J. Eur. Math. Soc. (Online first) DOI 10.4171/JEMS/1523

Alexander Dunn · Alexandru Zaharescu

The twisted second moment of modular half-integral weight L-functions

Received July 30, 2019; revised May 12, 2024

Abstract. Given a half-integral weight holomorphic Kohnen newform f on $\Gamma_0(4)$, we prove an asymptotic formula for large primes p with power saving error term for

$$\sum_{\chi \pmod p}^{*} |L(1/2, f, \chi)|^2.$$

Our result is unconditional, it does not rely on the Ramanujan–Petersson conjecture for the form f. This gives a very sharp Lindelöf-on-average result for Dirichlet series attached to Hecke eigenforms without an Euler product. The Lindelöf hypothesis for such series was originally conjectured by Hoffstein. There are two main inputs. The first is a careful spectral analysis of a highly unbalanced shifted convolution problem involving the Fourier coefficients of half-integral weight forms. The second input is a bound for sums of products of Salié sums in the Pólya–Vinogradov range. Half-integrality is fully exploited to establish such an estimate. We use the closed form evaluation of the Salié sum to relate our problem to the sequence $\alpha n^2 \pmod 1$. Our treatment of this sequence is inspired by work of Rudnick–Sarnak and the second author on the local spacings of αn^2 modulo 1.

Keywords: Kloosterman and Salié sums, half-integral weight automorphic forms, L-functions, twisted second moment, shifted convolution, local spacing statistics.

Contents

Alexander Dunn: School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA; adunn61@gatech.edu

Alexandru Zaharescu: Department of Mathematics, University of Illinois, Urbana, IL 61801, USA; Simon Stoilow Institute of Mathematics, Romanian Academy, 014700 Bucureşti, Romania; zaharesc@illinois.edu

Mathematics Subject Classification 2020: 11F66 (primary); 11L05, 11L07, 11F72, 11L15 (secondary).

1. Introduction and statement of results

Moments of L-functions play a central role in analytic number theory. Classical examples include the fourth moment of Riemann zeta

$$\int_0^T |\zeta(1/2+it)|^4\,dt = TP_4(\log T) + O_\varepsilon(T^{2/3+\varepsilon})$$

for a certain polynomial P_4 (see [25,49,68]), and the cuspidal analogue due to Good [20]

$$\int_0^T |L(1/2+it,f)|^2\,dt = TP_1(\log T) + O_\varepsilon(T^{2/3+\varepsilon})$$

for a certain polynomial P_1 depending on f.

The complexity of a moment computation for a family $\mathcal{F}$ of L-functions is measured by the quotient $r = \log\mathcal{C}/\log|\mathcal{F}|$, where $\mathcal{C}$ is the analytic conductor of each function in the family. The edge of current technology where one can hope to obtain an asymptotic with power saving error term is $r = 4$. Results in the case $r = 4$ can be found in a host of works, including Iwaniec–Sarnak [30], Kowalski–Michel–VanderKam [40] and Blomer [3].

From an adelic point of view, it is natural to replace the Archimedean twist $|\det|^{it}$ with a non-Archimedean twist by a Dirichlet character χ. Let $p > 2$ be prime, $\psi(p) := p-2$ denote the number of primitive characters modulo p, and

$$L(s,\chi) := \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s}, \qquad \operatorname{Re} s > 1,$$

be the usual Dirichlet L-function. In the breakthrough 2011 paper [65] Young proved for any $\varepsilon > 0$ that

$$\sideset{}{^*}\sum_{\chi \,(\mathrm{mod}\, p)} |L(1/2,\chi)|^4 = \psi(p)P_4(\log p) + O_\varepsilon(p^{1-\frac{1}{80}(1-2\theta)+\varepsilon}),$$

where $\psi(p) := p-2$ is the number of primitive Dirichlet characters modulo p, P_4 is a degree 4 polynomial and $\theta = 7/64$ is the best known exponent toward the Ramanujan–Petersson conjecture (due to Kim and Sarnak [32]) for Maass forms. The fourth moment of Dirichlet L-functions (for a general modulus $q \not\equiv 2 \pmod 4$) is a special case of the more general moment

$$\sideset{}{^*}\sum_{\chi \,(\mathrm{mod}\, q)} L(1/2, f\otimes\chi)\overline{L(1/2, g\otimes\chi)}, \tag{1.1}$$

where f, g are two fixed integral weight Hecke eigenforms (either holomorphic, Maass or Eisenstein) and could be either cuspidal or non-cuspidal. Here, $\{\lambda_f(n)\}_{n\geq 1}$ denotes the system of Hecke eigenvalues attached to f and

$$L(s, f \otimes \chi) := \sum_{n=1}^{\infty} \frac{\lambda_f(n)\chi(n)}{n^s}, \quad \operatorname{Re} s > 1. \tag{1.2}$$

Note that (1.2) has an Euler product when the weight of the form is integral. Striking progress has been made on the moment (1.1) in a sequence of works due to Blomer–Fouvry–Kowalski–Michel–Milićević–Sawin [5, 6, 9, 38]. An asymptotic for (1.1) (in the case $f = g$) with power saving error term appears in [6, Theorem 1.17]. The same family of twisted L-functions had also been previously studied in various contexts. One can see Chinta [11], Duke–Friedlander–Iwaniec [15], Gao–Khan–Ricotta [19], Stefanicki [62] and Hoffstein–Lee [46].

We would also like to highlight the recent 2022 breakthrough work of Li [45] that proves an asymptotic for the twisted second moment over the family of primitive quadratic Dirichlet characters (with fractional logarithmic power saving error term). This improved a result of Soundararajan and Young [61] that was conditional on the Generalised Riemann Hypothesis.

In this work we focus on the half-integral weight analogue of (1.1) when $f = g$. To enable subsequent discussion and introduce our results, we require some notation. More details are provided below in Section 4.1. For $j \in \mathbb{N}$, let $k := 1/2 + 2j$ be an odd half-integer. Suppose $f : \mathbb{H} \to \mathbb{C}$ is holomorphic, vanishes at all three cusps of $\Gamma_0(4)$, and satisfies

$$f(\gamma\tau) = \nu_\theta(\gamma)(c\tau + d)^k f(\tau) \quad \text{for all } \gamma \in \Gamma_0(4),$$

where ν_θ is the standard theta multiplier on $\Gamma_0(4)$. Let $\mathcal{S}_k(4)$ denote this space of cusp forms.

Let the Fourier expansion of f at ∞ be given by

$$f(\tau) := \sum_{n=1}^{\infty} b(n)e(n\tau) = \sum_{n=1}^{\infty} a(n)n^{\frac{k-1}{2}} e(n\tau). \tag{1.3}$$

For a prime $p > 2$ and a primitive character χ modulo p, define the twisted form

$$f_\chi(\tau) := \sum_{n=1}^{\infty} \chi(n)a(n)n^{\frac{k-1}{2}} e(n\tau),$$

of level $4p^2$. The twisted L-function is given by the Dirichlet series

$$L(s, f, \chi) := \sum_{n=1}^{\infty} \frac{a(n)\chi(n)}{n^s}, \quad \operatorname{Re} s > 1. \tag{1.4}$$

We have used a slightly different notation here to distinguish from the integral weight case discussed previously. Taking the Mellin transform of (1.4) one obtains a completed

L-function of degree 2 that has both a meromorphic continuation to all of $\mathbb{C}$ (in fact holomorphic, because f is cuspidal) and a functional equation, but is without an Euler product. The coefficients $a(n)$ are no longer multiplicative, except at squares.

For odd primes q, the Hecke operators $\mathcal{T}_{q^2}$ defined on $\mathcal{S}_k(4)$ (with $k = 1/2 + 2j$) are given by

$$\mathcal{T}_{q^2} f(\tau) := \sum_{n \geq 1} \left(b(q^2 n) + \left(\frac{n}{q}\right) q^{k-3/2} b(n) + q^{2k-2} b\left(\frac{n}{q^2}\right) \right) e(n\tau).$$

Here we have used the convention that $b(x) = 0$ unless $x \in \mathbb{Z}$. We call a half-integral weight cusp form a *Hecke cusp form* if $\mathcal{T}_{q^2} f = \lambda(q) f$ for all $q > 2$. One of the main tools for understanding half-integral weight forms and their coefficients is the Shimura lift [58]. Following Kohnen–Zagier [36], we focus on *Kohnen's plus subspace*. The behaviour of these forms under the Shimura lift is well understood. The Kohnen plus space $\mathcal{S}_k^+(4)$ (when $k = 1/2 + 2j$) is the subspace of $\mathcal{S}_k(4)$ consisting of forms whose Fourier coefficients satisfy

$$b(n) = 0 \quad \text{unless} \quad n \equiv 0, 1 \pmod 4. \tag{1.5}$$

This space has a basis consisting of simultaneous eigenfunctions of the $\mathcal{T}_{q^2}$ for odd q. As $k \to \infty$, asymptotically one-third of half-integral weight cusp forms lie in Kohnen's plus space by dimension considerations. Given a Hecke cusp form $f \in \mathcal{S}_k^+(4)$, one can normalise it so that its coefficients are totally real algebraic numbers [63]. Let d be a fundamental discriminant and

$$\psi_d(\bullet) := \left(\frac{d}{\bullet}\right). \tag{1.6}$$

There is no Euler product representation for (1.4), so one does not expect a Riemann hypothesis to hold. There are examples of Dirichlet series without an Euler product that fail to be subconvex at the centre point. Such an example is given in [12]. Let

$$D(s) := \sum_{n=1}^{\infty} \frac{\tau(n) \cos(2\pi n/q)}{n^s},$$

where $\tau(n)$ is the divisor function and q is a prime. This series has conductor q^2 and $D(1/2)$ gets as large as $\sqrt{q} \log q$ (convexity) as $q \to \infty$ through primes. This counterexample would suggest that the Euler product is crucial for subconvexity. However, in the case of automorphic L-functions attached to forms of integral weight, the Euler product is induced by the property that the attached form is a simultaneous eigenfunction for the Hecke operators. Jeffrey Hoffstein informally conjectured at Oberwolfach in 2011 that such a property was crucial in implying a Lindelöf hypothesis. Kıral [33] made the first progress towards a possible Lindelöf hypothesis. In particular, he proved that for primitive χ modulo p we have

$$L(1/2, f, \chi) \ll_{f,\varepsilon} p^{3/8+\theta/4+\varepsilon}, \tag{1.7}$$

where $\theta = 7/64$ is the Kim–Sarnak bound. Interestingly, any subconvex exponent $3/8 + \theta/4 < 1$ would be sufficient to obtain a power saving in our Theorem 1.1 below (cf. Remark 5.1 and the argument above it). Kıral's result also holds for more general moduli. For reference, the conductor here is $\asymp_k p^2$, so the exponent $3/8$ suggests a bound of Burgess quality. In this work we compute the “barrier” moment for this class of L-functions with power saving error term, that is, the moment that gives Lindelöf on average, but still yields the convexity bound for each individual L-value. Computing higher moments in this family certainly warrants further investigation to such an end. A mollified and/or amplified variant of the asymptotic second moment result in Theorem 1.1 (in addition to the first moment) would lead to a positive proportion of non-vanishing at centre point and subconvexity results. We leave this to the interested reader. One can also see [23] for applications of subconvex bounds in the level aspect for double Dirichlet series. Blomer [4] proved subconvex bounds of such series in the t-aspect on the critical line.

