Petersson/Kuznetsov formula, satisfying  $w_f^* \gg (kq)^{-\varepsilon}$  and  $w_j^* \gg (Tq)^{-\varepsilon}$ . These are the same weights used by Conrey and Iwaniec, up to a simple scaling, so we refer to [CI] for details on these weights. The sums are over newforms of level dividing  $q$ . A similar bound holds for the Eisenstein series, namely

$$\int_T^{T+\Delta} |L(1/2 + it, \chi_q)|^6 dt \ll \Delta T q (Tq)^\varepsilon. \quad (4.3)$$

All the twisted  $L$ -functions are newforms of level  $q^2$ .

In our work, we will assume  $T \gg q^\eta$  for some fixed  $\eta > 0$ ; in practice this will mean that expressions of the form  $O(T^{-A})$  with  $A > 0$  arbitrarily large are  $O((qT)^{-A'})$ , with  $A'$  arbitrarily large. We are able to restrict to this case because Conrey and Iwaniec showed an upper bound of the form  $T^A q^{1+\varepsilon}$  (with some fixed but unspecified  $A > 0$ ) for the cubic moment, so if  $T \ll q^\eta$  with  $\eta$  arbitrarily small, then their bound is satisfactory for Theorem 1.1. We will also suppose that  $T^{\eta'} \asymp \Delta$  for some  $\eta' > 0$  fixed but arbitrarily small; note that if we prove (4.1)–(4.3) for such  $\Delta$ , then it extends to larger values of  $\Delta$  automatically by dissecting the longer interval into these shorter pieces.

In the case of Maass forms, it is technically convenient to introduce the function

$$h(t) = \frac{1}{\cosh(\frac{t-T}{\Delta})} + \frac{1}{\cosh(\frac{t+T}{\Delta})}.$$

This choice of  $h$  is analytic for  $|\operatorname{Im}(t)| < (\pi/2)\Delta$  (which we may assume is large), nonnegative on  $\mathbb{R} \cup \{iy : -1 \leq y \leq 1\}$ , and even. Furthermore,  $h(t) \gg 1$  for  $T \leq t \leq T + \Delta$ , and  $h(t)$  is very small outside of  $|t \mp T| \leq \Delta$ . To prove (4.2), it therefore suffices to show

$$\sum_{t_j} w_j^* h(t_j) L(u_j \times \chi_q, 1/2)^3 \ll \Delta T q (Tq)^\varepsilon.$$

For the holomorphic case, it is also convenient to sum over  $k$  with a smooth weight function but in this case we can take  $h(k) = w(\frac{k-1-2T}{\Delta})$  where  $w$  is any smooth, nonnegative,

compactly supported function where  $w(t) = 1$  for  $1 \leq t \leq 2$ , and zero for  $t \leq 1/2$  and  $t \geq 3$  (here we chose to center  $w$  at  $2T + 1$  to simplify some later formulas).

After some initial steps, eventually our method handles the Maass and holomorphic cases in a unified way. We shall more carefully treat the Maass case since it is a bit more complicated.

### 5. Approximate functional equation and separation of variables

Since all the  $L$ -functions under consideration have  $+1$  as the sign in their functional equation, by [IK, Theorem 5.3] we have

$$L(u_j \times \chi_q, 1/2) = 2 \sum_{n=1}^{\infty} \frac{\lambda_j(n)\chi_q(n)}{\sqrt{n}} V_{t_j}(n/q),$$

where

$$V_r(y) = \frac{1}{2\pi i} \int_{(1)} \frac{\Gamma(\frac{1/2+s+ir}{2})\Gamma(\frac{1/2+s-ir}{2})}{\Gamma(\frac{1/2+ir}{2})\Gamma(\frac{1/2-ir}{2})} G(s)(\pi y)^{-s} \frac{ds}{s}.$$

Here  $G(s)$  is an even, analytic function satisfying  $G(0) = 1$ . We choose  $G(s) = e^{s^2}$  (this  $G$  should not be confused with (3.3)). In order to more simply sum over  $t_j$ , we wish to separate the variables  $r$  and  $y$  in  $V_r(y)$ . By symmetry, we may as well suppose  $r > 0$ .

By Stirling, if  $|\text{Im}(z)| \rightarrow \infty$  (with fixed real part), but  $|s| \ll |z|^{1/2}$ , then

$$\frac{\Gamma(z+s)}{\Gamma(z)} = z^s \left( 1 + \sum_{k=1}^N \frac{P_k(s)}{z^k} + O\left(\frac{(1+|s|)^{2N+2}}{|z|^{N+1}}\right) \right)$$

for certain polynomials  $P_k(s)$  of degree  $\leq 2k$ . Since  $G(s)$  has exponential decay, we may truncate at  $|\text{Im}(s)| \ll T^\epsilon$  with only a small error in  $V_r(y)$ . With these assumptions on  $s$  and  $r$ , we have

$$\frac{\Gamma(\frac{1/2+s+ir}{2})\Gamma(\frac{1/2+s-ir}{2})}{\Gamma(\frac{1/2+ir}{2})\Gamma(\frac{1/2-ir}{2})} = (r/2)^s \left( 1 + \sum_{k=1}^N \frac{P_k(s)}{r^k} + O\left(\frac{(1+|s|)^{2N+2}}{r^{N+1}}\right) \right)$$

for a different collection of polynomials  $P_k(s)$ . For convenience, set  $P_0(s) = 1$ .

Hence

$$V_r(y) = \sum_{k=0}^N r^{-k} \frac{1}{2\pi i} \int_{(1)} G(s) P_k(s) \left(\frac{r}{2\pi y}\right)^s \frac{ds}{s} + O\left(r^{-N-1} \left(1 + \frac{y}{r}\right)^{-A}\right), \tag{5.1}$$

where the extra factor  $(1 + y/r)^{-A}$  arises from moving the contour to  $\text{Re}(s) = A$  if  $y \geq r$ , and to  $\text{Re}(s) = -1$  if  $y \leq r$ . We further refine (5.1) by approximating  $r$  by  $T$ , which is a good approximation since in our application  $h(r)$  is very small unless  $|r - T| \ll \Delta \log T$ ,

and  $\Delta$  is a small power of  $T$ . We may assume that  $|s(r - T)/T| \ll 1$ , so that by expansion into Taylor series we have

$$r^s = T^s e^{s \log(1 + \frac{r-T}{T})} = T^s \sum_{l=0}^N Q_l(s) \left(\frac{r-T}{T}\right)^l + O\left((1 + |s|)^{N+1} \left(\frac{|r-T|}{T}\right)^{N+1}\right)$$

for certain polynomials  $Q_l(s)$  of degree  $\leq l$ . We then derive for  $r \asymp T$  that

$$V_r(y) = \sum_{k=0}^N \sum_{l=0}^N \frac{1}{T^k} \left(\frac{r-T}{T}\right)^l V_{k,l} \left(\frac{y}{T}\right) + O\left(\left(\frac{1 + |r-T|}{T}\right)^{N+1} \left(1 + \frac{y}{T}\right)^{-A}\right),$$

where  $V_{k,l}$  is a function of the form

$$V_{k,l}(y) = \frac{1}{2\pi i} \int_{(1)} P_{k,l}(s) (2\pi y)^{-s} G(s) \frac{ds}{s} \tag{5.2}$$

for some polynomial  $P_{k,l}$  of degree  $\leq 2k + l$ .

In light of this form of the approximate functional equation, it suffices to show

$$\sum_{t_j}^* h_{k,l}(t_j) w_j^* \sum_{n_1, n_2, n_3=1}^{\infty} \frac{\chi_q(n_1 n_2 n_3) \lambda_j(n_1) \lambda_j(n_2) \lambda_j(n_3)}{(n_1 n_2 n_3)^{1/2}} \prod_{i=1}^3 V_i\left(\frac{n_i}{Tq}\right) + (\text{cts}) \ll \Delta(qT)^{1+\varepsilon}, \tag{5.3}$$

where (cts) represents the obvious continuous spectrum contribution (which is also non-negative), and

$$h_{k,l}(r) = ((r - T)^l T^{-k-l} + (-r - T)^l T^{-k-l}) h(r), \tag{5.4}$$

and  $V_i$  are functions of the form (5.2). To prepare this for the Kuznetsov formula, we use the Hecke relation  $\lambda_j(n_1) \lambda_j(n_2) = \sum_{d|(n_1, n_2)} \lambda_j(n_1 n_2 / d^2)$ , implying that (5.3) equals

$$\sum_{(d,q)=1} d^{-1} \sum_{n_1, n_2, n_3} V_1\left(\frac{n_1}{d^{-1}Tq}\right) V_2\left(\frac{n_2}{d^{-1}Tq}\right) V_3\left(\frac{n_3}{Tq}\right) \frac{\chi_q(n_1 n_2 n_3)}{(n_1 n_2 n_3)^{1/2}} \times \sum_j^* h_{k,l}(t_j) w_j^* \lambda_j(n_1 n_2) \lambda_j(n_3). \tag{5.5}$$

The Kuznetsov formula says

$$\sum_j^* h_{k,l}(t_j) w_j^* \lambda_j(n_1 n_2) \lambda_j(n_3) + (\text{Continuous}) = D \delta_{n_1 n_2 = n_3} + \sum_{\pm} q \sum_{c \equiv 0 \pmod{q}} \frac{S(n_1 n_2, \pm n_3; c)}{c} B^{\pm} \left(\frac{4\pi \sqrt{n_1 n_2 n_3}}{c}\right),$$

where  $D$  is the size of the diagonal term. We have chosen our weights  $w_j^*$  so that  $D \asymp \Delta Tq$ , and it is easy to see that the bound arising from the diagonal terms is of the desired magnitude. Here

$$B^+(x) = \frac{i}{\pi} \int_{-\infty}^{\infty} \left( \frac{J_{2ir}(x) - J_{-2ir}(x)}{\cosh(\pi r)} \right) h_{k,l}(r)r \, dr, \tag{5.6}$$

$$B^-(x) = \frac{4}{\pi^2} \int_{-\infty}^{\infty} K_{2ir}(x) \sinh(\pi r) h_{k,l}(r)r \, dr. \tag{5.7}$$

To prove Theorem 1.1 (that is, the Maass and Eisenstein cases of the theorem), it suffices to show the following

**Theorem 5.1.** *Let  $1 \ll N_1, N_2, N_3 \ll (qT)^{1+\varepsilon}$ , and let each  $w_i$  be a smooth weight function with support on  $x \asymp N_i$ , and satisfying  $w_i^{(k)}(x) \ll N_i^{-k}$ . Suppose that  $B$  is given by (5.6) or (5.7), with  $h = h_{k,l}$  of the form (5.4). Then*

$$\sum_{n_1, n_2, n_3} w_1(n_1)w_2(n_2)w_3(n_3)\chi_q(n_1n_2n_3) \sum_{c \equiv 0 \pmod{q}} \frac{S(n_1n_2, \pm n_3; c)}{c} B\left(\frac{4\pi\sqrt{n_1n_2n_3}}{c}\right) \ll (N_1N_2N_3)^{1/2} \Delta T (qT)^\varepsilon. \tag{5.8}$$

This follows by applying a smooth dyadic partition of unity to each  $n_i$ -sum.