This family of L-functions attached to half-integral weight Kohnen newforms has also been studied in other contexts. In 2020, Lester and Radziwiłł under the Generalised Riemann Hypothesis proved that half-integral weight holomorphic Hecke forms in Kohnen’s plus space satisfy Quantum Unique Ergodicity (QUE) [44].

We use the spectral theory of automorphic forms and a delicate analysis of the distribution of αn^2 modulo 1 to prove the following moment result.

Theorem 1.1. *Let $\varepsilon > 0$, $j \in \mathbb{N}$, and f be a holomorphic cuspidal newform of weight $k := 1/2 + 2j$ on $\Gamma_0(4)$ such that*

- *f lies in Kohnen’s plus space,*
- *f is a simultaneous Hecke eigenform for all $\mathcal{T}_{q^2}$ with $q > 2$ prime,*
- *f is normalised so that its Fourier coefficients are totally real algebraic numbers.*

As $p \to \infty$ through primes $p \equiv 1 \pmod 4$, we have

$$\sum_{\chi \,(\mathrm{mod}\, p)}^{*} |L(1/2, f, \chi)|^2 = c_1(f)\psi(p)\log(p) + c_2(f)\psi(p) + O_{f,\varepsilon}(p^{1-\frac{1}{600}+\varepsilon}), \tag{1.8}$$

where $\psi(p) := p - 2$ is the number of primitive Dirichlet characters modulo p, and $c_1(f), c_2(f) \in \mathbb{R}$ are constants depending only on f. Furthermore, we have $c_1(f) > 0$.

Remark 1.1. The constants $c_1(f)$ and $c_2(f)$ are given in (5.15) and (5.16) respectively.

Remark 1.2. Other cases of Theorem 1.1, i.e. when $k = 3/2 + 2j$ and/or $p \equiv 3 \pmod 4$ can be established by a mild adaption of the methods in this paper. It is technically convenient to restrict attention to the case $k = 1/2 + 2j$ and $p \equiv 1 \pmod 4$.

Remark 1.3. We emphasise that the purpose of this paper was to break the moral “convexity barrier” by establishing a power saving error term in Theorem 1.1. Optimality of the power saving is not pursued in this paper.

Remark 1.4. There are other interesting potential variants of Theorem 1.1. The methods of the paper should be easily adapted to prove a moment with summand of the form $L(1/2, f, \chi)\overline{L(1/2, g, \chi)}$ with f, g both Kohnen newforms and orthogonal to one another. The main term should have magnitude $\psi(p)$ (with no $\log p$) in this case. Another variant is a moment with summand $|L(1/2, f, \chi)|^2$ where f is a non-cuspidal metaplectic Eisenstein series. This appears to be more involved because the Fourier coefficients of half-integer weight Eisenstein series are essentially quadratic Dirichlet L-functions. Unlike the case of Young [65], the convolution structure of the divisor function cannot be used in this case. Another interesting variant is a moment with summand $L(1/2, f, \chi)\overline{L(1/2, g, \chi)}$ where f is a Kohnen newform and g is its Shimura correspondent.

Theorem 1.1 depends on the following bound for a short sum of products of Salié sums. The result and its proof are of independent interest, because its origins are a bilinear form in Salié sums.

Theorem 1.2. *Let p be a prime with $p \equiv 1 \pmod 4$. Suppose $\varepsilon > 0$, $p^{1/2-1/10} \leq N \leq p^{1/2+1/10}$, $1 \leq M \leq p/2$ and $c \in \mathbb{F}_p^{\times}$. Then*

$$\sum_{N \leq n_1, n_2 \leq 2N} \Big| \sum_{M \leq m \leq 2M} S(m, cn_1, p)\overline{S(m, cn_2, p)} \Big| \\ \ll_{\varepsilon} p^{\varepsilon}\big(MN^2 p^{1-\frac{1}{27}} + MNp^{\frac{3}{2}-\frac{1}{27}} + N^2 p^{\frac{3}{2}-\frac{1}{27}} + Np^{2-\frac{1}{27}} + N^{\frac{1}{2}} p^{2+\frac{23}{108}} + p^{\frac{5}{2}-\frac{1}{27}}\big), \tag{1.9}$$

where $S(m, n, p)$ denotes the usual unnormalised Salié sum (cf. (4.24)), and the implied constant depends only on ε.

Theorem 1.2 gives a non-trivial power saving over the trivial bound in the Pólya–Vinogradov range $M, N \sim p^{1/2+o(1)}$ (cf. (2.3)). In a subsequent joint work, both authors with Kerr and Shparlinski [18] improved Theorem 1.2 using an alternative argument that exploited the geometry of numbers and additive combinatorics. One can also see [31] for further improvements, as well as a generalisation to higher order Salié sums. An arithmetic application of bilinear forms in Salié sums to the equidistribution modulo 1 of roots to the quadratic congruence $x^2 \equiv p \pmod q$ with q a large prime and p varying over primes $p \leq q$ is also given in [18]. Average versions (over q) of these applications are given in [60]. An average version (over the modulus q) of Theorem 1.1 with power saving error term was proved by the second author jointly with Shkredov and Shparlinski [59].

2. High level sketch

We work with normalised forms whose Fourier coefficients are totally real algebraic numbers to emulate the integral weight setting as much as possible. The natural starting point is an approximate functional equation for the product of L-functions

$$L(s, f, \chi)\overline{L(s, f, \chi)} = L(s, f, \chi)L(\overline{s}, f, \overline{\chi}).$$

After summing the approximate functional equation over all primitive χ modulo p using orthogonality and extracting the main terms, one obtains expressions roughly of the form

$$\text{(A)} \quad \frac{1}{p} \sum_{\substack{mn \leq p^2 \\ m \neq n}} a(m)\left(\frac{m}{p}\right)a(n)\left(\frac{n}{p}\right) \quad \text{and} \quad \text{(B)} \quad \sum_{\substack{mn \leq p^2 \\ m \equiv n \ (\mathrm{mod}\ p) \\ m \neq n}} a(m)a(n), \tag{2.1}$$

where $a(m)$ denotes the Fourier coefficients of the holomorphic half-integer weight cusp form f. The twisted terms in (A) appear because the theta multiplier causes the second term in the approximate functional equation for $L(s, f, \chi)L(\overline{s}, f, \overline{\chi})$ to contain Gauss sums attached to the character $\chi(\frac{\bullet}{p})$.

If one knew the Ramanujan–Petersson conjecture for the Fourier coefficients $a(n)$ of f (cf. (4.9)), then applying this bound pointwise to (2.1) would yield the "trivial" bound of $O(p^{1+\varepsilon})$. We will beat this bound by a power savings in p, without recourse to the Ramanujan–Petersson conjecture. The Ramanujan–Petersson conjecture for the Fourier coefficients of f is tantamount to the Lindelöf hypothesis for the L-function attached to quadratic twists of the Shimura correspondent of f by the Kohnen–Zagier formula [37] (cf. (4.7)).

We reserve the discussion here for (B). The terms in (A) can then be effectively handled using results implicit in the work of Kıral [33]. There are two well known ways to interpret this double summation. One point of view is to cast it as a shifted convolution problem involving Fourier coefficients of the half-integral weight form. This is useful when the sizes of the variables are not too far apart. Another option is to consider it a sum over the Fourier coefficients of a half-integral weight cusp form in arithmetic progressions. This has utility when one variable is significantly larger than the other.

To be precise, we restrict the variables to $n \asymp N$ and $m \asymp M$ where $N \geq M$ by symmetry and $NM = p^2$. On the one hand, we can apply Voronoi summation in the inner sum of

$$\sum_{m \asymp M} a(m) \sum_{\substack{n \asymp N \\ n \equiv m \,(\mathrm{mod}\, p)}} a(n),$$

obtaining an expression roughly of the form

$$\frac{N}{p^2} \sum_{m \asymp M} a(m) \sum_{n \asymp p^2/N} a(n) S(m, n, p), \tag{2.2}$$

where $S(m, n, p)$ denotes the usual unnormalised Salié sum (cf. (4.24)). Using Rankin–Selberg bounds and the evaluation of the Salié sum, we obtain a bound of $Mp^{1/2}$, which is admissible if $M \leq p^{1/2-\delta}$ (or equivalently $N \geq p^{3/2+\delta}$) for some fixed $\delta > 0$.

On the other hand, we can interpret the problem as an averaged shifted convolution sum

$$\sum_{r \asymp N/p} \sum_{\substack{n \asymp N, m \asymp M \\ n - m = rp}} a(m)a(n).$$

We detect the equality using additive characters and apply Jutila's circle method to set up the problem. One of the key steps is to perform Voronoi summation in both the m and n summations. The collision of the two theta multipliers (evaluated at opposite sign) in this process has the net effect of twisting by the quadratic character $\chi := (\frac{4\ell_1\ell_2}{\bullet})$, and essentially returns us to a weight zero setting. This feature can also be seen in another way. The function $V_{\ell_1,\ell_2}(\tau) := f(\ell_1\tau)\overline{f(\ell_2\tau)}y^k$ has nebentypus χ on $\Gamma_0(4\ell_1\ell_2)$. The standard approach to the shifted convolution problem $\ell_1 n - \ell_2 m = h$ would be to obtain the spectral decomposition of $\langle V_{\ell_1,\ell_2}, \mathcal{P}_h(\cdot,s)\rangle$, where $\mathcal{P}_h(\tau,s)$ is an appropriate Poincaré series (see [33]).

We get an expression roughly of the form

$$\frac{M^2(\ell_1\ell_2)}{C^3}\sum_{\substack{b\\|b|\asymp\mathcal{K}}}\sum_{\substack{\ell_1 n-\ell_2 m=b\\ \ell_1 n\asymp C^2N/M^2\\ \ell_2 m\asymp C^2/M}} a(m)a(n)\sum_{\ell_1\ell_2|c}\frac{K(b,h,c,\chi)}{c}\Phi\left(4\pi\frac{\sqrt{|b|h}}{c}\right),$$

where $C := N^{1000}$, Φ is some smooth weight function and the $K(b,h,c,\chi)$ are the usual weight zero Kloosterman sums twisted by a quadratic character. Note that the size of C has no bearing on the eventual bounds that are obtained. The Kuznetsov formula can then be applied to the summation over c to decompose it into the contributions from the holomorphic, Maass and Eisenstein spectrums. Here we are able to use the analysis of Blomer and Milićević [9]. A crucial input in this analysis is a flexible version of the large sieve for Maass forms due to Blomer and Milićević [9, Theorem 13] that allows for extra divisibility conditions. This idea leads to a bound roughly of the shape of $Np^{-1/2}$, rather than one of the quality $Np^{\theta-1/2}$, where θ is the best known exponent toward the Ramanujan–Petersson conjecture for weight zero Maass forms. A precise version of this bound is stated in Proposition 5.2. We also develop the analogous flexible large sieve bounds for coefficients of Eisenstein series attached to even Dirichlet characters, generalising the one for trivial nebentypus due to Blomer–Harcos–Michel in [7]. The computational technology for Eisenstein series developed by Kıral–Young [34] and Young [66] is useful for this.

Analogously to [9], it remains to close the small gap where $M = p^{1/2+o(1)}$ and $N = p^{3/2+o(1)}$, referred to as the *critical range*. We define the sets

$$\mathcal{N}_0(f) := \{n \in \mathbb{N} : 0 \leq |a(n)| \leq 1\},$$

and for all $r \geq 1$,

$$\mathcal{N}_r(f) := \{n \in \mathbb{N} : 2^{r-1} < |a(n)| \leq 2^r\}.$$

We break (2.2) into $O(\log^2 p)$ subsums

$$\frac{N}{p^2}\sum_{m\asymp M} a(m)\sum_{\substack{n\asymp p^2/N\\ n\in\mathcal{N}_r(f)}} a(n)S(m,n,p). \tag{2.3}$$

Observe that $M = p^{1/2+o(1)}$ and $p^2/N = p^{1/2+o(1)}$, and so we are left with estimating a bilinear form involving Salié sums in the Pólya–Vinogradov range. Power saving bounds for bilinear forms in Kloosterman sums and generalised Kloosterman sums in the Pólya–Vinogradov range have been stunningly proved by Kowalski–Michel–Sawin [38,39] using deep algebro-geometric techniques. We emphasise that the techniques of [38, 39] do not apply to the case of Salié sums (the monodromy group of the Salié sums is too small to make the arguments work). The elementary nature of Salié sums requires a completely different approach of diophantine nature. Applying Cauchy–Schwarz to the m sum in (2.3) and using Rankin–Selberg bounds we arrive at

$$\frac{NM^{1/2}}{p^2}\Bigg(\sum_{\substack{n_1,n_2\asymp p^2/N\\ n_1,n_2\in\mathcal{N}_r(f)}}|a(n_1)a(n_2)|\Bigg|\sum_{m\asymp M}S(m,n_1,p)\overline{S(m,n_2,p)}\Bigg|\Bigg)^{1/2}. \tag{2.4}$$

At this stage it is tempting to invoke the Ramanujan–Petersson conjecture to handle the Fourier coefficients using a sup norm in (2.4). Instead, we estimate (2.4) in two different ways depending on the size of r. For r large, we estimate trivially. The main feature here is that the Rankin–Selberg bound (4.4) implies that

$$|\mathcal{N}_r(f)\cap[0,X]|\ll_\varepsilon\frac{X^{1+\varepsilon}}{2^{2r}}.$$

This guarantees that (2.3) is

$$\ll\frac{Mp^{1/2}}{2^r},$$

which saves over the trivial bound as long as r is large enough.

For r small, we take the sup norm in (2.4), extend the summation on n by positivity, and use the closed form evaluation of Salié sums in terms of Weyl sums to obtain

$$\frac{2^rNM^{1/2}}{p^{3/2}}\Bigg(\sum_{n_1,n_2\asymp p^2/N}\Bigg|\sum_{M\le m\le 2M}\sum_{\substack{u,v\ (\mathrm{mod}\ p)\\ u^2\equiv mn_1\ (\mathrm{mod}\ p)\\ v^2\equiv mn_2\ (\mathrm{mod}\ p)}}e\Big(\frac{2(u+v)}{p}\Big)\Bigg|\Bigg)^{1/2}. \tag{2.5}$$

Recall that $M=p^{1/2+o(1)}$ and $p^2/N=p^{1/2+o(1)}$ here. The strategy now is to obtain cancellation in the short m summation by utilising the short average over n_1 and n_2. For simplicity, we consider a restricted version of the sum in (2.5) whose variables m,n_1 and n_2 satisfy

$$\Big(\frac{n_1}{p}\Big)=\Big(\frac{n_2}{p}\Big)=\Big(\frac{m}{p}\Big)=1.$$

The other case is analogous. For $\ell\in\mathbb{F}_p^\times$, define

$$A_\ell:=\sum_{M\le m\le 2M}\sum_{t^2\equiv m\ (\mathrm{mod}\ p)}e\Big(\frac{2t\ell}{p}\Big),$$

and

$$\mathbb{S}_\ell:=\{(u,v)\in(\mathbb{F}_p^\times)^2:(u^2,v^2)\ (\mathrm{mod}\ p)\in[p^2/N,p^2/N]\times[p^2/N,p^2/N]$$
$$\text{and } u+v\equiv\ell\ (\mathrm{mod}\ p)\}.$$

The triangle inequality asserts that the restricted version of the bracketed sum in (2.5) is

$$\le\sum_{\ell\ (\mathrm{mod}\ p)}|A_\ell|\,|\mathbb{S}_\ell|. \tag{2.6}$$

We focus on the non-trivial case when $\ell \in \mathbb{F}_p^\times$. The exponential sums A_ℓ are too short to complete, so we focus on $\mathbb{S}_\ell$, whose elements still capture the averaging over n_1 and n_2. Recall that $(u, v) \in \mathbb{S}_\ell$ are solutions to the linear equation

$$u + v \equiv \ell \ (\mathrm{mod}\ p), \tag{2.7}$$

whose squares lie in a short interval. Algebraically manipulating (2.7) we see that $(u, v) \in \mathbb{S}_\ell$ must satisfy the polynomial congruence

$$\overline{\ell}^2(u^2 - v^2)^2 + \ell^2 \equiv 2(u^2 + v^2) \ (\mathrm{mod}\ p). \tag{2.8}$$

We set

$$\alpha_\ell := \overline{\ell}^2/p \in \mathbb{Q}/\mathbb{Z} \quad \text{and} \quad \beta_\ell := \ell^2/p \in \mathbb{Q}/\mathbb{Z}.$$

Thus (2.8) implies

$$\|\alpha_\ell(u^2 - v^2)^2 + \beta_\ell\| \le 8p/N, \tag{2.9}$$

where $\| \bullet \|$ denotes the distance to the closest integer. Therefore pairs $(u, v) \in \mathbb{S}_\ell$ produce elements of the sequence $\{\alpha_\ell n^2\}_{0 \le n \le N}$ modulo 1 and lie in a cluster around $-\beta_\ell$. For given n and ℓ, there is at most one pair $(u, v) \in \mathbb{S}_\ell$ that corresponds to them.

The local spacing distribution of the sequence αn^2 for α irrational has been extensively studied in the literature. A classical result of Rudnick and Sarnak [52] states that for all integers $d \ge 2$ and almost all real α, the pair correlation of the sequence αn^d mod 1 is Poissonian. This is in contrast with the case $d = 1$, where it is well known that for all α and all N, the gaps between consecutive elements of αn mod 1, $1 \le n \le N$, can take at most three values.

Returning to the case $d = 2$, Rudnick, Sarnak, and the second author [53,67] show that for sufficiently well approximable numbers α, the m-level correlations and consecutive spacing are Poissonian along subsequences. For $\alpha = \sqrt{2}$, these types of conjectures are supported numerically [10] because of their close connection to the distribution between neighbouring levels of a generic integrable quantum system. It is also shown in [53] that when $b/p \in \mathbb{Q}$, the sequence bn^2/p also has a local Poissonian distribution when the number of points sampled is in certain ranges (in terms of p). In shorter ranges, they are able to show that such a phenomenon dramatically fails for some b. Moreover, some of the clusters of these points in these sequences are so dense that they are capable of making the 5-level (and all higher level) correlations diverge. One key aspect is that, although our α's are rational, our intuition comes from the case of α irrational in [53]. Thus, we will not work with the numbers α_ℓ themselves to analyse the cluster of points in (2.9), but instead consider various convergents to their respective continued fractions. We pay close attention to the size of the denominators of these convergents.

We fix $\boldsymbol{\delta} := (\delta_2, \delta_3, \delta_4, \delta_5) \in (0, 1)^4$ such that $\delta_2 < \delta_3 < \delta_4$ (in reality, we will need to have more parameters at our disposal). This vector is chosen appropriately in the course of the proof of Theorem 1.2. To control the size of $\mathbb{S}_\ell$, it is necessary to have control over the discrepancy of the sequence $\alpha_\ell n^2$. The natural strategy is to use the Erdős–Turán theorem in conjunction with Weyl's inequality for exponential sums whose argument is a quadratic polynomial. For this to work, the continued fraction expansion of $h\alpha_\ell$ for all $h \in [1, p^{\delta_5}]$ must have a convergent with denominator trapped in $[p^{\delta_2}, p^{\delta_3}]$ say. This leads us to essentially partition the summation variable in (2.6) into three sets,

$$\ell \in \mathbb{F}_p^\times := \mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3,$$

which are described below. The subset $\mathcal{H}_1$ contains exactly those ℓ described and so the sum over such ℓ in (2.6) can be handled.