Next we reduce the holomorphic case of Theorem 1.1 (that is, (1.4)) to a variation on Theorem 5.1. The separation of variables in the approximate functional equation is quite similar to the Maass case, so we omit the details. In this way, we quickly reduce to estimating an expression of the form (5.5) but with the sum over  $t_j$  replaced with a sum of the form

$$\sum_{k \equiv a \pmod{4}} w\left(\frac{k-1-2T}{\Delta}\right) \sum_{f \in B_k(q)} w_f^* \lambda_f(n_1n_2) \lambda_f(n_3). \tag{5.9}$$

By the Petersson formula, (5.9) equals

$$D\delta_{n_1n_2=n_3} + q \sum_{c \equiv 0 \pmod{q}} \frac{S(n_1n_2, n_3; c)}{c} B^{\text{holo}}\left(\frac{4\pi\sqrt{n_1n_2n_3}}{c}\right),$$

where  $D \asymp \Delta Tq$ , and

$$B^{\text{holo}}(x) = T \sum_{k \equiv a \pmod{4}} w\left(\frac{k-1-2T}{\Delta}\right) J_{k-1}(x). \tag{5.10}$$

Here the factor  $T$  arises because we chose the weights  $w_f^*$  so that  $\sum_{f \in B_k(q)} w_f^* \asymp k \asymp T$  (often in the literature, e.g. [Iw2, Theorem 3.6], the weights do not grow with  $k$  so here it is necessary to multiply by  $T$ ). At this point we have reduced the proof of (1.4) to extending Theorem 5.1 to hold for  $B = B^{\text{holo}}$ .

## 6. Exponential integrals

In Section 8 below we require extensive information on some oscillatory integrals. The stationary phase estimates we need are in principle standard, but the error terms occurring in the literature are generally not good enough for our purposes here. The issue is that Theorem 1.1 has to hold uniformly in  $(q, T)$ -parameter space. The weight function analysis is entirely in the  $T$ -aspect, and it has to commute with the  $q$ -aspect analysis. In practice this means that we require the error terms in the  $T$ -aspect to be very strong, as otherwise if  $T$  is only a small power of  $q$  then this error term is not much smaller than the main term, and we cannot apply savings in the  $q$ -aspect on the error term from the  $T$ -aspect. For this reason, we will quote some convenient results of [BKY] that have sufficiently strong errors.

First we begin with a useful definition.

**Definition 6.1.** Suppose that  $f(x_1, \dots, x_n)$  is a smooth function on  $\mathbb{R}^n$ . We say that  $f$  is *inert* if

$$x_1^{i_1} \dots x_n^{i_n} f^{(i_1, \dots, i_n)}(x_1, \dots, x_n) \ll 1,$$

with an implied constant depending on  $i_1, \dots, i_n$  and with the superscript denoting partial differentiation.

**Remark 6.2.** In practice we require that the implied list of constants is *uniform* in terms of certain parameters (e.g.,  $q, T, \Delta, N_i, C$ , but not necessarily  $\varepsilon$ ). It is then appropriate to say that  $f$  is *uniformly inert* (in terms of those parameters).

We remark that an inert function that is also say Schwartz class (e.g., with compact support) can have its variables separated almost for free, in the sense that

$$f(x_1, \dots, x_n) = \int_{\mathbb{R}^n} \widehat{f}(y_1, \dots, y_n) e(x_1 y_1) \dots e(x_n y_n) dy_1 \dots dy_n,$$

where  $\widehat{f}(y_1, \dots, y_n) \ll (1 + |y_1|)^{-A} \dots (1 + |y_n|)^{-A}$ . There exists a similar Mellin formula, of course. Note that the product of two inert functions is also inert (with new implied constants). Also if  $f(t)$  is inert and  $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ , then

$$g(x_1, \dots, x_n) = f\left(\prod_{i=1}^n (x_i / X_i)^{\alpha_i}\right) \quad (6.1)$$

is inert (with uniformity in the  $X_i$  but not the  $\alpha_i$ ). Virtually all our constructions of inert functions are variations of (6.1).

Next we synthesize both Lemma 8.1 and Proposition 8.2 of [BKY] along with some simplified choices of parameters, with the following

**Lemma 6.3.** *Suppose that  $w$  is a smooth weight function with compact support on  $[X, 2X]$ , satisfying  $w^j(t) \ll X^{-j}$  for  $X \gg 1$  (in particular,  $w$  is inert with uniformity in  $X$ ). Also suppose that  $\phi$  is smooth and satisfies  $\phi^{(j)}(t) \ll Y/X^j$  for some  $Y \gg X^\varepsilon$ . Let*

$$I = \int_{-\infty}^{\infty} w(t) e^{i\phi(t)} dt.$$

- (1) If  $\phi'(t) \gg Y/X$  for all  $t$  in the support of  $w$ , then  $I \ll_A Y^{-A}$  for  $A$  arbitrarily large.
- (2) If  $\phi''(t) \gg Y/X^2$  for all  $t$  in the support of  $w$ , and there exists  $t_0 \in \mathbb{R}$  such that  $\phi'(t_0) = 0$  (note  $t_0$  is necessarily unique), then

$$I = \frac{e^{i\phi(t_0)}}{\sqrt{\phi''(t_0)}} F(t_0) + O(Y^{-A}),$$

where  $F$  is an inert function (depending on  $A$ , but uniformly in  $X$  and  $Y$ ) supported on  $t_0 \asymp X$ .

Technically, [BKY] did not show that  $F$  is inert in all auxilliary variables. This more robust statement is proved in [KY, Lemma 5.5], which we reproduce here, with some minor simplifications.

**Lemma 6.4.** *Suppose  $w$  is inert in  $t_1, \dots, t_d$ , supported on  $t_1 \asymp X$ ,  $t_i \asymp X_i$  for  $i = 2, \dots, d$ , and  $\phi$  satisfies*

$$\frac{\partial^{a_1+\dots+a_d}}{\partial t_1^{a_1} \dots \partial t_d^{a_d}} \phi(t_1, \dots, t_d) \ll \frac{Y}{X^{a_1}} \frac{X^{a_2+\dots+a_d}}{X_2^{a_2} \dots X_d^{a_d}}$$

on the support of  $\phi$ . Assume that the conditions in Lemma 6.3(2) hold for  $t = t_1$  (uniformly in  $t_2, \dots, t_d$ ), and that  $t_0$  satisfies

$$\frac{\partial^{b_2+\dots+b_d}}{\partial t_2^{b_2} \dots \partial t_d^{b_d}} t_0 \ll_{b_2, \dots, b_d} \frac{t_0}{X_2^{b_2} \dots X_d^{b_d}} \tag{6.2}$$

for  $t_0 \asymp X$ . Then  $F$  is inert in  $t_2, \dots, t_d$ .

A simple yet common situation is when  $t_0$  is monomial in the other variables, meaning

$$t_0 = ct_2^{\alpha_2} \dots t_d^{\alpha_d}, \tag{6.3}$$

where the  $\alpha_i$  are fixed real numbers and  $c$  is some constant (depending on the tuple  $T$ ). It is easy to check that if  $t_0$  satisfies (6.3), then it satisfies (6.2). All the applications of Lemma 6.3 in this paper will have the stationary point of the form (6.3).

As a fairly direct consequence of Lemma 6.3, we shall obtain the following asymptotic formula for a certain 2-dimensional oscillatory integral. This will be used in Section 8.

**Lemma 6.5.** *Suppose  $\alpha, \beta, \gamma \in \mathbb{R}$ , let  $X \gg 1$ , and suppose that  $f_1(t_1)$  and  $f_2(t_2)$  are (uniformly) inert functions with support on  $t_1 \asymp X_1$ ,  $t_2 \asymp X_2$  with  $X_1, X_2 \gg X$ . If  $\alpha/(X_1 X_2) \gg X$ ,  $\beta X_1 \gg X$ ,  $\gamma X_2 \gg X$ , then*

$$\begin{aligned} & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1(t_1) f_2(t_2) e\left(-\frac{\alpha}{t_1 t_2} - \beta t_1 - \gamma t_2\right) dt_1 dt_2 \\ &= \left(\frac{X_1 X_2}{\beta \gamma}\right)^{1/2} e(-3(\alpha \beta \gamma)^{1/3}) f_3(\alpha, \beta, \gamma) + O_A((X_1 + X_2) X^{-A}), \end{aligned} \tag{6.4}$$

where  $f_3$  is an inert function (depending on  $A$ , uniform in  $X_1, X_2, X$ ), having support on

$$\alpha/\beta \asymp X_1^2 X_2, \quad \alpha/\gamma \asymp X_1 X_2^2. \tag{6.5}$$

More generally, if

$$\frac{|\alpha|}{X_1 X_2} \gg X, \quad |\beta| X_1 \gg X, \quad |\gamma| X_2 \gg X, \tag{6.6}$$

then the integral in (6.4) is  $O((X_1 + X_2)X^{-A})$  if  $\alpha, \beta, \gamma$  do not all have the same sign. If  $\alpha, \beta, \gamma < 0$ , then (6.4) remains valid, with the convention  $(-1)^{1/3} = -1$ . Finally, if exactly one or exactly two of the inequalities (6.6) are valid, then the integral in (6.4) is  $O((X_1 + X_2)X^{-A})$ .

*Proof.* The basic idea is to use stationary phase analysis in each variable. Let us first examine the  $t_1$ -integral, under the assumption  $\alpha, \beta, \gamma > 0$ . Let  $Z_1 = \alpha/(X_1 X_2) + \beta X_1$ . By repeated integration by parts (Lemma 6.3(1)), the integral is small (namely,  $O_A(Z_1^{-A}) = O_A(X^{-A})$ ) except possibly if  $\alpha/(X_1 X_2) \asymp \beta X_1$ . Furthermore, this argument shows the integral is very small unless  $\alpha$  and  $\beta$  have the same sign (so by symmetry,  $\alpha$  and  $\gamma$  have the same sign too). There exists a stationary point at  $t_1^0 = \sqrt{\alpha/(\beta t_2)}$ . Therefore, Lemma 6.3 gives

$$\int_{-\infty}^{\infty} f_1(t_1) e\left(-\frac{\alpha}{t_1 t_2} - \beta t_1\right) dt_1 = \frac{X_1^{1/2}}{\beta^{1/2}} e\left(-2\frac{\sqrt{\alpha\beta}}{\sqrt{t_2}}\right) F(t_2) + O(Z_1^{-A}),$$

where  $F(t_2)$  is a function that is inert in terms of  $\alpha, \beta, t_2$  (see (6.1) and surrounding discussion), and has support on an interval of the form  $t_2 \asymp \alpha/(\beta X_1^2)$ .