The next subset of ℓ we consider are those such that there exists an $h_\ell \in [1, p^{\delta_5}]$ (it may depend on ℓ) such that $h_\ell \alpha_\ell$ has no convergent with denominator in the larger interval $[p^{\delta_2}, p^{\delta_4}]$. There is no toggle to control the size of $\mathbb{S}_\ell$ here, but $h_\ell \alpha_\ell$ has two consecutive convergents whose denominators have a large gap. This is a somewhat rare event, and an argument with standard inequalities from continued fractions indeed forces the size of $\mathcal{H}_2$ to be small.

This leaves the third and final set $\mathcal{H}_3$ to consider. For $\ell \in \mathcal{H}_3$, there exists an $h_\ell \in [1, p^{\delta_5}]$ such that $h_\ell \alpha_\ell$ has no convergent with denominator in $[p^{\delta_2}, p^{\delta_3}]$, but guaranteed to have one, say $a_\ell^\star / b_\ell^\star$, with $b_\ell^\star \in (p^{\delta_3}, p^{\delta_4}]$. Here we must study the sizes of $\mathcal{H}_3$ and $\mathbb{S}_\ell$ (which is equivalent to analysing $\alpha_\ell n^2$) simultaneously. We make this precise now. For each ℓ, denote

$$\mathbb{V}_\ell := \{0 \le n \le p^2/N : \|\alpha_\ell n^2 + \beta_\ell\| \le 8p/N\},$$

and for each $p^{\delta_3} \le U \le p^{\delta_4}$ and $0 \le V \le p^2/N$ we define

$$\mathcal{E}(U, V) := \{\ell \in \mathcal{H}_3 : b_\ell^\star \in [U, 2U] \text{ and } |\mathbb{V}_\ell| \in [V, 2V]\}.$$

Thus

$$\sum_{\ell \in \mathcal{H}_3} |A_\ell|\,|\mathbb{S}_\ell| \ll M \log^2 p \max_{p^{\delta_3} \le U \le p^{\delta_4}} \max_{1 \ll V \le p^2/N} V \cdot |\mathcal{E}(U, V)|. \tag{2.10}$$

We can assume V is moderately large by trivial considerations. Thus we need to bound $V \cdot |\mathcal{E}(U, V)|$. For each $\ell \in \mathcal{E}(U, V)$, we now construct an algebraic set $\mathfrak{C}_\ell \subseteq \mathbb{F}_p^3$ with restricted variables. Arrange the numbers $n_{\ell,j} \in \mathbb{V}_\ell$ with order

$$0 \le n_{\ell,1} < n_{\ell,2} < \cdots < n_{\ell,|\mathbb{V}_\ell|} \le p^2/N. \tag{2.11}$$

The average consecutive gap between these numbers is

$$\frac{p^2}{N|\mathbb{V}_\ell|} \asymp \frac{p^2}{NV}.$$

More than $|\mathbb{V}_\ell|/2$ consecutive gaps are less than or equal to $2p^2/(N|\mathbb{V}_\ell|)$. By the pigeonhole principle there exists an integer $1 \le d_\ell \le 2p^2/(N|\mathbb{V}_\ell|)$ that is repeated as a consecutive gap at least $|\mathbb{V}_\ell|^2 N/(4p^2)$ times. Thus we consider

$$\begin{aligned}\mathfrak{C}_\ell := \{(n, A, B) \in [1, p^2/N] \times [-8p^2/N, 8p^2/N]^2 : \overline{\ell}^2 n^2 + \ell^2 \equiv A \pmod{p} \\ \text{and } \overline{\ell}^2 (n + d_\ell)^2 + \ell^2 \equiv B \pmod{p}\},\end{aligned} \tag{2.12}$$

and

$$\mathfrak{U}(U,V) := \bigcup_{\ell \in \mathcal{E}(U,V)} \{\ell\} \times \mathfrak{C}_\ell \subseteq \mathbb{F}_p^4,$$

and perform an overall count of points in this last set in order to obtain a contradiction, unless the bound in the statement of Theorem 1.2 holds. On the one hand, we know that the size of this set should be large by construction. On the other hand, the nature of the set forces it to be thin enough. This eventually leads to the proof of Theorem 1.2.

The anatomy of the paper is as follows. Section 4 gives background material on holomorphic half-integer weight modular forms. Section 5 contains the main argument to establish Theorem 1.1. The main terms are extracted using Kronecker's first limit formula, and auxiliary bounds for sums of Fourier coefficients that are needed are also listed there. Propositions 5.1 and 5.2 are the main inputs used in the proof of Theorem 1.1. Proposition 5.2 uses spectral techniques. The relevant automorphic preliminaries are contained in Section 6, and Proposition 5.2 is proved in Section 7. Proposition 5.1 contains the critical range bounds and takes Theorem 1.2 as input. Theorem 1.2 is proved in Section 8.

3. Conventions

All implied constants in proofs are allowed to depend on $\varepsilon > 0$ (possibly different in each instance), $f \in \mathcal{S}_k(4)$, and the fixed smooth functions introduced in various partitions of unity. The square root $\sqrt{\ }$ denotes the principal branch of the square root.

4. Automorphic preliminaries I (half-integral weight)

4.1. Holomorphic cusp forms and L-functions

For $\tau := x + iy \in \mathbb{H}$, define $q := e^{2\pi i \tau}$. Let

$$\theta(\tau) := \sum_{n=-\infty}^{\infty} q^{n^2} \quad \text{and} \quad \eta(\tau) := q^{1/24} \prod_{n=1}^{\infty} (1 - q^n)$$

be the fundamental theta functions. Define the theta multiplier ν_θ on $\Gamma_0(4)$ by

$$\theta(\gamma\tau) = \nu_\theta(\gamma)\sqrt{c\tau + d}\,\theta(\tau) \quad \text{for } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(4),$$

and the eta multiplier ν_η on $\mathrm{SL}_2(\mathbb{Z})$ by

$$\eta(\gamma\tau) = \nu_\eta(\gamma)\sqrt{c\tau + d}\,\eta(\tau) \quad \text{for } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}).$$

The theta multiplier is given by the formula

$$\nu_\theta(\gamma) = \varepsilon_d^{-1}\left(\frac{c}{d}\right), \tag{4.1}$$

where $(\frac{\bullet}{\bullet})$ denotes the Kronecker symbol and

$$\varepsilon_d := \begin{cases} 1 & \text{if } d \equiv 1 \pmod 4, \\ i & \text{if } d \equiv 3 \pmod 4. \end{cases} \tag{4.2}$$

For $j \geq 1$ and $k := 1/2 + 2j$, let $\mathcal{S}_k(4)$ denote the space of holomorphic cusp forms of weight k, level 4 and trivial nebentypus. If $f \in \mathcal{S}_k(4)$, then $f : \mathbb{H} \to \mathbb{C}$ is holomorphic, vanishes on all three cusps of $\Gamma_0(4)$, and satisfies

$$f(\gamma\tau) = \nu_\theta(\gamma)(c\tau + d)^k f(\tau) \quad \text{for all } \gamma \in \Gamma_0(4).$$

For $f, g \in \mathcal{S}_k(4)$, define the *Petersson inner product*

$$\langle f, g\rangle := \int_{\Gamma_0(4)\backslash\mathbb{H}} y^k f(\tau)\overline{g(\tau)}\, d\mu(\tau), \quad d\mu(\tau) := \frac{dx\, dy}{y^2}.$$

Recall that $\mathcal{S}_k(4)$ becomes a Hilbert space with the inner product defined above.

Let the Fourier expansion of f at ∞ be given by

$$f(\tau) = \sum_{n=1}^{\infty} b(n)e(n\tau) = \sum_{n=1}^{\infty} a(n)n^{\frac{k-1}{2}} e(n\tau). \tag{4.3}$$

It follows from [17, p. 786] that for $X \geq 1$ we have

$$\sum_{n \leq X} |a(n)|^2 \ll_f X \log X. \tag{4.4}$$

A Wilton-type bound also follows from [17, p. 786],

$$\sum_{n \leq X} a(n)e(n\alpha) \ll_f X^{1/2} \log^2 X, \tag{4.5}$$

where the implied constant is uniform with respect to $\alpha \in \mathbb{R}$.

For odd primes q, the Hecke operators $\mathcal{T}_{q^2}$ defined on $\mathcal{S}_k(4)$ (with $k = 1/2 + 2j$) are defined by

$$\mathcal{T}_{q^2} f(\tau) := \sum_{n \geq 1} \left(b(q^2 n) + \left(\frac{n}{q}\right) q^{k-3/2} b(n) + q^{2k-2} b\left(\frac{n}{q^2}\right)\right) e(n\tau).$$

Here we have used the convention that $b(x) = 0$ unless $x \in \mathbb{Z}$. We call a half-integral weight cusp form a *Hecke cusp form* if $\mathcal{T}_{q^2} f = \lambda(q) f$ for all $q > 2$. The *Kohnen plus space* $\mathcal{S}_k^+(4)$ denotes the subspace of $\mathcal{S}_k(4)$ consisting of cusp forms f whose Fourier coefficients satisfy

$$b(n) = 0 \quad \text{unless} \quad n \equiv 0, 1 \pmod 4.$$

The plus space has a basis of simultaneous eigenfunctions of the $\mathcal{T}_{q^2}$ for q odd. For $f \in \mathcal{S}_k^+(4)$, there exists a Hecke cusp form $g \in \mathcal{S}_{2k-1}(1)$ (via the Shimura lift), where $\mathcal{S}_{2k-1}(1)$ denotes the space of integer weight cusp forms of weight $2k - 1$ and trivial nebentypus. This correspondence has the property that $H_q g = \lambda(q) g$, where H_q denotes the usual integer weight Hecke operator on $\mathcal{S}_{2k-1}(1)$. By the strong multiplicity 1 theorem, this determines f up to scalar multiplication. Write the Fourier expansion of g as

$$g(\tau) = \sum_{n=1}^{\infty} c(n) e(n\tau),$$

and normalise so that $c(1) = 1$. We can normalise f so that its coefficients lie in the field generated over $\mathbb{Q}$ by the coefficients $c(n)$ by [63, Proposition 2.3.1], and hence are totally real algebraic numbers. We have the coefficient relation

$$b(d\delta^2) = b(d) \sum_{e|\delta} \mu(e) e^{k-3/2} \psi_d(e) c\left(\frac{\delta}{e}\right), \tag{4.6}$$

where ψ_d was defined in (1.6). Recalling that both f and g have been normalised, the Kohnen–Zagier formula [37, Theorem 1] asserts

$$\frac{b(d)^2}{\frac{1}{6}\langle f, f\rangle} = \frac{(k-3/2)!}{\pi^{k-1/2}} d^{k-1} \frac{L(g \otimes \psi_d, 1/2)}{\langle g, g\rangle}. \tag{4.7}$$

Combining (4.6) and (4.7) we see that the Lindelöf hypothesis for all quadratic twists of the Shimura lift of f implies the Ramanujan–Petersson conjecture

$$b(n) \ll_{f,\varepsilon} n^{\frac{k-1}{2}+\varepsilon} \quad \text{for all } n \in \mathbb{N}. \tag{4.8}$$

Recalling the normalisation in (1.3), we obtain

$$a(n) \ll_{f,\varepsilon} n^{\varepsilon}. \tag{4.9}$$

It is well known that (cf. [26, (1.1)])

$$a(n) \ll_{f,\varepsilon} n^{1/4+\varepsilon} \quad \text{for all } n \in \mathbb{N}. \tag{4.10}$$

There has been considerable progress toward (4.9). Iwaniec [26] proved that

$$a(n) \ll_{f,\varepsilon} n^{3/14+\varepsilon}, \tag{4.11}$$

for all squarefree n. Conrey and Iwaniec [13] improved (4.11) to a Weyl-type subconvex bound

$$a(n) \ll_{f,\varepsilon} n^{1/6+\varepsilon}, \tag{4.12}$$

for all squarefree n. They achieved this bound by estimating a moment involving the L-values $L(1/2, g \otimes \psi_n)^3$ summed over all primitive cusp forms g of level dividing n and fixed integral weight. If f is a Hecke cusp form then (4.12) can be extended to all $n \in \mathbb{N}$ via (4.6).

Define the operator on $\mathcal{S}_k(4)$ by

$$(W_4 f)(\tau) := (2i\tau)^{-k} f\left(\frac{-1}{4\tau}\right).$$

Note that W_4 is an involution. Since W_4 commutes with each $\mathcal{T}_{p^2}$, a Kohnen newform is also an eigenfunction of W_4 with eigenvalue $\varepsilon(f) = \pm 1$ by the strong multiplicity 1 theorem for the plus space.

Let $Q \in \mathbb{N}$ and χ be a Dirichlet character modulo Q of conductor $Q^\star$. Then if

$$f(\tau) = \sum_{n=1}^{\infty} b(n) q^n \in \mathcal{S}_k(4),$$

define the χ-twist by

$$f_\chi(\tau) := \sum_{n=1}^{\infty} b(n)\chi(n) q^n \in S_k(4Q^{\star 2}, \chi^2).$$

The L-function of the twist f_χ (recall the normalisation in (1.3)) is defined by

$$L(s, f, \chi) := \sum_{n=1}^{\infty} \frac{a(n)\chi(n)}{n^s}, \quad \operatorname{Re} s > 1.$$

After taking the Mellin transform as in [33, p. 694], the completed L-function of f_χ is given by

$$L^*(s, f, \chi) = 2^{-1}\pi^{-k/2-s}(4Q^2)^{s/2}\Gamma\left(\frac{s+\frac{k-1}{2}}{2}\right)\Gamma\left(\frac{s+\frac{k+1}{2}}{2}\right)L(s, f, \chi). \tag{4.13}$$

We now give some details regarding the functional equation of $L^*(s, f, \chi)$ in the case $(Q, 4) = 1$ following [33]. Observe that f_χ can be realised as an average over additive twists

$$f_\chi(\tau) = \frac{1}{\mathcal{G}_{\overline{\chi}}(1; Q)} \sum_{u \,(\mathrm{mod}\, Q)} \overline{\chi}(u) f\left(\tau + \frac{u}{Q}\right). \tag{4.14}$$

Here for $c, n \in \mathbb{N}$ with $Q \mid c$, we denote the Gauss sum

$$\mathcal{G}_\chi(n; c) := \sum_{d=1}^{c}{}^{*} \chi(d) e\left(\frac{nd}{c}\right), \tag{4.15}$$

where $*$ denotes that the summation is over all d modulo c such that $(d, c) = 1$. Using (4.14), we can rewrite (4.13) as

$$L^*(s, f, \chi) = \frac{1}{\mathcal{G}_{\overline{\chi}}(1; Q)} \sum_{u \,(\mathrm{mod}\, Q)} \overline{\chi}(u) L^*\left(s, f, \frac{u}{Q}\right), \tag{4.16}$$

where

$$L^*\left(s, f, \frac{u}{Q}\right) := (4Q^2)^{s/2} \int_0^\infty f\left(iy + \frac{u}{Q}\right) y^{s+\frac{k-1}{2}} \frac{dy}{y}.$$

Let

$$\widetilde{\psi}_Q(\bullet) := \left(\frac{\bullet}{Q}\right).$$

Suppose u and v are any integers satisfying $4uv \equiv -1 \pmod{Q}$. A computation using the matrix identity

$$\begin{pmatrix} 1 & u/Q \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1/Q \\ 4Q & 0 \end{pmatrix} = \begin{pmatrix} (4uv+1)/Q & u \\ 4v & Q \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 4 & 0 \end{pmatrix}\begin{pmatrix} 1 & v/Q \\ 0 & 1 \end{pmatrix},$$

and the fact that

$$W_4 f = \varepsilon(f) f,$$

gives the relation

$$L^*\left(s, f, \frac{u}{Q}\right) = \varepsilon(f)\varepsilon_Q^{-2k}\left(\frac{v}{Q}\right)L^*\left(1-s, f, \frac{v}{Q}\right). \tag{4.17}$$

Observe that (4.16) and (4.17) imply the functional equation

$$L^*(s, f, \chi) = \varepsilon^*(f, \chi)L^*(1-s, f, \chi, \chi\widetilde{\psi}_Q), \tag{4.18}$$

where $\varepsilon^*(f, \chi)$ is a quantity of absolute value 1,

$$\varepsilon^*(f, \chi) := \varepsilon(f)\varepsilon_Q^{-2k}\chi(-4),$$

and the L-function defined on the right of (4.18) is defined by

$$\begin{aligned} &L^*(s, f, \chi, \chi\widetilde{\psi}_Q) \\ &\quad := (4Q^2)^{s/2}(2\pi)^{-(s+\frac{k-1}{2})}\Gamma\left(s+\frac{k-1}{2}\right)\frac{1}{\mathcal{G}_{\overline{\chi}}(1;Q)}\sum_{n=1}^{\infty}\frac{a(n)\mathcal{G}_{\chi\widetilde{\psi}_Q}(n;Q)}{n^s}. \end{aligned} \tag{4.19}$$

The Fourier coefficients in (1.3) are normalised so as to make (4.18) symmetric about $s = 1/2$.

A computation following [28, Chapter 5] shows that (4.18) implies an approximate functional equation. Let $V : \mathbb{R}_{>0} \to \mathbb{R}$ be defined by

$$V(x) = \frac{1}{2\pi i}\int_{(3)}\frac{x^{-z}}{(2\pi)^z}\frac{\Gamma(z+k/2)}{\Gamma(k/2)}\frac{dz}{z}.$$

We have

$$\begin{aligned} L(1/2, f, \chi) = &\sum_{m=1}^{\infty}\frac{a(m)\chi(m)}{\sqrt{m}}V\left(\frac{m}{2Q}\right) \\ &+ \frac{\varepsilon^*(f, \chi)}{\mathcal{G}_{\overline{\chi}}(1;Q)}\sum_{m=1}^{\infty}\frac{a(m)\mathcal{G}_{\chi\widetilde{\psi}_Q}(m;Q)}{\sqrt{m}}V\left(\frac{m}{2Q}\right). \end{aligned} \tag{4.20}$$

4.2. *Voronoi summation*

Let $V : (0, \infty) \to \mathbb{C}$ be a smooth function with compact support. Define the Hankel transform

$$\mathring{V}(y) := 2\pi i^k \int_0^{\infty} V(x)J_{k-1}(4\pi\sqrt{xy})\, dx,$$

where J denotes the usual J-Bessel function. Note that $\mathring{V}$ depends on k but is not displayed in the notation. We now see that $\mathring{V}$ is a Schwartz function. By [8, Section 2.6] we have

$$\int_0^\infty V(x) J_{k-1}(4\pi\sqrt{xy})\,dx \\ = \left(-\frac{1}{2\pi\sqrt{y}}\right)^j \int_0^\infty \frac{\partial^j}{\partial x^j}\left(V(x)x^{-\frac{k-1}{2}}\right)x^{\frac{k-1+j}{2}} J_{k-1+j}(4\pi\sqrt{xy})\,dx, \qquad (4.21)$$

for any $j \in \mathbb{N}_0$. One can then differentiate repeatedly under the integral sign in (4.21) using [21, (8.471.2)].

The next lemma follows from [17, p. 792].