Next we insert this expansion into the  $t_2$ -integral, so we need to evaluate

$$\int_{-\infty}^{\infty} e\left(-2\frac{\sqrt{\alpha\beta}}{\sqrt{t_2}} - \gamma t_2\right) f_2(t_2) F(t_2) dt_2. \tag{6.7}$$

Again, (6.7) is  $O(Z_2^{-A})$ , where  $Z_2 = \alpha/(X_1 X_2) + \gamma X_2$ , except possibly if  $X_2 \asymp t_2^0 := (\alpha\beta)^{1/3}/\gamma^{2/3}$ , in which case there is a stationary point at  $t_2^0$ . So we conclude that (6.7) equals

$$\frac{X_2^{1/2}}{\gamma^{1/2}} e(-3(\alpha\beta\gamma)^{1/3}) f(\alpha, \beta, \gamma) + O(X^{-A}),$$

where  $f$  is inert (uniformly in all relevant variables), and is supported on  $\alpha/(X_1 X_2) \asymp \beta X_1$ , and  $X_2 \asymp (\alpha\beta)^{1/3}/\gamma^{2/3}$ . Note that this latter estimate can be replaced by  $\alpha/\gamma \asymp X_1 X_2^2$ , which is more symmetric. Putting everything together, and simplifying, we obtain (6.4). It is easy to derive the same formula in case  $\alpha, \beta, \gamma < 0$  by conjugation.

The final sentence of Lemma 6.5 was proved implicitly along the way, since at least one of  $Z_1$  and  $Z_2$  will be large under these conditions.  $\square$

### 7. Analytic properties of $B$

Our goal in this section is to deduce some useful estimates for  $B^\pm(x)$  and  $B^{\text{holo}}(x)$ . Even more valuable is an integral representation for  $B$  that allows us to unify the different cases.



**Lemma 7.1.** *Let  $B^+(x)$  be given by (5.6). Then for  $x \ll T$ , we have  $B^+(x) \ll \Delta x$ . Furthermore, there exists a function  $g$  depending on  $\Delta$  and  $T$ , satisfying  $g^{(j)}(x) \ll_{j,A} (1 + |x|)^{-A}$ , and such that*

$$B^+(x) = \Delta T \int_{|v| \leq \Delta^\varepsilon / \Delta} \cos(x \cosh(v)) e\left(\frac{vT}{\pi}\right) g(\Delta v) dv + O(T^{-A}). \tag{7.1}$$

Furthermore,  $B^+(x) \ll T^{-A}$  unless  $x \gg \Delta T^{1-\varepsilon}$ . In addition,  $\frac{d^k}{dx^k} B^+(x)$  is bounded by a polynomial (depending on  $k$ ) of  $\Delta, T, x$ , and  $x^{-1}$ .

One can find the asymptotic behavior of  $B^+$  given by Jutila and Motohashi [JM, (3.19)], which essentially shows that for  $x \gg \Delta T^{1-\varepsilon}$  it is of the shape  $\frac{\Delta T}{\sqrt{x}} \cos(x + \phi(x, T))$ , where  $\phi = -2T^2/x + \dots$ , with the dots indicating lower-order terms (again, with  $x \gg \Delta T^{1-\varepsilon}$ ). This comes from asymptotically evaluating the  $v$ -integral in (7.1) by stationary phase. Our plan is to simply retain this  $v$ -integral until a later stage, and apply the stationary phase method at the very end.

The bound  $B^+(x) \ll \Delta x$  for  $x \ll T$  is used only for very large values of  $c$ , giving a way to initially truncate the sum over  $c$ . Note that the error term in (7.1) is problematic for very large values of  $c$ , since one needs to recover the convergence of the sum of Kloosterman sums.

*Proof of Lemma 7.1.* Firstly, we show that  $B^+(x) \ll \Delta x$  for  $x \ll T$ . For this, we use the integral representation

$$B^+(x) = \frac{2i}{\pi} \int_{-\infty}^{\infty} \frac{J_{2ir}(x)}{\cosh(\pi r)} h(r) r dr,$$

shift the contour to  $\text{Re}(2ir) = 1 + \delta, 0 < \delta < 1$ , and apply the uniform bound (for  $t$  real and  $x > 0$ )  $J_{1+\delta+2it}(x) \ll \cosh(\pi t) \left(\frac{x}{1+|t|}\right)^{1+\delta}$ , which in turn follows directly from [GR, 8.411.4]. This gives

$$B^+(x) \ll x |h(-i/2)| + \int_{-\infty}^{\infty} \left(\frac{x}{1+|t|}\right)^{1+\delta} \left| h\left(-\frac{i}{2} - \frac{i\delta}{2} + it\right) \right| \cdot \left| -\frac{i}{2} - \frac{i\delta}{2} + it \right| dt \ll \Delta x,$$

as no other poles are crossed besides  $r = -i/2$ .

Now we treat the case where  $x$  is not extremely small. We have (see [GR, 8.411.11])

$$\frac{J_{2ir}(x) - J_{-2ir}(x)}{\cosh(\pi r)} = \tanh(\pi r) \frac{2}{\pi i} \int_{-\infty}^{\infty} \cos(x \cosh(v)) e\left(\frac{rv}{\pi}\right) dv,$$

so

$$B^+(x) = \int_{-\infty}^{\infty} \cos(x \cosh(v)) \frac{4}{\pi^2} \int_0^{\infty} r h(r) \tanh(\pi r) e\left(\frac{rv}{\pi}\right) dr dv. \tag{7.2}$$

The inner integral over  $r$  is essentially the Fourier transform of a function effectively supported on  $r = T + O(\Delta \log T)$  (outside of this region,  $h$  is exponentially small in terms of  $T$ ), so that

$$B^+(x) = \Delta T \int_{-\infty}^{\infty} \cos(x \cosh(v)) e\left(\frac{vT}{\pi}\right) g(\Delta v) dv + O(T^{-A}),$$

where  $g$  is a function satisfying  $g^{(k)}(x) \ll_{k,A} (1 + |x|)^{-A}$ . By the rapid decay of  $g$ , the contribution to the integral from  $|v| \gg \Delta^{-1+\varepsilon}$  is also  $O(T^{-A})$ , so (7.1) holds. Next we argue that  $B^+(x) \ll T^{-A}$  unless  $x \gg \Delta T^{1-\varepsilon}$ . This follows from Lemma 6.3 (repeated integration by parts). Furthermore, a minor variation of these estimates shows  $\frac{d^k}{dx^k} B^+(x) \ll_{k,A} T^{-A}$  unless  $x \gg \Delta T^{1-\varepsilon}$ , for  $k = 1, 2, \dots$

In order to show  $\frac{d^k}{dx^k} B^+(x)$  is polynomially bounded by  $\Delta, T, x$ , and  $x^{-1}$ , we take the integral representation (7.2) and treat  $|v| \leq 1$  and  $|v| > 1$  separately. For  $|v| \leq 1$  we simply differentiate inside the integral sign, and use  $\cosh(v) \leq \cosh(1)$  in this range; everything else is polynomially bounded in the other parameters. For  $|v| \geq 1$ , we cannot so simply differentiate inside the integral sign because this introduces powers of  $\cosh(v)$ , which causes convergence problems. To get around this, we simply first repeatedly integrate by parts to save a large enough power of  $e^{\pi|v|}$  to cancel these powers of  $\cosh(v)$ . An alternative approach would be to use the recursion formulas for  $\frac{d}{dx} J_\nu(x)$ .  $\square$

**Lemma 7.2.** *Let  $B^-(x)$  be given by (5.7). Then for  $x \ll T$ , we have  $B^-(x) \ll_\delta \Delta T^\delta x^{1-\delta}$  for any  $0 < \delta < 1$ . Furthermore, there exists a function  $g$  depending on  $\Delta$  and  $T$ , satisfying  $g^{(j)}(x) \ll_{j,A} (1 + |x|)^{-A}$ , and such that*

$$B^-(x) = \Delta T \int_{|v| \leq \Delta^\varepsilon / \Delta} \cos(x \sinh(v)) e\left(\frac{vT}{\pi}\right) g(\Delta v) dv + O(T^{-A}). \tag{7.3}$$

Furthermore,  $B^-(x) \ll (x + T)^{-A}$  unless  $x \asymp T$ . In addition,  $\frac{d^k}{dx^k} B^-(x)$  is bounded by a polynomial (depending on  $k$ ) in  $\Delta, T, x$ , and  $x^{-1}$ .

*Proof.* We quote the relation [GR, 8.486.10]

$$K_\nu(x) = \frac{x}{2\nu} K_{\nu+1}(x) - \frac{x}{2\nu} K_{\nu-1}(x),$$

which implies

$$B^-(x) = \frac{x}{i\pi^2} \int_{-\infty}^{\infty} (K_{1+2ir}(x) - K_{1-2ir}(x)) \sinh(\pi r) h(r) dr. \tag{7.4}$$

Next by [GR, 8.432.5], we have for  $\text{Re}(v) \geq -1/2$ ,

$$K_\nu(x) = \frac{\Gamma(\nu + 1/2)(2/x)^\nu}{\Gamma(1/2)} \int_0^\infty \frac{\cos(xv)}{(v^2 + 1)^{\nu+1/2}} dv,$$

so that if  $\text{Re}(v) = \delta > 0$ , for  $y \in \mathbb{R}$  we have, by Stirling and a trivial bound,

$$|K_{\delta+2iy}(x)| \ll_\delta \frac{(1 + |y|)^\delta}{x^\delta \cosh(\pi y)}. \tag{7.5}$$

For the part of (7.4) with  $K_{1+2ir}(x)$ , we shift the contour to  $\text{Re}(1 + 2ir) = \delta > 0$ , and apply (7.5). A similar procedure (shifting the other direction) works for the part with  $K_{1-2ir}(x)$ , so in all we obtain the bound

$$B^-(x) \ll_\delta x^{1-\delta} \int_{-\infty}^{\infty} (1 + |y|)^\delta \exp\left(-\left|\frac{y - T}{2\Delta}\right|\right) dy \ll_\delta \Delta T^\delta x^{1-\delta}.$$

Next we derive (7.3). For this we use [GR, 8.432.4], which states

$$K_{2ir}(x) \cosh(\pi r) = \frac{1}{2} \int_{-\infty}^{\infty} \cos(x \sinh v) \cos(2rv) dv.$$