Lemma 4.1. *Let $c \in \mathbb{N}$ such that $4 \mid c$ and*

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(c).$$

Let $V : (0,\infty) \to \mathbb{C}$ be a smooth function with compact support. Suppose $k = 1/2 + 2j$ with $j \in \mathbb{N}$ and let $a(n)$ denote the normalised Fourier coefficients of $f \in \mathcal{S}_k(4)$ as in (1.3). Then for $X > 0$ we have

$$\sum_n a(n) e\left(\frac{an}{c}\right) V\left(\frac{n}{X}\right) = \frac{X}{c}\nu_\theta(\gamma) \sum_n a(n) e\left(-\frac{dn}{c}\right) \mathring{V}\left(\frac{n}{c^2/X}\right).$$

4.3. *Half-integral weight Kloosterman sums and Salié sums*

Let $\kappa, c, m, n \in \mathbb{N}$ such that $\kappa \equiv 1 \pmod 2$ and $4 \mid c$. Then for any Dirichlet character χ modulo c define

$$K_{\kappa,\chi}(m,n;c) := \sideset{}{^*}\sum_{d \,(\mathrm{mod}\, c)} \varepsilon_d^{-\kappa}\left(\frac{c}{d}\right)\chi(d) e\left(\frac{md + n\overline{d}}{c}\right). \qquad (4.22)$$

When $\chi = \mathbf{1}_c$ in (4.22) we suppress the subscript. For $q \in \mathbb{N}$ with $q \equiv 1 \pmod 2$ and any Dirichlet character Ψ modulo q define the twisted sums

$$S_\Psi(m,n;q) = \sideset{}{^*}\sum_{x \,(\mathrm{mod}\, q)} \left(\frac{x}{q}\right)\Psi(x) e\left(\frac{mx + n\overline{x}}{q}\right). \qquad (4.23)$$

When $\Psi = \mathbf{1}_q$ in (4.23), we recover the well known Salié sum, and suppress the subscript. These sums have a closed form evaluation, unlike the Kloosterman sums attached to the trivial multiplier. Sarnak [55, p. 90] asserts that this phenomenon is the finite analogue of the Bessel function being an elementary function when its order is an odd half-integer. For example,

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\sin x, \quad J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\cos x.$$

When $q = p$ a prime and $(mn, p) = 1$, [54] gives

$$S(m,n;p) = \begin{cases} \left(\frac{n}{p}\right)\varepsilon_p\sqrt{p} \displaystyle\sum_{\substack{x \,(\mathrm{mod}\, p)\\ x^2\equiv mn \,(\mathrm{mod}\, p)}} e\left(\frac{2x}{p}\right) & \text{if } \left(\frac{mn}{p}\right) = 1,\\ 0 & \text{if } \left(\frac{mn}{p}\right) = -1, \end{cases} \tag{4.24}$$

where ε_d is given by (4.2).

We have the following useful twisted multiplicativity lemma.

Lemma 4.2. *Suppose $c = qr$ with $r \equiv 0 \pmod 4$ and $(q, r) = 1$ any Dirichlet character Ψ is a Dirichlet character modulo c. Let Ψ_r and Ψ_q are Dirichlet characters modulo r and q respectively such that $\Psi = \Psi_r\Psi_q$. Then*

$$K_{\kappa,\Psi}(m,n;c) = K_{\kappa-q+1,\Psi_r}(m\overline{q}, n\overline{q}; r)S_{\Psi_q}(m\overline{r}, n\overline{r}; q).$$

4.4. Eisenstein series and Rankin–Selberg L-functions

We give a brief background on the Eisenstein series and Rankin–Selberg L-functions relevant to the main terms appearing Theorem 1.1 and in Section 5.1. One can consult Section 6 below and [27, Chapter 13] for more details. Let

$$E_\infty(\tau,s) := \sum_{\gamma\in\Gamma_0(4)_\infty\backslash\Gamma_0(4)} (\mathrm{Im}\,\gamma\tau)^s, \quad \mathrm{Re}\, s > 1, \tag{4.25}$$

denote the weight zero Eisenstein series of level 4 attached to the cusp ∞. This Eisenstein series has a meromorphic continuation to all of $\mathbb{C}$ with its only pole in the region $\mathrm{Re}\, s \geq 1/2$ being simple and at $s = 1$, with residue

$$\mathrm{Res}_{s=1}\, E_\infty(\tau,s) = \frac{1}{\mathrm{Vol}(\Gamma_0(4)\backslash\mathbb{H})} = \frac{1}{2\pi}. \tag{4.26}$$

For $f \in \mathcal{S}_k(4)$, consider the Rankin–Selberg L-function

$$L(s, f \times \overline{f}) = \sum_{n=1}^{\infty} \frac{|a(n)|^2}{n^s} \quad \text{for } \mathrm{Re}\, s > 1.$$

The analytic continuation of $L(s, f \times \overline{f})$ is afforded by the above Eisenstein series. In particular, [27, Proposition 13.1] asserts (after taking into account the normalisations in the first two displays in [27, p. 233]):

$$(4\pi)^{-s-(k-1)}\Gamma(s+k-1)L(s, f\times\overline{f}) = \int_{\Gamma_0(4)\backslash\mathbb{H}} y^k|f(\tau)|^2 E_\infty(\tau,s)\, d\mu(\tau), \quad \mathrm{Re}\, s > 1. \tag{4.27}$$

The function $L(s, f \times \overline{f})$ also satisfies a vector functional equation ($L(s, f \times \overline{f})$ is one of the vector entries) with scattering matrix $\Phi(s, \chi_0)$, where χ_0 denotes the principal character modulo 4 [27, Theorem 13.4]. By [27, p. 240] the relevant scattering matrix has finite order and hence $L(s, f \times \overline{f})$ has polynomial growth in fixed vertical strips of $\mathbb{C}$ by the Phragmén–Lindelöf principle. Consulting [27, (13.34)], there is a simple pole at $s = 1$ with

$$\operatorname{Res}_{s=1} L(s, f \times \overline{f}) = \frac{(4\pi)^k}{\Gamma(k)} \frac{1}{\operatorname{Vol}(\Gamma_0(4)\backslash\mathbb{H})} \int_{\Gamma_0(4)\backslash\mathbb{H}} y^k |f(\tau)|^2 \, d\mu(\tau).$$

We will need the pole and constant term in the Laurent expansion of $E_\infty(\tau, s)$. We use Möbius inversion and the bijection [48, Lemma 7.1.6 (1)]

$$\Gamma_0(4)_\infty \backslash \Gamma_0(4) \simeq \{(c, d) : c \equiv 0 \pmod 4,\ (c, d) = 1 \text{ and } d > 0\},$$

in (4.25) to obtain

$$E_\infty(\tau, s) := \frac{1}{2} \frac{1}{4^s} \frac{1}{L(2s, \chi_0)} \left(\sum_{c,d}{}' \frac{(4y)^s}{|c(4\tau)+d|^{2s}} - \frac{1}{2^s} \sum_{c,d}{}' \frac{(2y)^s}{|c(2\tau)+d|^{2s}} \right), \quad \operatorname{Re} s > 1, \tag{4.28}$$

where χ_0 denotes the principal character modulo 4 and $'$ denotes the exclusion of $(c, d) = (0, 0)$. Both of the series on the right of (4.28) have meromorphic continuation to all of $\mathbb{C}$ with only simple poles at $s = 1$ [43, p. 273]. A computation that uses the Taylor expansion of 2^{-s} at $s = 1$, and that applies Kronecker's first limit formula [43, p. 273] to the right side of (4.28) gives (for each $\tau \in \mathbb{H}$) the Laurent expansion

$$\begin{aligned} E_\infty(\tau, s) = \frac{1}{2} \frac{1}{4^s} \frac{1}{L(2s, \chi_0)} \Bigg(\frac{\pi/2}{s-1} &+ \pi\left(\gamma - \frac{5}{2}\log 2\right) \\ &+ 2\pi \log\left(y^{-1/4} \left| \frac{\eta(2\tau)}{\eta(4\tau)^2} \right| \right) + O_\tau(s-1) \Bigg) \end{aligned} \tag{4.29}$$

for $s \in \mathbb{C}$ such that $|s - 1| < 1$.

4.5. *Functional equation for the second moment*

Here we take $Q = p$ a prime with $p \equiv 1 \pmod 4$, χ a primitive character modulo p such that $\chi \neq \widetilde{\psi}_p$. Thus $\chi\widetilde{\psi}_p$ is primitive, so we may appeal to the properties of Gauss sums attached to primitive characters. Let $f \in \mathcal{S}_k^+(4)$ be a simultaneous Hecke eigenform, so it is automatically an eigenfunction of W_4. Also suppose f is normalised so that its coefficients are totally real algebraic numbers (cf. Section 4.1). Applying [1, Theorem 8.15] we write (4.20) as

$$\begin{aligned} L(1/2, f, \chi) = \sum_{m=1}^{\infty} &\frac{a(m)\chi(m)}{\sqrt{m}} V\left(\frac{m}{2p}\right) \\ &+ \frac{\varepsilon^*(f, \chi)}{\mathcal{G}_{\overline{\chi}}(1; p)} \sum_{m=1}^{\infty} \frac{a(m)\overline{\chi}\widetilde{\psi}_p(m)\mathcal{G}_{\chi\widetilde{\psi}_p}(1; p)}{\sqrt{m}} V\left(\frac{m}{2p}\right). \end{aligned} \tag{4.30}$$

A computation with (4.30) using the fact that the Fourier coefficients are real shows that

$$\overline{L(1/2, f, \chi)} = L(1/2, f, \overline{\chi}).$$

Define $W : \mathbb{R}_{>0} \to \mathbb{R}$ by

$$W(x) = \frac{1}{2\pi i} \int_{(3)} \frac{x^{-z}}{(2\pi)^{2z}} \frac{\Gamma(z+k/2)^2}{\Gamma(k/2)^2} \frac{dz}{z}.$$

A computation following [28, Chapter 5] shows that (4.18) implies a second approximate functional equation

$$L(1/2, f, \chi)L(1/2, f, \overline{\chi}) = \sum_{m,n} \frac{a(m)\overline{\chi}(m)a(n)\chi(n)}{\sqrt{mn}} W\left(\frac{mn}{4p^2}\right) + \sum_{m,n} \frac{a(m)\overline{\chi}(m)\widetilde{\psi}_p(m)a(n)\chi(n)\widetilde{\psi}_p(n)}{\sqrt{mn}} W\left(\frac{mn}{4p^2}\right). \tag{4.31}$$

Note that we have for all $A > 0$ and $j \geq 0$ we have

$$W^{(j)}(x) \ll_{A,j} (1+x)^{-A}. \tag{4.32}$$

5. The core argument

We have the following orthogonality lemma.

Lemma 5.1. *Let p be prime and $\widetilde{\psi}_p = (\frac{\bullet}{p})$. For $m, n \in \mathbb{N}$ with $(nm, p) = 1$, we have*

$$\sideset{}{^*}\sum_{\substack{\chi \,(\mathrm{mod}\, p)\\ \chi \neq \widetilde{\psi}_p}} \overline{\chi}(m)\chi(n) = \sum_{\substack{d \mid p\\ d \mid (m-n)}} \phi(d)\mu\left(\frac{p}{d}\right) - \widetilde{\psi}_p(m)\widetilde{\psi}_p(n). \tag{5.1}$$

Recalling (1.7) we have

$$|L(1/2, f, \widetilde{\psi}_p)|^2 \ll_\varepsilon p^{3/4+\theta/2+\varepsilon}. \tag{5.2}$$

Summing (4.31) over all primitive characters $\chi \neq \widetilde{\psi}_p$ and applying Lemma 5.1 we obtain

$$\sideset{}{^*}\sum_{\substack{\chi \,(\mathrm{mod}\, p)\\ \chi \neq \widetilde{\psi}_p}} L(1/2, f, \chi)L(1/2, f, \overline{\chi}) = \mathcal{D}_1 + \mathcal{D}_2 + \mathcal{E}, \tag{5.3}$$

where

$$\mathcal{D}_1 := \sum_{d \mid p} \phi(d)\mu\left(\frac{p}{d}\right) \sum_{\substack{m \equiv n \,(\mathrm{mod}\, d)\\ (mn,p)=1}} \frac{a(m)a(n)}{\sqrt{mn}} W\left(\frac{mn}{4p^2}\right), \tag{5.4}$$

$$\mathcal{D}_2 := \sum_{d \mid p} \phi(d)\mu\left(\frac{p}{d}\right) \sum_{\substack{m \equiv n \,(\mathrm{mod}\, d)\\ (mn,p)=1}} \frac{a(m)\widetilde{\psi}_p(m)a(n)\widetilde{\psi}_p(n)}{\sqrt{mn}} W\left(\frac{mn}{4p^2}\right), \tag{5.5}$$

$$\mathcal{E} := -\sum_{m,n} \frac{a(m)\widetilde{\psi}_p(m)a(n)\widetilde{\psi}_p(n)}{\sqrt{mn}} W\left(\frac{mn}{4p^2}\right) - \sum_{\substack{m,n\\ (mn,p)=1}} \frac{a(m)a(n)}{\sqrt{mn}} W\left(\frac{mn}{4p^2}\right). \tag{5.6}$$

Combining (5.2)–(5.6) we obtain

$$\sum_{\chi \,(\mathrm{mod}\, p)}^{*} L(1/2, f, \chi)L(1/2, f, \overline{\chi}) = \mathcal{D}_1 + \mathcal{D}_2 + \mathcal{E} + O(p^{3/4+\theta/2+\varepsilon}). \tag{5.7}$$

Remark 5.1. Note that any subconvex exponent $3/4 + \theta/2 < 1$ in (5.2) would be sufficient.

5.1. Main terms

The main terms come from the diagonal $m = n$ in $\mathcal{D}_1$ and $\mathcal{D}_2$. This gives

$$\mathcal{M}(f, p) := 2\psi(p) \sum_{\substack{n \\ (n,p)=1}} \frac{a(n)^2}{n} W\left(\frac{n^2}{4p^2}\right). \tag{5.8}$$

Using (4.10) and (4.32), we can write

$$\mathcal{M}(f, p) := 2\psi(p) \sum_{n} \frac{a(n)^2}{n} W\left(\frac{n^2}{4p^2}\right) + O(p^{1/2+\varepsilon}). \tag{5.9}$$

Using Mellin inversion we obtain

$$\mathcal{M}(f, p) := 2\frac{\psi(p)}{2\pi i} \int_{(3)} 4^s L(1 + 2s, f \times \overline{f}) p^{2s} \widehat{W}(s)\, ds + O(p^{1/2+\varepsilon}) \tag{5.10}$$

where

$$\widehat{W}(s) = \frac{1}{(2\pi)^{2s}} \frac{\Gamma(k/2 + s)^2}{\Gamma(k/2)^2} \frac{1}{s}.$$

We shift the contour in (5.10) to $\operatorname{Re} s = -1/4 + \varepsilon$ and pick up a double pole at $s = 0$. Recalling (4.27) and applying the residue theorem, we see that

$$\mathcal{M}(f, p) = 2\psi(p) \lim_{s \to 0} \frac{d}{ds} \left(\int_{\Gamma_0(4)\backslash \mathbb{H}} y^k |f(\tau)|^2 h(\tau, s)\, d\mu(\tau) \right) + O(p^{1/2+\varepsilon}), \tag{5.11}$$

where

$$h(\tau, s) := s^2 \frac{(4\pi)^{2s+k}}{\Gamma(2s + k)} \widehat{W}(s) 4^s p^{2s} E_\infty(\tau, 1 + 2s).$$

We interchange the derivative and the integral by [42, Lemma 1.1, p. 409] and [2, Theorem 2, p. 130]. We then interchange the limit and integral using uniform convergence. Thus, for each fixed $\tau = x + iy \in \mathbb{H}$, it suffices to compute

$$y^k |f(\tau)|^2 \lim_{s \to 0} \frac{d}{ds} h(\tau, s). \tag{5.12}$$

Recalling (4.29) we have

$$sE_\infty(\tau, 1 + 2s) = \frac{1}{2} \frac{1}{4^{1+2s}} \frac{1}{L(2 + 4s, \chi_0)} \left(\frac{\pi}{4} + \left(\pi \left(\gamma - \frac{5}{2} \log 2 \right) \right.\right.$$
$$\left.\left. + 2\pi \log\left(y^{-1/4} \left| \frac{\eta(2\tau)}{\eta(4\tau)^2} \right| \right) \right) s + O_\tau(s^2) \right). \tag{5.13}$$

Thus

$$h(\tau,s) = \frac{(4\pi)^k\Gamma(s+k/2)^2p^{2s}}{8\Gamma(k/2)^2\Gamma(2s+k)}\frac{1}{L(2+4s,\chi_0)} \\ \times\left(\frac{\pi}{4}+\left(\pi\left(\gamma-\frac{5}{2}\log 2\right)+2\pi\log\left(y^{-1/4}\left|\frac{\eta(2\tau)}{\eta(4\tau)^2}\right|\right)\right)s+O_\tau(s^2)\right). \tag{5.14}$$

Performing (5.12) using (5.14) and the product rule, and substituting the result into (5.11), we obtain the constants

$$c_1(f) := \frac{(4\pi)^k}{\pi\Gamma(k)}\int_{\Gamma_0(4)\backslash\mathbb{H}} y^k|f(\tau)|^2\,d\mu(\tau), \tag{5.15}$$

$$c_2(f) := \frac{(4\pi)^{k+1}}{\pi^2\Gamma(k)}\int_{\Gamma_0(4)\backslash\mathbb{H}} y^k|f(\tau)|^2\log\left(y^{-1/4}\left|\frac{\eta(2\tau)}{\eta(4\tau)^2}\right|\right)d\mu(\tau) \\ +\frac{(4\pi)^{k-1}}{\Gamma(k)}\left(-\frac{8\log 2}{3}+\frac{4\Gamma'(k/2)}{\Gamma(k/2)}-\frac{4\Gamma'(k)}{\Gamma(k)}-\frac{48\zeta'(2)}{\pi^2}+8\left(\gamma-\frac{5}{2}\log 2\right)\right) \\ \times\int_{\Gamma_0(4)\backslash\mathbb{H}} y^k|f(\tau)|^2\,d\mu(\tau), \tag{5.16}$$

in Theorem 1.1.

5.2. *Error terms from $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{E}$*

Let $V_{1,2}:(0,\infty)\to\mathbb{R}_{\geq 0}$ be smooth functions compactly supported on $[1,2]$ that satisfy

$$V_{1,2}^{(j)}(x) \ll_j (\log 5p)^{2j} \ll_{j,\varepsilon} p^{\varepsilon}. \tag{5.17}$$

In (5.4)–(5.6) we place a smooth partition of unity, perform Mellin inversion and truncate the resulting integrals as in [9, Section 5]. We localise the variables to $M\leq m\leq 2M$ and $N\leq n\leq 2N$ satisfying

$$N\geq M\geq 1 \quad \text{(by symmetry)} \quad \text{and} \quad 1\leq MN\leq p^{2+\varepsilon}. \tag{5.18}$$

We find it is sufficient to bound $O(\log^2 p)$ sums of the shape

$$\mathcal{S}_{N,M,p,d} := \frac{d}{(MN)^{1/2}}\sum_{\substack{m\equiv n \ (\mathrm{mod}\, d)\\ m\neq n\\ (mn,p)=1}} a(m)a(n)V_1\left(\frac{m}{M}\right)V_2\left(\frac{n}{N}\right), \tag{5.19}$$

and

$$\widetilde{\mathcal{S}}_{N,M,p,d} := \frac{d}{(MN)^{1/2}}\sum_{\substack{m\equiv n \ (\mathrm{mod}\, d)\\ m\neq n\\ (mn,p)=1}} a(m)\widetilde{\psi}_p(m)a(n)\widetilde{\psi}_p(n)V_1\left(\frac{m}{M}\right)V_2\left(\frac{n}{N}\right), \tag{5.20}$$

for each $d = 1$ or p. Using (4.9) we have the bound

$$\mathcal{S}_{N,M,p,d}, \widetilde{\mathcal{S}}_{N,M,p,d} \ll (MN)^{1/2+\varepsilon}. \tag{5.21}$$

We will not make use of (5.21). We have the weaker, but unconditional version in the next lemma.