The integral does not converge absolutely, but it does converge uniformly as one can integrate by parts once to save a factor  $1/\sinh(v)$ . Thus we derive

$$B^-(x) = \int_{-\infty}^{\infty} \cos(x \sinh(v)) \frac{1}{\pi^2} \int_0^{\infty} rh(r) \tanh(\pi r) e\left(\frac{rv}{\pi}\right) dr dv.$$

Compare this with (7.2). At this point, all the arguments from Lemma 7.1 carry over almost without change, so we omit the details. The main difference is that  $\sinh(v) \sim v$  for  $v = o(1)$ , so by Lemma 6.3,  $B^-(x)$  is small unless  $x \asymp T$ .  $\square$

Next we consider  $B^{\text{holo}}(x)$  defined originally by (5.10). By [Iw2, pp. 85–86], we have

$$B^{\text{holo}}(x) = \frac{T}{4} \int_{-\infty}^{\infty} W(t)c(t) dt,$$

where

$$W(t) = \int_{-\infty}^{\infty} w\left(\frac{y - 2T}{\Delta}\right) e(ty) dy,$$

$$c(t) = -2i \sin(x \sin(2\pi t)) + 2i^{-a} \sin(x \cos(2\pi t)).$$

By direct calculation,  $W(t) = \Delta e(2tT) \widehat{w}(-\Delta t)$ , and therefore there exists a function  $g$  such that  $g^{(j)}(x) \ll (1 + |x|)^{-A}$  so that

$$B^{\text{holo}}(x) = \Delta T \int_{-\infty}^{\infty} e\left(\frac{vT}{\pi}\right) g(\Delta v) (2i \sin(x \sin(v)) + 2i^{-a} \sin(x \cos(v))) dv.$$

We can thus decompose  $B^{\text{holo}}(x)$  into two pieces that are precisely of the form (7.1) and (7.3) but with  $T$  replaced by  $T'$ ,  $\cosh(v)$  replaced by  $\cos(v)$ ,  $\sinh(v)$  replaced by  $\sin(v)$ , and  $\cos(x)$  replaced by  $\sin(x)$ . In this holomorphic case it is quite easy to show  $B^{\text{holo}}(x) \ll Tx$  since  $\sin(x) \ll x$ .

We close this section by applying these results to the cubic moment problem. Recall that we wish to show (5.8). With  $w_1, w_2, w_3$  as in Theorem 5.1, define

$$S_{\pm}(N_1, N_2, N_3; C; f) = \sum_{\substack{c \asymp C \\ c \equiv 0 \pmod{q}}} \sum_{n_1, n_2, n_3} \chi_q(n_1 n_2 n_3) S(n_1 n_2, \pm n_3; c) w_1(n_1) w_2(n_2) w_3(n_3) f\left(\frac{4\pi \sqrt{n_1 n_2 n_3}}{c}\right).$$

(7.6)

Theorem 5.1 amounts to a bound on this  $S$  with  $f = B$ . Using the weak bound  $B(x) \ll Tx^{3/4}$ , and the Weil bound for Kloosterman sums, we obtain

$$S_{\pm}(N_1, N_2, N_3; C; B) \ll T(N_1 N_2 N_3)^{11/8} (qT)^{\varepsilon} C^{-1/4+\varepsilon},$$

which is satisfactory if  $C$  is a large power of  $qT$ , using  $N_i \ll (qT)^{1+\varepsilon}$ . Therefore, it suffices to bound  $S$  when  $C \ll (qT)^A$  for some fixed but large  $A$ . For a similar reason, it suffices to bound the terms with  $B(x)$  replaced by  $B_0(x)$  defined by

$$B_0(x) = \Delta T \int_{|v| \leq \Delta^\varepsilon / \Delta} e(x\phi(v)) e\left(\frac{vT}{\pi}\right) g(\Delta v) dv, \tag{7.7}$$

where

$$\phi(v) \in \pm\{\cos(v), \cosh(v), \sin(v), \sinh(v)\}. \tag{7.8}$$

By this discussion, we have reduced the proofs of (4.1)–(4.3) (and hence Theorem 1.1) to the following

**Proposition 7.3.** *Let  $B_0$  be a function of the form (7.7), where  $g$  is a function satisfying  $g^{(j)}(x) \ll_{j,A} (1 + |x|)^{-A}$ . Then for  $N_i \ll (qT)^{1+\varepsilon}$  and  $C \ll (qT)^A$  with some fixed  $A$ ,*

$$S(N_1, N_2, N_3; C; B_0) \ll (N_1 N_2 N_3)^{1/2} \Delta T (qT)^\varepsilon. \tag{7.9}$$

### 8. Analytic separation of variables

Following the method of [CI], we begin by applying Poisson summation to each of  $n_1, n_2, n_3$  modulo  $c$  in (7.6). That is,

$$S_\pm(N_1, N_2, N_3; c; B_0) = \sum_{m_1, m_2, m_3 \in \mathbb{Z}} G_\pm(m_1, m_2, m_3; c) K(m_1, m_2, m_3, c), \tag{8.1}$$

where

$$G_\pm(m_1, m_2, m_3; c) = c^{-3} \sum_{a_1, a_2, a_3 \pmod{c}} \chi_q(a_1 a_2 a_3) S(a_1 a_2, \pm a_3; c) e_c(a_1 m_1 + a_2 m_2 + a_3 m_3) \tag{8.2}$$

and

$$K(m_1, m_2, m_3, c) = \int_{\mathbb{R}^3} w_1(t_1) w_2(t_2) w_3(t_3) B_0(4\pi \sqrt{t_1 t_2 t_3} / c) e_c(-m_1 t_1 - m_2 t_2 - m_3 t_3) dt_1 dt_2 dt_3. \tag{8.3}$$

By changing variables  $a_3 \rightarrow -a_3$ , one derives

$$G_-(m_1, m_2, m_3; c) = \chi_q(-1) G_+(m_1, m_2, -m_3; c),$$

which will allow us to mainly focus on the  $+$  sign case. Let  $G = G_+$  as shorthand.

The sum  $G$  was studied extensively by Conrey and Iwaniec [CI], so we will quote their results. Our work differs from [CI] in the nature of the weight function  $K$ . Conrey and Iwaniec showed, for  $T$  essentially bounded, that  $K$  has a phase  $e_c(-m_1 m_2 m_3)$ , and

is small except for  $m_i \ll (N_1 N_2 N_3)^{1/2} / N_i$ . The difficulty in extending their work is that for  $T$  larger, one needs to more carefully treat the lower-order terms in the phase of  $K$ . More generally, what we show is that  $K$  still has a phase as mentioned above, but has a lower-order phase that still contains the factor  $m_1 m_2 m_3 / c$  in a block. This still allows us to use a Mellin technique to separate these four variables with a single integral. The important properties of  $K$  are summarized in

**Lemma 8.1.** *Suppose that  $|m_i| \asymp M_i$  for each  $i$ , and  $c \asymp C$ .*

(1) *Suppose  $\phi(v)$  is  $\pm \cos(v)$  or  $\pm \cosh(v)$ . Then*

$$K(m_1, m_2, m_3, c) = \frac{C^{3/2} \Delta T (N_1 N_2 N_3)^{1/2} e_c(-m_1 m_2 m_3)}{(M_1 M_2 M_3)^{1/2}} L(m_1, m_2, m_3, c) + O\left(T^{-A} \prod_{i=1}^3 (1 + |m_i|)^{-2}\right), \tag{8.4}$$

where  $L$  is a function that takes the form

$$L(m_1, m_2, m_3, c) = \frac{1}{V} \int_{|y| \leq T^\varepsilon} m_1^{iy_1} m_2^{iy_2} m_3^{iy_3} c^{iy_4} \int_{|u| \ll U} \ell(u, \mathbf{y}) \left(\frac{m_1 m_2 m_3}{c}\right)^{iu} du d\mathbf{y}, \tag{8.5}$$

where  $\mathbf{y} = (y_1, y_2, y_3, y_4)$ ,  $V = T$ , and

$$U = \frac{T^2 C}{(N_1 N_2 N_3)^{1/2}}. \tag{8.6}$$

Here  $\ell(u, \mathbf{y}) \ll 1$  does not depend on  $c$  or the  $m_i$ . Furthermore,  $L$  vanishes unless

$$C \ll \frac{(N_1 N_2 N_3)^{1/2}}{\Delta^{1-\varepsilon} T} \quad \text{and} \quad M_i \asymp \frac{(N_1 N_2 N_3)^{1/2}}{N_i}, \quad i = 1, 2, 3. \tag{8.7}$$

(2) *Suppose  $\phi(v)$  is  $\pm \sin(v)$  or  $\pm \sinh(v)$ . Then*

$$K(m_1, m_2, m_3, c) = \frac{C^{3/2} \Delta T (N_1 N_2 N_3)^{1/2} e_c(m_1 m_2 m_3)}{(M_1 M_2 M_3)^{1/2}} L(m_1, m_2, m_3, c) + O\left(T^{-A} \prod_{i=1}^3 (1 + |m_i|)^{-2}\right), \tag{8.8}$$

with the following parameters. If  $M_i N_i / C \gg T^\varepsilon$  for some  $i$ , then  $L$  is defined as (8.5) but with  $V = T$ ,  $U = X^{1/3} T^{2/3}$ , and  $X = M_1 M_2 M_3 / C$ . In addition,  $L$  vanishes unless

$$C \asymp \frac{(N_1 N_2 N_3)^{1/2}}{T}, \quad M_i \ll \frac{(N_1 N_2 N_3)^{1/2}}{N_i \Delta^{1-\varepsilon}}, \tag{8.9}$$

and  $M_1 N_1 \asymp M_2 N_2 \asymp M_3 N_3$ . If  $M_i N_i / C \ll T^\varepsilon$  for all  $i$  then  $K$  has the same form as (8.4) with  $L$  defined as (8.5) but with  $V = T^{-\varepsilon} X^{-1/2}$ ,  $U = T^\varepsilon$ , and  $X = M_1 M_2 M_3 / C$ .

**Remarks.** Here the  $y$ -integral is practically harmless, so in effect Lemma 8.1 expresses  $K$  in terms of an integral of length  $U$ , and has all variables  $m_1, m_2, m_3, c$  separated. The expressions (8.4) and (8.8) are identical except for the sign in  $e_c(\mp m_1 m_2 m_3)$ ; in both cases, this phase cancels a factor coming from the calculation of  $G_{\pm}(m_1, m_2, m_3; c)$  defined by (8.2) below. Actually, the appearance of  $e_c(m_1 m_2 m_3)$  in (8.8) is artificial in the sense that this phase is of lower order than the phase implicit in  $L(m_1, m_2, m_3, c)$  and we could have equally well showed that  $e_c(m_1 m_2 m_3)L(m_1, m_2, m_3, c)$  has a representation in the form (8.5).

A careful reader may notice that we have not stated an expression for  $K$  that is a true analog of (3.4). The reason for this omission is that for the large sieve method, it is desirable to have the variables  $m_1, m_2, m_3$ , and  $c$  separated in multiplicative (Mellin) form, and so developing an asymptotic expansion of the new phase of  $L$  is at odds with this goal. It should be possible to develop the correct form of (3.4) by applying the stationary phase method to (8.16). Instead, the key idea is that in (8.15), the variables  $m_1, m_2, m_3, c$  only occur in the block form  $m_1 m_2 m_3 / c$ . It then suffices to understand the Mellin transform of  $\Phi$  in a somewhat crude form (upper bounds suffice), which is a bit easier than forming an asymptotic evaluation with many lower-order terms.

*Proof of Lemma 8.1.* As our first step, we integrate by parts three times in each of the  $t_i$  for which  $m_i \neq 0$ , allowing us to obtain a crude bound of the form

$$K(m_1, m_2, m_3, c) \ll P(T, \Delta, N_1, N_2, N_3, c) \prod_{i=1}^3 (1 + |m_i|)^{-3},$$

where  $P$  is some fixed polyomial. This bound is sufficient for Lemma 8.1 when some  $m_i$  is  $\gg T^{A'}$  for some large  $A'$  depending polynomially on  $A$ .

For the rest of the proof, suppose that  $|m_i| \ll T^{A'}$  for some  $A'$ , and each  $i$ . Then in the expansion for  $B(x)$  we may assume  $x \gg T$ , since otherwise  $B_0$  is extremely small, and we obtain the desired bound for Lemma 8.1.

Before we jump into more intricate analysis, we can fix the sizes of  $C$  as stated in (8.7) and (8.9) using  $x \asymp (N_1 N_2 N_3)^{1/2} / C$ , and Lemmas 7.1 and 7.2.

It is a bit awkward to directly treat  $K$  by stationary phase in each  $t_i$ , so instead we use the following workaround. By the change of variables  $t_3 = u / (t_1 t_2)$ , we obtain

$$K(m_1, m_2, m_3, c) = \int_0^\infty B_0\left(\frac{4\pi\sqrt{u}}{c}\right) I(u) du + O(T^{-A}),$$

where

$$I(u) = \int_0^\infty \int_0^\infty w_3\left(\frac{u}{t_1 t_2}\right) w_1(t_1) w_2(t_2) e_c\left(-\frac{m_3 u}{t_1 t_2} - m_1 t_1 - m_2 t_2\right) \frac{dt_1 dt_2}{t_1 t_2}.$$

The asymptotic expansion of  $I(u)$  will be derived from Lemma 6.5. This saves the more difficult case of stationary phase with  $B_0$  for last when the integral is one-dimensional. This feature pleasantly allows us to unify cases for as long as possible since we do not use any properties of  $B_0$  until the later stages.

Suppose that

$$M_i N_i / C \gg T^\varepsilon, \quad i = 1, 2, 3, \tag{8.10}$$

so that the integrand defining  $I(u)$  is oscillatory. The opposite case is somewhat easier and we will return to it later. Lemma 6.5 then shows that

$$I(u) = \frac{c}{(m_1 m_2 N_1 N_2)^{1/2}} e\left(-\frac{3(m_1 m_2 m_3 u)^{1/3}}{c}\right) w_4(u) + O(T^{-A}),$$

where  $w_4$  is a smooth function supported on  $u \asymp N_1 N_2 N_3$ , depending on  $m_1, m_2, m_3, c$ , etc., but that is inert in all variables. Furthermore, we derive from (6.5) that  $w_4$  vanishes unless

$$m_1 N_1 \asymp m_2 N_2 \asymp m_3 N_3. \tag{8.11}$$

Hence we derive

$$K = \left( \frac{c \Delta T}{(m_1 m_2 N_1 N_2)^{1/2}} \int_{|v| \leq \Delta^\varepsilon / \Delta} e\left(\frac{vT}{\pi}\right) g(\Delta v) \times \int_0^\infty e\left(\frac{2\sqrt{u}}{c} \phi(v) - \frac{3(m_1 m_2 m_3 u)^{1/3}}{c}\right) w_4(u) du dv \right) + O(T^{-A}). \tag{8.12}$$

The  $u$ -integral can also be analyzed by stationary phase, as (8.10) and Lemma 6.3(1) show that it is small unless a stationary point exists, which implies

$$|\phi(v)| \asymp |m_1 m_2 m_3|^{1/3} (N_1 N_2 N_3)^{-1/6} \tag{8.13}$$

(additionally,  $\phi(v)$  must have the same sign as  $m_1 m_2 m_3$ ). Note that if  $\phi(v) = \pm \cos(v)$  or  $\pm \cosh(v)$  then  $|\phi(v)| \asymp 1$ , so this implies  $|m_1 m_2 m_3| \asymp (N_1 N_2 N_3)^{1/2}$ , which by (8.11) leads to the assumption  $M_i \asymp (N_1 N_2 N_3)^{1/2} / N_i$  in Lemma 8.1(1).

However, if  $\phi(v) = \pm \sin(v)$  or  $\pm \sinh(v)$ , then  $|v| \asymp |m_1 m_2 m_3|^{1/3} (N_1 N_2 N_3)^{-1/6}$ , and since  $|v| \leq \Delta^{-1+\varepsilon}$ , we conclude that  $|m_1 m_2 m_3| \ll (N_1 N_2 N_3)^{1/2} \Delta^{-3+\varepsilon}$ . Again using (8.11) leads to the upper bound on  $M_i$  in (8.9).

Assuming (8.13), the stationary point at  $u_0 = (m_1 m_2 m_3)^2 / \phi(v)^6$  potentially lies inside the support of  $w_4$ , so

$$\int_0^\infty e\left(\frac{2\sqrt{u}}{c} \phi(v) - \frac{3(m_1 m_2 m_3 u)^{1/3}}{c}\right) w_4(u) du = \frac{c^{1/2} (N_1 N_2 N_3)^{5/6}}{|m_1 m_2 m_3|^{1/6}} e\left(-\frac{m_1 m_2 m_3}{c \phi(v)^2}\right) w_5(v) + O(T^{-A}), \tag{8.14}$$

where  $w_5$  is inert in terms of  $v$  as well as the  $m_i$  and  $c$ , and  $w_5$  has support on (8.13). The fact that  $w_5$  is inert in terms of  $v$  perhaps requires some discussion. We naturally obtain an inert function in terms of  $\phi(v)$ , but since  $\phi(v)$  has bounded derivatives for  $|v| \leq 1$ , we do in fact obtain an inert function of  $v$ .

By inserting (8.14) into (8.12), and using (8.11) to obtain a more symmetric expression, we derive

$$K(m_1, m_2, m_3, c) = \frac{c^{3/2} \Delta T (N_1 N_2 N_3)^{1/2}}{|m_1 m_2 m_3|^{1/2}} e\left(\mp \frac{m_1 m_2 m_3}{c}\right) \Phi\left(\frac{m_1 m_2 m_3}{c}\right) + O(T^{-A}), \tag{8.15}$$

where

$$\Phi(x) = \int_{|v| \leq \Delta^\epsilon / \Delta} e\left(\frac{vT}{\pi}\right) g(\Delta v) e(x(\pm 1 - \phi(v)^{-2})) w_6(v) dv, \tag{8.16}$$

and  $w_6$  is another inert function having the same properties as  $w_5$ . Here the  $\mp$  sign in (8.15) is  $-$  for the  $\cos(v)$  or  $\cosh(v)$  cases, and  $+$  for the  $\sin(v)$  or  $\sinh(v)$  cases, and the sign in (8.16) respects these.

Finally, we shall use the Mellin technique to analyze  $\Phi(x)$ .

**Lemma 8.2.** *Suppose that  $g$  is a function satisfying  $g^{(j)}(x) \ll_{j,A} (1 + |x|)^{-A}$ ,  $\phi(v)$  is given via (7.8), and  $\Phi$  is defined by (8.16) for some inert function  $w_6$ . Then for  $x \asymp X$ ,*

$$\Phi(x) = \frac{1}{T} \int_{|t| \ll U} \lambda_{X,T}(t) x^{it} dt + O(T^{-A}),$$

where  $\lambda_{X,T}$  and  $U$  depend on  $X, T$ , and the choice of  $\phi(v)$ , and satisfy  $\lambda_{X,T}(t) \ll 1$  and

$$\begin{cases} U = T^2/X, & \phi(v) = \pm \cosh(v) \text{ or } \pm \cos(v), \\ U = X^{1/3} T^{2/3}, & \phi(v) = \pm \sinh(v) \text{ or } \pm \sin(v). \end{cases} \tag{8.17}$$

*Proof.* Suppose that  $1 \leq Y < y < 2Y$ , and let  $w(y)$  be a smooth function such that  $w(y) = 1$  on this interval, and  $w(y) = 0$  for  $y < Y/2$  and  $y > 3Y$ . By the Mellin inversion formula, for  $Y < y < 2Y$ , we have

$$e(y) = w(y)e(y) = \int_{-\infty}^{\infty} f(t)y^{it} dt, \quad f(t) = \frac{1}{2\pi} \int_0^{\infty} w(y)e(y)y^{-it} \frac{dy}{y}.$$

Integration by parts shows  $f(t) \ll (|t| + Y)^{-A}$  unless  $t \asymp Y$ , in which case  $f(t) \ll Y^{-1/2}$ , by stationary phase. We sometimes write  $f = f_Y$  to help us remember the parameter associated to  $f$ .

We will use this in  $\Phi(x)$ , but we need to treat the two types of  $\phi$  a bit differently. If  $\phi(v) = \pm \sin(v)$  or  $\pm \sinh(v)$ , we have  $|\phi(v)| \asymp |v| = o(1)$ , so  $x|1 + \phi(v)^{-2}| \asymp X/|v|^2 \gg T^\epsilon$ , and thus letting  $|v_0| = |m_1 m_2 m_3|^{1/3} (N_1 N_2 N_3)^{-1/6}$  (recall  $v$  is supported for  $|v| \asymp |v_0|$ ), we have

$$\begin{aligned} \Phi(x) &= \int_{|t| \asymp X/|v_0|^2} f_{X/v_0^2}(t) x^{it} \left( \int_{|v| \leq \Delta^\epsilon / \Delta} e\left(\frac{vT}{\pi}\right) g(\Delta v) (\phi(v)^{-2} + 1)^{it} w_6(v) dv \right) dt \\ &\quad + O(T^{-A}). \end{aligned}$$

We need to bound the inner integral over  $v$ . By a Taylor expansion, we have

$$(\phi(v)^{-2} + 1)^{it} = e^{-2it \log v + it(d_2 v^2 + d_4 v^4 + \dots)}$$



for certain constants  $d_i$  (depending only on the choice of  $\sin(v)$  or  $\sinh(v)$ ), so the  $v$ -integral is small except near the stationary point  $|v| \asymp t/T$ . Note that  $|v_0| \asymp |t|/T \asymp X/(v_0^2 T)$ , so that  $|v_0| \asymp (x/T)^{1/3}$ . As a pleasant consistency check, note that  $x \asymp M_1 M_2 M_3 / C$ , and so  $(x/T)^{1/3} \asymp (M_1 M_2 M_3)^{1/3} (N_1 N_2 N_3)^{-1/6}$ , using (8.9). The length of the  $t$ -integral is therefore seen to be  $X/|v_0|^2 \asymp X^{1/3} T^{2/3}$ , consistent with the claimed size of  $U$  in the second line of (8.17). In the inner  $v$ -integral, the second derivative of the phase is of size  $t/v_0^2$ , so we have

$$\int_{|v| \leq \Delta^\varepsilon / \Delta} e\left(\frac{vT}{\pi}\right) g(\Delta v) (1 + \phi(v)^{-2})^{it} w_6(v) dv \ll \frac{|v_0|}{\sqrt{|t|}},$$

and therefore  $f_{X/v_0^2}(t) |v_0| / \sqrt{|t|} \ll |v_0|^2 / (\sqrt{X} \sqrt{|t|}) \asymp |v_0|^3 / X \asymp T^{-1}$ , leading to the bound on  $\lambda_{X,T}(t)$  in the second line of (8.17).

Next we treat  $\phi(v) = \pm \cos(v)$  or  $\pm \cosh(v)$ . Here we have, for certain constants  $d'_i$ ,

$$1 - \phi(v)^{-2} = \frac{1}{2} v^2 + d'_4 v^4 + d'_6 v^6 + \dots,$$

so in this case we may initially restrict the  $v$ -integral so that  $v \asymp T/x$  (prior to applying the  $t$ -integral formula for  $e(y)$ ), using Lemma 6.3 again. In this range,  $x(1 - \phi(v)^{-2}) \asymp T^2/X$ , so that

$$\begin{aligned} \Phi(x) &= \int_{|t| \asymp T^2/X} f_{T^2/X}(t) x^{it} \left( \int_{|v| \asymp T/X} e\left(\frac{vT}{\pi}\right) g(\Delta v) (1 - \phi(v)^{-2})^{it} w_6(v) dv \right) dt \\ &\quad + O(T^{-A}). \end{aligned}$$

Now we have

$$(1 - \phi(v)^{-2})^{it} = e^{2it \log v + it \log(1 + c_4 v^2 + c_6 v^4 + \dots)}.$$

The second derivative of the phase is of size  $t/v^2 \asymp X$ . Hence

$$\int_{|v| \leq \Delta^\varepsilon / \Delta} e\left(\frac{vT}{\pi}\right) g(\Delta v) (1 - \phi(v)^{-2})^{it} w_6(v) dv \ll X^{-1/2},$$

and  $f_{T^2/X}(t) X^{-1/2} \ll T^{-1}$ . □

Now we apply Lemma 8.2 to (8.15), where recall that we restrict to  $c \asymp C$ , and  $|m_i| \asymp M_i$ . Then we have  $X = C^{-1} M_1 M_2 M_3$ , and so in the  $\pm \cos(v)$  or  $\pm \cosh(v)$  cases, we have  $U = T^2/X = T^2 C / M_1 M_2 M_3 \asymp T^2 C / (N_1 N_2 N_3)^{1/2}$ , so we have shown (8.4). The  $\pm \sin(v)$  or  $\pm \sinh(v)$  cases are similar.

Finally, we consider the case where

$$M_i N_i / C \ll T^\varepsilon. \tag{8.18}$$

In the  $\pm \cos(v)$  or  $\pm \cosh(v)$  cases, we claim that  $K$  is very small. To see this, we go back to the original definition (8.3), combined with (7.7). We also recall that  $B(x)$  is small unless  $x \gg \Delta T^{1-\varepsilon}$ , which means that each  $t_i$ -integral is oscillatory. Therefore, Lemma 6.3 shows that  $K$  is small unless  $M_i \asymp (N_1 N_2 N_3)^{1/2} / N_i$ ; in other words, the

inequalities in (8.7) remain valid. But then

$$M_i N_i / C \asymp (N_1 N_2 N_3)^{1/2} / C \gg \Delta T^{1-\varepsilon},$$

which is not consistent with (8.18) (meaning  $K$  is small).

Next we study the  $\pm \sin(v)$  or  $\pm \sinh(v)$  cases. Again we use (8.3) and (7.7), and change variables  $t_i \rightarrow N_i t_i$ , to obtain

$$\begin{aligned}
 K &= \Delta T (N_1 N_2 N_3) \int_{\mathbb{R}^3} \left( \prod_{i=1}^3 w_i(N_i t_i) e_c(-m_i N_i t_i) \right) \\
 &\quad \times \int_{|v| \leq \Delta^\varepsilon / \Delta} e\left( \frac{2\sqrt{N_1 N_2 N_3}}{c} \sqrt{t_1 t_2 t_3} \phi(v) \right) e\left( \frac{vT}{\pi} \right) g(\Delta v) dv dt_1 dt_2 dt_3 + O(T^{-A}).
 \end{aligned}
 \tag{8.19}$$

The  $v$ -integral is small unless  $\sqrt{N_1 N_2 N_3} / C \asymp T$ . The phase of each  $t_i$ -integral is then of size  $T|v| + O(T^\varepsilon)$ , by using (8.18). Thus if  $|v| \gg T^{-1+\varepsilon}$ , then repeated integration by parts shows  $K$  is small, so in (8.19) we may shorten the  $v$ -integral by assuming  $|v| \ll T^{-1+\varepsilon}$ , without creating a new error term. Next we artificially multiply  $K$  by  $e_c(-m_1 m_2 m_3) e_c(m_1 m_2 m_3)$ , and use a Mellin transform to separate the variables in  $e_c(-m_1 m_2 m_3)$  (we keep the plus sign part as it is, since it is visible in (8.8)). The cost in doing this is an integral of length  $O(T^\varepsilon)$ , by (8.18). Thus in all we obtain an expression of the form

$$K = \Delta (N_1 N_2 N_3) e_c(m_1 m_2 m_3) L(m_1, m_2, m_3, c) + O(T^{-A}),$$

where  $L$  is of the form (8.5) with  $V = T^{-\varepsilon}$  and  $U = T^\varepsilon$ . To put this into the form of (8.4), we use

$$\Delta N_1 N_2 N_3 \asymp \Delta T (N_1 N_2 N_3)^{1/2} C = \frac{C^{3/2} \Delta T (N_1 N_2 N_3)^{1/2}}{(M_1 M_2 M_3)^{1/2}} \left( \frac{M_1 M_2 M_3}{C} \right)^{1/2},$$

which is now of the desired form by absorbing  $X^{-1/2}$  into the  $V$  appearing in (8.5).  $\square$

### 9. Arithmetic separation of variables and the large sieve

The material in this section is logically independent of Section 8.

According to [CI, Lemma 10.2], we have an evaluation of  $G$  as follows. Let  $c = qr$  with  $q$  squarefree, and suppose

$$(m_3, r) = 1 \quad \text{and} \quad (m_1 m_2, q, r) = 1. \tag{9.1}$$

Then

$$\begin{aligned}
 G(m_1, m_2, m_3; c) \\
 &= e_c(m_1 m_2 m_3) \frac{h \chi_{kl}(-1)}{r q^2 \phi(k)} R_k(m_1) R_k(m_2) R_k(m_3) H(\overline{r h k m_1 m_2 m_3}; l),
 \end{aligned}
 \tag{9.2}$$

where  $h = (r, q)$ ,  $k = (m_1 m_2 m_3, q)$ ,  $l = q/(hk)$ , and  $R_k(m) = S(0, m; k)$  is the Ramanujan sum. If the coprimality conditions (9.1) are not satisfied then  $G$  vanishes. Here

$$H(w; q) = \sum_{u, v \pmod{q}} \chi_q(uv(u+1)(v+1)) e_q((uv-1)w). \tag{9.3}$$

Also [CI, (11.7)] gives

$$H(w; q) = \sum_{q_1 q_2 = q} \mu(q_1) \chi_{q_1}(-1) H^*(\overline{q_1} w; q_2), \tag{9.4}$$

where

$$H^*(w; q) = \sum_{\substack{u, v \pmod{q} \\ (uv-1, q) = 1}} \chi_q(uv(u+1)(v+1)) e_q((uv-1)w),$$

and from [CI, (11.9)],

$$H^*(w; q) = \frac{1}{\phi(q)} \sum_{\psi \pmod{q}} \tau(\overline{\psi}) g(\chi, \psi) \psi(w). \tag{9.5}$$

**Lemma 9.1** (Hybrid large sieve, [Ga, Theorem 2]). *Suppose  $U \geq 1$ , and let  $a_n$  be a sequence of complex numbers. Then*

$$\int_{-U}^U \sum_{\psi \pmod{q}} \left| \sum_{n \leq N} a_n \psi(n) n^{iu} \right|^2 du \ll (qU + N) \sum_{n \leq N} |a_n|^2.$$

**Lemma 9.2.** *Suppose that  $q$  is odd and squarefree, and let  $\alpha_{m_1}, \beta_{m_2}, \gamma_{m_3}, \delta_r$  be any sequence of complex numbers. Suppose  $(b_1 b_2, q) = 1$ ,  $a$  is a nonzero real number, and  $U \geq 1$ . Then*

$$\begin{aligned} & \int_{|u| \leq U} \left| \sum_{\substack{m_1, m_2, m_3 \\ m_i \asymp M_i}} \sum_{r \ll R} \alpha_{m_1} \beta_{m_2} \gamma_{m_3} \delta_r H^*(\overline{b_1 r} b_2 m_1 m_2 m_3; q) \left( \frac{m_1 m_2 m_3}{ar} \right)^{iu} \right| du \\ & \ll q^{1/2+\varepsilon} (qU + M_1 M_2)^{1/2} (qU + M_3 R)^{1/2} \left( \sum_{m_1, m_2, m_3, r} |\alpha_{m_1} \beta_{m_2} \gamma_{m_3} \delta_r|^2 \right)^{1/2}. \end{aligned} \tag{9.6}$$

The implied constant depends only on  $\varepsilon$ . Furthermore, (9.6) holds with  $H^*$  replaced by  $H$ . In addition, if we restrict the sums over  $m_3$  and  $r$  so that  $(m_3, r) = 1$ , then the left hand side of (9.6) (with either  $H^*$  or  $H$ ) is

$$\ll q^{1/2+\varepsilon} (qU + M_1 M_2)^{1/2} (qU + M_3 R)^{1/2} \left( \sum_{d, m_1, m_2, m_3, r} d^{1+\varepsilon} |\alpha_{m_1} \beta_{m_2} \gamma_{dm_3} \delta_{dr}|^2 \right)^{1/2} \tag{9.7}$$

This is a variation on [CI, Lemma 11.1] where the main difference is that here we have an integral which detects orthogonality in the archimedean aspect.

*Proof of Lemma 9.2.* By (9.5), the left hand side of (9.6) equals

$$\int_{|u| \leq U} \left| \frac{1}{\phi(q)} \sum_{\psi \pmod{q}} \sum_{\substack{m_1, m_2, m_3 \\ m_i \asymp M_i}} \sum_{r \ll R} \alpha_{m_1} \beta_{m_2} \gamma_{m_3} \delta_r \tau(\bar{\psi}) g(\chi, \psi) \psi(\bar{b}_1 \bar{r} b_2 m_1 m_2 m_3) \right. \\ \left. \times \left( \frac{m_1 m_2 m_3}{ar} \right)^{iu} \right| du, \tag{9.8}$$

which we arrange into a bilinear form by grouping together  $m_1$  and  $m_2$  as well as  $m_3$  and  $r$ ; this shows that (9.8) is bounded by

$$\int_{|u| \leq U} \sum_{\psi \pmod{q}} \frac{|g(\chi, \psi) \tau(\bar{\psi})|}{\phi(q)} \left| \sum_{m_1, m_2} \alpha_{m_1} \beta_{m_2} \psi(m_1 m_2) (m_1 m_2)^{iu} \right| \\ \times \left| \sum_{m_3, r} \gamma_{m_3} \delta_r \psi(\bar{r} m_3) \left( \frac{m_3}{r} \right)^{iu} \right| du. \tag{9.9}$$

Conrey and Iwaniec showed  $|g(\chi, \psi)| \ll q^{1+\varepsilon}$ , and of course  $|\tau(\bar{\psi})| \leq q^{1/2}$  (see e.g. [IK, Lemma 3.1]). By Cauchy’s inequality and some rearrangements, this is

$$\ll q^{1/2+\varepsilon} \left( \int_{|u| \leq U} \sum_{\psi \pmod{q}} \left| \sum_{n \ll M_1 M_2} a_n \psi(n) n^{iu} \right|^2 du \right)^{1/2} \\ \times \left( \int_{|u| \leq U} \sum_{\psi \pmod{q}} \left| \sum_{n \ll M_3 R} b_n \psi(n) n^{iu} \right|^2 du \right)^{1/2},$$

where  $(a_n)$  is the Dirichlet convolution of  $(\alpha_{m_1})$  and  $(\beta_{m_2})$ , and likewise for  $(b_n)$ . By Lemma 9.1, we obtain the bound (9.6), as desired.

The case with  $(m_3, r) = 1$  follows similar lines. In (9.9) we use Möbius inversion, writing  $\sum_{d|(m_3, r)} \mu(d)$  to detect this condition. Then we move the sum over  $d$  to the outside, and apply Cauchy’s inequality, in the form  $|\sum_d a_d|^2 \leq \zeta(1 + \varepsilon) \sum_d d^{1+\varepsilon} |a_d|^2$ . The remaining steps are identical to the previous case.

To replace  $H^*$  by  $H$ , we use (9.4) and apply (9.6) or (9.7), whichever is appropriate.  $\square$

**Lemma 9.3.** *Let conditions be as in Lemma 9.2. Then*

$$\int_{|u| \leq U} \left| \sum_{\substack{m_1, m_2, m_3 \\ m_i \asymp M_i}} \sum_{r \asymp R} \alpha_{m_1} \beta_{m_2} \gamma_{m_3} \delta_r G(m_1, m_2, m_3, qr) e_{-qr}(m_1 m_2 m_3) \left( \frac{m_1 m_2 m_3}{qr} \right)^{iu} \right| du \\ \ll \frac{q^{1/2+\varepsilon}}{Rq^2} (qU + M_1 M_2)^{1/2} (qU + M_3 R)^{1/2} \left( \sum_{d, m_1, m_2, m_3, r} d^{1+\varepsilon} |\alpha_{m_1} \beta_{m_2} \gamma_{d m_3} \delta_{dr}|^2 \right)^{1/2}. \tag{9.10}$$

Remark: If  $\gamma_{m_3} \ll 1$  and  $\delta_r \ll 1$ , which is all that we require in our use of Lemma 9.3, then the extra sum over  $d$  does not change the bound arising from  $d = 1$ .

Proof. By (9.2), the left hand side of (9.10) is of the form

$$\int_{|u| \leq U} \left| \sum_{\substack{m_1, m_2, m_3 \\ m_i \asymp M_i}}^* \sum_{r \asymp R} \alpha_{m_1} \beta_{m_2} \gamma_{m_3} \delta_r \frac{h \chi_{kl}(-1)}{r q^2 \phi(k)} \right. \\ \left. \times R_k(m_1) R_k(m_2) R_k(m_3) H(\overline{r h k} m_1 m_2 m_3; l) \left( \frac{m_1 m_2 m_3}{q r} \right)^{iu} \right| du,$$

where  $h, k, l$  are functions of the other variables, and the star on the sum indicates that (9.1) holds. Lemma 9.2 immediately shows that the most important terms  $h = k = 1, l = q$  give the stated bound, but of course we need to treat all cases. We may assume that the coefficients are zero unless  $m_i \asymp M_i$ , and  $r \asymp R$ . Next we write  $r = hr'$  where  $h | q$  and  $(r', q/h) = 1$ . Note that (9.1) means  $(m_1, h) = (m_2, h) = (m_3, h) = 1$  and  $(m_3, r') = 1$ . We also move the sum over  $k$  to the outside, showing that the left hand side of (9.10) is

$$\ll \sum_{hk|q} \frac{h}{R q^2 \phi(k)} \int_{|u| \leq U} \left| \sum_{\substack{m_1, m_2, m_3 \\ (m_1 m_2 m_3, q) = k}}^* \sum_{(r', m_3) = 1}^* \alpha_{m_1} \beta_{m_2} \gamma_{m_3} \delta_{hr'} \right. \\ \left. \times R_k(m_1) R_k(m_2) R_k(dm_3) H(\overline{r' h^2 k} m_1 m_2 m_3; l) \left( \frac{m_1 m_2 m_3}{r'} \right)^{iu} \right| du, \quad (9.11)$$

where the stars on the sums indicate the already listed coprimality conditions. The key point is that these conditions are only of the form  $(m'_i, s) = 1$  where  $s$  is an integer independent of the other  $m'_i$  and  $r'$  (and similarly for the condition on  $r'$ ). In this way, we can absorb these conditions into the definition of the coefficients without altering the upper bound on their magnitude.

The next problem is that the condition  $(m_1 m_2 m_3, q) = k$  is in terms of the product of the  $m_i$ , and so these variables are not separated. However, this is easily solved as follows. Since  $q$  is squarefree, for fixed  $k | q$ , the condition  $(m_1 m_2 m_3, q) = k$  means  $(m_1 m_2 m_3, k) = k$  and  $(m_1, q/k) = (m_2, q/k) = (m_3, q/k) = 1$ . We can parameterize the solutions to  $(m_1 m_2 m_3, k) = k$  as follows. Suppose  $(m_1, k) = k_1$ , and write  $m_1 = k_1 m'_1$  and  $k = k_1 k'$ , so  $(m_1, k') = 1$ . Now  $(m_2 m_3, k') = k'$ , so we can repeat this argument giving say  $(m_2, k') = k_2, m_2 = k_2 m'_2, k' = k_2 k_3$ , and therefore  $(m_3, k_3) = k_3$ , so we may write  $m_3 = k_3 m'_3$ . In this way we separate the  $m'_i$ .

Thus (9.11) is

$$\ll \sum_{\substack{hk|q \\ k=k_1 k_2 k_3}} \frac{h}{R q^2 \phi(k)} \int_{|u| \leq U} \left| \sum_{m'_1, m'_2, m'_3}^* \sum_{\substack{r' \asymp R/h \\ (r', m'_3) = 1}}^* \alpha_{k_1 m'_1} \beta_{k_2 m'_2} \gamma_{k_3 m'_3} \delta_{hr'} \right. \\ \left. \times R_k(m'_1 k_1) R_k(m'_2 k_2) R_k(m'_3 k_3) H(\overline{r' h^2 m'_1 m'_2 m'_3}; l) \left( \frac{m_1 m_2 m_3}{r'} \right)^{iu} \right| du.$$

Using  $|R_k(m)| \leq (m, k)$ , we easily see that

$$\sum_{m'_1 \ll M_1/k_1}^* |\alpha_{k_1 m'_1}|^2 |R_{k_1 k_2 k_3}(k_1 m'_1)|^2 \leq k_1^2 \sum_{m_1 \ll M_1} |\alpha_{m_1}|^2,$$

and similarly for  $m'_2$  and  $m'_3$ . Therefore, Lemma 9.2 shows that the left hand side of (9.10) is

$$\begin{aligned} &\ll \sum_{h|q} \sum_{k_1 k_2 k_3 = k|q} \frac{h}{Rq^2 \phi(k)} \left(\frac{q}{hk}\right)^{1/2+\varepsilon} \left(\frac{qU}{hk} + \frac{M_1 M_2}{k_1 k_2}\right)^{1/2} \left(\frac{qU}{hk} + \frac{M_3 R}{k_3 h}\right)^{1/2} \\ &\quad \times \left(k^2 \sum_{d, m_1, m_2, m_3, r} d^{1+\varepsilon} |\alpha_{m_1} \beta_{m_2} \gamma_{dm_3} \delta_{dr}|^2\right)^{1/2}. \end{aligned}$$

One easily checks that larger values of  $h$  and  $k$  do not create a larger error term than the case with  $h = k = 1$ , so the proof is complete.  $\square$

**10. Completion of the proof**

In this section we combine estimates from Sections 8 and 9, and prove Proposition 7.3, which in turn implies Theorem 1.1.

*Proof of Proposition 7.3.* Using (8.1), Lemma 8.1, and (9.2), we obtain

$$\begin{aligned} S_{\pm}(N_1, N_2, N_3; C; B_0) &\ll \sum_{M_1, M_2, M_3 \text{ dyadic}} \int_{|y| \leq T^\varepsilon} \frac{C^{1/2} \Delta T(N_1 N_2 N_3)^{1/2}}{(M_1 M_2 M_3)^{1/2} V} \\ &\quad \times \int_{|u| \ll U} \left| \sum_{m_i \asymp M_i} \sum_{r \asymp C/q} m_1^{iy_1} m_2^{iy_2} m_3^{iy_3} c^{iy_4} G(m_1, m_2, m_3, qr) e_{qr}(-m_1 m_2 m_3) \right. \\ &\quad \left. \times \left(\frac{m_1 m_2 m_3}{qr}\right)^{iu} du \right|, \end{aligned}$$

plus an error term of size  $O(T^{-A})$ . Recall we wish to show the bound (7.9). We need this for the three choices of  $U$  and  $V$  appearing in Lemma 8.1, so we do not specialize them yet. By Lemma 9.3 (and the following remark), we bound this by

$$\frac{C^{1/2} \Delta T(N_1 N_2 N_3)^{1/2}}{(M_1 M_2 M_3)^{1/2} V} \frac{q^{1/2+\varepsilon} T^\varepsilon}{Cq} (qU + M_1 M_2)^{1/2} \left(qU + \frac{M_3 C}{q}\right)^{1/2} \left(\frac{M_1 M_2 M_3 C}{q}\right)^{1/2}.$$

Without using any specializations yet, this simplifies as

$$\Delta T(N_1 N_2 N_3)^{1/2} (qT)^\varepsilon \left[ \frac{1}{qV} (qU + M_1 M_2)^{1/2} \left(qU + \frac{M_3 C}{q}\right)^{1/2} \right]. \tag{10.1}$$

As the desired bound  $\Delta T(N_1 N_2 N_3)^{1/2} (qT)^\varepsilon$  is already visible in front, we treat the part of (10.1) inside the square brackets.

First consider case (1) from Lemma 8.1 where  $M_i \asymp (N_1 N_2 N_3)^{1/2}/N_i$ ,  $V = T$ , and  $U$  and  $C$  are given by (8.6) and (8.7). Here we have  $qU \ll (qT/\Delta)T^\varepsilon$ ,  $M_3 C/q \ll \frac{N_1 N_2}{\Delta T q} T^\varepsilon \ll (qT/\Delta)T^\varepsilon$  (recalling  $N_i \ll (qT)^{1+\varepsilon}$ ), and  $M_1 M_2 \asymp N_3 \ll (qT)^{1+\varepsilon}$ . Thus (10.1) simplifies as

$$\Delta^{1/2} T (N_1 N_2 N_3)^{1/2} (qT)^\varepsilon.$$

Next consider case (2) from Lemma 8.1 with  $M_i N_i/C \gg T^\varepsilon$ . Here  $X = M_1 M_2 M_3/C$ ,  $V = T$ ,  $U = X^{1/3} T^{2/3}$ , (8.9) holds, and  $M_1 N_1 \asymp M_2 N_2 \asymp M_3 N_3$ . Using (8.9), we easily deduce  $X \ll T \Delta^{-3+\varepsilon}$ . This implies  $U \ll (T/\Delta)T^\varepsilon$ , exactly as in (1). Similarly,  $M_3 C/q \ll \frac{N_1 N_2}{q T \Delta} T^\varepsilon \ll (qT)^{1+\varepsilon}$ , as in (1), and  $M_1 M_2 \ll N_3 \Delta^{-2} T^\varepsilon$ , which is slightly smaller than in (1). It is now clear that (2) gives the same bound as (1).

Finally, we treat the case of  $M_i N_i/C \ll T^\varepsilon$  (which holds for all  $i$ , otherwise  $K$  is small); then the expression in brackets is, using  $X \ll T^\varepsilon$ ,

$$\ll \frac{T^\varepsilon}{q} (q + M_1 M_2)^{1/2} \left( q + \frac{M_3 C}{q} \right)^{1/2} \ll \frac{T^\varepsilon}{q} \left( q + \frac{C^2}{N_1 N_2} \right)^{1/2} \left( q + \frac{C^2}{N_3 q} \right)^{1/2}.$$

It still remains true that  $C \asymp (N_1 N_2 N_3)^{1/2}/T$ , so we quickly bound the expression in brackets by  $(qT)^\varepsilon$ , as desired.  $\square$

*Acknowledgements.* I would like to thank Ian Petrow and the referee for careful readings and extensive comments that improved the paper.

This material is based upon work supported by the National Science Foundation under agreement No. DMS-1101261. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

## References

- [BKY] Blomer, V., Khan, R., Young, M.: Distribution of mass of holomorphic cusp forms. *Duke Math. J.* **162**, 2609–2644 (2013) [Zbl 1312.11028](#) [MR 3127809](#)
- [Bl] Blomer, V.: Subconvexity for twisted  $L$ -functions on  $GL(3)$ . *Amer. J. Math.* **134** (2012), 1385–1421. [Zbl 1297.11046](#) [MR 2975240](#)
- [BH] Blomer, V., Harcos, G.: Hybrid bounds for twisted  $L$ -functions. *J. Reine Angew. Math.* **621**, 53–79 (2008) [Zbl 1193.11044](#) [MR 2431250](#)
- [Bu] Burgess, D. A.: On character sums and  $L$ -series, I. *Proc. London Math. Soc.* **12**, 193–206 (1962) [Zbl 0106.04004](#) [MR 0132733](#)
- [CI] Conrey, J. B., Iwaniec, H.: The cubic moment of central values of automorphic  $L$ -functions. *Ann. of Math. (2)* **151**, 1175–1216 (2000) [Zbl 0973.11056](#) [MR 1779567](#)
- [D] Duke, W.: Hyperbolic distribution problems and half-integral weight Maass forms. *Invent. Math.* **92**, 73–90 (1988) [Zbl 0628.10029](#) [MR 0931205](#)
- [Ga] Gallagher, P. X.: A large sieve density estimate near  $\sigma = 1$ . *Invent. Math.* **11**, 329–339 (1970) [Zbl 0219.10048](#) [MR 0279049](#)
- [GR] Gradshteyn, I. S., Ryzhik, I. M.: *Table of Integrals, Series, and Products*. 6th ed., Academic Press, San Diego, CA (2000) [Zbl 0981.65001](#) [MR 1773820](#)
- [Gu] Guo, J.: On the positivity of the central critical values of automorphic  $L$ -functions for  $GL(2)$ . *Duke Math. J.* **83**, 157–190 (1996) [Zbl 0861.11032](#) [MR 1388847](#)

- [HM] Harcos, G., Michel, P.: The subconvexity problem for Rankin–Selberg  $L$ -functions and equidistribution of Heegner points. II. *Invent. Math.* **163**, 581–655 (2006) [Zbl 1111.11027](#) [MR 2207235](#)
- [H-B1] Heath-Brown, D. R.: Hybrid bounds for Dirichlet  $L$ -functions. *Invent. Math.* **47**, 149–170 (1978) [Zbl 0362.10035](#) [MR 0485727](#)
- [H-B2] Heath-Brown, D. R.: Hybrid bounds for Dirichlet  $L$ -functions. II. *Quart. J. Math. Oxford Ser. (2)* **31**, 157–167 (1980) [Zbl 0396.10030](#) [MR 0576334](#)
- [HW] Huxley, M. N., Watt, N.: Hybrid bounds for Dirichlet’s  $L$ -function. *Math. Proc. Cambridge Philos. Soc.* **129**, 385–415 (2000) [Zbl 0970.11033](#) [MR 1780495](#)
- [Iv] Ivić, A.: On sums of Hecke series in short intervals. *J. Théor. Nombres Bordeaux* **13**, 453–468 (2001) [Zbl 0994.11020](#) [MR 1879668](#)
- [Iw1] Iwaniec, H.: Fourier coefficients of modular forms of half-integral weight. *Invent. Math.* **87**, 385–401 (1987) [Zbl 0606.10017](#) [MR 0870736](#)
- [Iw2] Iwaniec, H., Topics in Classical Automorphic Forms. *Grad. Stud. Math.* 17, Amer. Math. Soc. (1997) [Zbl 0905.11023](#) [MR 1474964](#)
- [IK] Iwaniec, H., Kowalski, E.: *Analytic Number Theory*. Amer. Math. Soc. Colloq. Publ. 53, Amer. Math. Soc. (2004) [Zbl 1059.11001](#) [MR 2061214](#)
- [JM] Jutila, M., Motohashi, Y.: Uniform bound for Hecke  $L$ -functions. *Acta Math.* **195**, 61–115 (2005) [Zbl 1098.11034](#) [MR 2233686](#)
- [KS] Katok, S., Sarnak, P.: Heegner points, cycles and Maass forms. *Israel J. Math.* **84**, 193–227 (1993) [Zbl 0787.11016](#) [MR 1244668](#)
- [KY] Kiral, E., Young, M.: The fifth moment of modular  $L$ -functions. [arXiv:1701.07507](#) (2007)
- [KZ] Kohlen, W., Zagier, D.: Values of  $L$ -series of modular forms at the center of critical strip. *Invent. Math.* **64**, 175–198 (1981) [Zbl 0468.10015](#) [MR 0629468](#)
- [Li] Li, X. Q.: Bounds for  $GL(3) \times GL(2)$   $L$ -functions and  $GL(3)$   $L$ -functions. *Ann. of Math.* **173**, 301–336 (2011) [Zbl 1320.11046](#) [MR 2753605](#)
- [LMY] Liu, S.-C., Masri, R., Young, M. P.: Subconvexity and equidistribution of Heegner points in the level aspect. *Compos. Math.* **149**, 1150–1174 (2013) [Zbl 1329.11046](#) [MR 3078642](#)
- [Lu] Lu, Q.: Bounds for the spectral mean value of central values of  $L$ -functions. *J. Number Theory* **132**, 1016–1037 (2012) [Zbl 1320.11047](#) [MR 2890524](#)
- [Pen] Peng, Z.: Zeros and central values of automorphic  $L$ -functions. Princeton PhD thesis (2001) [MR 2701928](#)
- [Pet] Petrow, I.: A twisted Motohashi formula and Weyl-subconvexity for  $L$ -functions of weight two cusp forms. *Math. Ann.* **363** 175–216 (2015) [Zbl 1331.11035](#) [MR 3394377](#)
- [W] Waldspurger, J.-L.: Sur les coefficients de Fourier des formes modulaires de poids demi-entier. *J. Math. Pures Appl. (9)* **60**, 375–484 (1981) [Zbl 0431.10015](#) [MR 0646366](#)
- [Y] Young, M.: The quantum unique ergodicity conjecture for thin sets. *Adv. Math.* **286**, 958–1016 (2016) [Zbl 1326.81281](#) [MR 3415701](#)
- [Z] Zhang, S. W.: Gross–Zagier formula for  $GL_2$ . *Asian J. Math.* **5**, 183–290 (2001) [Zbl 1111.11030](#) [MR 1868935](#)