Lemma 5.2. *For all $N \geq 20M$ we have*

$$\mathcal{S}_{N,M,p,d}, \widetilde{\mathcal{S}}_{N,M,p,d} \ll M^{1/2+\varepsilon} N^{3/4+\varepsilon}.$$

Proof. We have

$$\mathcal{S}_{N,M,p,d}, \widetilde{\mathcal{S}}_{N,M,p,d} \ll \frac{d}{(MN)^{1/2}} \sum_{\substack{M \leq m \leq 2M \\ (m,p)=1}} |a(m)| \sum_{\substack{n \equiv m \ (\mathrm{mod}\ d) \\ N \leq n \leq 2N \\ (n,p)=1}} |a(n)|.$$

Observe that if $d = p$ and $N \leq p/2$, then the congruence condition $m \equiv n \pmod{d}$ implies that $m = n$. However, this is not possible since $M \leq m \leq 2M$, $N \leq n \leq 2N$, and $N \geq 20M$. In all other cases we have $\#\{N \leq n \leq 2N : n \equiv 0 \pmod{d}\} \asymp N/d$, and so

$$\mathcal{S}_{N,M,p,d}, \widetilde{\mathcal{S}}_{N,M,p,d} \ll \frac{d}{(MN)^{1/2}} \frac{N^{5/4}}{d} \sum_{M \leq m \leq 2M} |a(m)|,$$

where the last inequality follows from positivity and (4.10). The lemma now follows by applying Cauchy–Schwarz and then (4.4). ∎

Remark 5.2. Instead of (4.10), we could have applied the Conrey–Iwaniec bound (4.12) in the proof of Lemma 5.2 to obtain an improved bound. Optimality is not the purpose of this paper and we prefer to show that our method works (i.e. obtaining a power saving error term in Theorem 1.1) with weaker inputs.

We now remove the greatest common divisor condition in (5.19). The removal of the gcd condition in (5.20) is automatic because of the presence of the character.

Lemma 5.3. *For all M, N satisfying (5.18) and $d = 1$ or p we have*

$$\mathcal{S}_{N,M,p,d} := \frac{d}{(MN)^{1/2}} \sum_{\substack{m \equiv n \ (\mathrm{mod}\ d) \\ m \neq n}} a(m)a(n)V_1\left(\frac{m}{M}\right)V_2\left(\frac{n}{N}\right) + O(p^{1/2+\varepsilon}). \tag{5.22}$$

Proof. We prove (5.22) by estimating the contribution from pairs in the set

$$\mathcal{B} := \{(m,n) : m \equiv 0 \pmod{p} \text{ or } n \equiv 0 \pmod{p} \text{ such that } m \neq n\}.$$

If $d = 1$ and $1 \leq M \leq (p-1)/2$, then there are no pairs with $m \equiv 0 \pmod{p}$. Pairs (m,n) with $n \equiv 0 \pmod{p}$ contribute $O(p^{1/2+\varepsilon})$ by (4.10), Cauchy–Schwarz (on the m sum) and (4.4).

If $1 \le M \le (p-1)/2$ and $d = p$ then there are no (m,n) such that $m \equiv n \equiv 0 \pmod{p}$.

Now suppose that $(p-1)/2 \le M \le p^{1+\varepsilon/2}$ and $d = 1$. The contribution from all (m,n) with $m \equiv 0 \pmod{p}$ and $n \not\equiv 0 \pmod{p}$ is $O(p^{1/2+\varepsilon})$ by a similar argument to the above. The contribution from all (m,n) with $n \equiv 0 \pmod{p}$ and $m \not\equiv 0 \pmod{p}$ is at most $O(p^{1/2+\varepsilon})$ by a similar argument. The contribution from (m,n) with $m \equiv n \equiv 0 \pmod{p}$ is negligible by (4.10).

If $(p-1)/2 \le M \le p^{1+\varepsilon/2}$ and $d = p$ then the contribution from (m,n) with $m \equiv n \equiv 0 \pmod{p}$ is $O(p^{1/2+\varepsilon})$ by (4.10). ■

Lemma 5.4. *For all M, N satisfying (5.18) we have*

$$\mathcal{S}_{N,M,p,1} \ll_\varepsilon p^{1/2+\varepsilon}.$$

Proof. We write (5.22) as

$$\begin{aligned}\mathcal{S}_{N,M,p,1} = {} & \frac{1}{(MN)^{1/2}}\left(\sum_m a(m)V_1\left(\frac{m}{M}\right)\right)\left(\sum_n a(n)V_2\left(\frac{n}{N}\right)\right)\\ & - \frac{1}{(MN)^{1/2}}\sum_m a(m)^2 V_1\left(\frac{m}{M}\right)V_2\left(\frac{m}{N}\right) + O(p^{1/2+\varepsilon}). \end{aligned} \tag{5.23}$$

Applying partial summation, (4.4) and (4.5) guarantee that all but the last term in (5.23) are $O(p^\varepsilon)$. ■

Lemma 5.5. *Let $f \in \mathcal{S}_k(4)$ have normalised coefficients $a(n)$. Let χ be a primitive character modulo p and $V : (0,\infty) \to \mathbb{R}_{\ge 0}$ be a smooth function with support contained in $[1,2]$. Then for $X \ge 1$ we have*

$$\sum_m a(m)\chi(m)V\left(\frac{m}{X}\right) \ll_\varepsilon X^{1/2}p^{3/8+\theta/4+\varepsilon} + X^{3/8}p^{1/2+\varepsilon},$$

where θ represents the best progress toward the Ramanujan–Petersson conjecture for weight zero Maass forms.

Proof. Let S, S_j for $j = 1,2,3$ be defined as in [33, Proposition 3]. Each S_j is an averaged shifted convolution sum, depending on p, X, f, ℓ_1 and ℓ_2. Let $L \ge 1$ be a parameter to be chosen later. Taking χ' to be χ in [33, (11)] yields

$$\frac{L^2}{(\log L)^2}\left|\sum_m a(m)\chi(m)V\left(\frac{m}{X}\right)\right|^2 \le \phi(p)\sum_{\substack{L\le \ell_1,\ell_2 \le 2L\\ \ell_j \text{ prime}}} \chi(\ell_1)\overline{\chi(\ell_2)}(S_1+S_2+S_3). \tag{5.24}$$

Note the square is missing in [33, (11)]. Invoking the bounds for the S_j given in [33, Proposition 4 and Theorem 16], the right side of (5.24) is bounded by

$$\ll p(XL + X^{1/2}L^3 + X^{1+\varepsilon}L^{3+\varepsilon}p^{\theta-1/2}).$$

Note that the statement of [33, Theorem 16] has a typographic error (the inequality is missing a Q^θ). However, the correct bound is stated at the end of proof (cf. [33, p. 713]) and that is the one used here. Thus

$$\left|\sum_m a(m)\chi(m)V\left(\frac{m}{X}\right)\right| \ll L^\varepsilon\left(\frac{p^{1/2}X^{1/2}}{L^{1/2}} + L^{1/2}(X^{1/4}p^{1/2} + X^{1/2}p^{1/4+\theta/2})\right).$$

Choosing

$$L := \frac{p^{1/2}X^{1/2}}{X^{1/4}p^{1/2} + X^{1/2}p^{1/4+\theta/2}} \geq 1$$

yields the result. ■

Lemma 5.6. *For M, N satisfying (5.18) we have*

$$\widetilde{S}_{N,M,p,1} \ll_\varepsilon p^\varepsilon((MN)^{1/4}p^{3/8+\theta/4} + N^{1/4}M^{3/16}p^{7/16+\theta/8} + (MN)^{3/16}p^{1/2}).$$

Proof. We write (5.20) as

$$\begin{aligned}\widetilde{S}_{N,M,p,1} = {} & \frac{1}{(MN)^{1/2}}\left(\sum_m a(m)\widetilde{\psi}_p(m)V_1\left(\frac{m}{M}\right)\right)\left(\sum_n a(n)\widetilde{\psi}_p(n)V_2\left(\frac{n}{N}\right)\right) \\ & - \frac{1}{(MN)^{1/2}}\sum_{\substack{m \\ (m,p)=1}} a(m)^2 V_1\left(\frac{m}{M}\right)V_2\left(\frac{m}{N}\right).\end{aligned} \tag{5.25}$$

Observe that (4.4) implies that

$$\frac{1}{(MN)^{1/2}}\sum_{\substack{m \\ (m,p)=1}} |a(m)|^2 V_1\left(\frac{m}{M}\right)V_2\left(\frac{m}{N}\right) \ll p^\varepsilon. \tag{5.26}$$

By the Cauchy–Schwarz inequality and (4.4) we have

$$\sum_y a(y)\widetilde{\psi}_p(y)V_1\left(\frac{y}{X}\right) \ll X^{1+\varepsilon}. \tag{5.27}$$

Using Lemma 5.5, (5.27) and the fact

$$\min(A+B, C) \leq \sqrt{AC} + \sqrt{BC} \quad \text{for } A, B, C > 0,$$

we obtain

$$\begin{aligned}\sum_y a(y)\psi_p(y)V_2\left(\frac{y}{X}\right) &\ll \min(X^{1/2}p^{3/8+\theta/4+\varepsilon} + X^{3/8}p^{1/2+\varepsilon}, X^{1+\varepsilon}) \\ &\ll p^\varepsilon(X^{3/4}p^{3/16+\theta/8} + X^{11/16}p^{1/4}).\end{aligned} \tag{5.28}$$

After applying the triangle inequality in (5.25), we insert (5.26) and (5.28) into (5.25) to obtain the result. ■

We are now left to treat

$$\widetilde{\mathcal{S}}_{N,M,p,p} = \mathcal{S}_{N,M,p,p}. \tag{5.29}$$

It will be convenient to consider $M \leq N \leq 20M$ and $N \geq 20M$ separately.

Lemma 5.7. *For all M, N satisfying (5.18) and $M \leq N \leq 20M$, we have*

$$\mathcal{S}_{N,M,p,p} \ll_\varepsilon p^{1/2+\theta+\varepsilon}. \tag{5.30}$$

Proof. Lemma 5.3 and the bounds for S_2 and S_3 in [33, p. 713] together imply (5.30)

$$\mathcal{S}_{N,M,p,p} \ll_\varepsilon p^{1/2+\theta+\varepsilon}. \qquad \blacksquare$$

Now consider $N \geq 20M$. The following lemma in conjunction with parts of Lemma 5.2 and Proposition 5.2 below will essentially serve as a weaker, but unconditional replacement for the "trivial bound" in (5.21).

Lemma 5.8. *Let $\delta_0, \delta_1 > 0$ and suppose M, N satisfy (5.18) as well as $M \ll p^{1/2-\delta_0}$ and $N \gg p^{1+\delta_1}$. Then*

$$\mathcal{S}_{N,M,p,p} \ll_{\varepsilon,\delta_0,\delta_1} p^\varepsilon \Big(p^{3/4} M^{1/2} + \frac{p^{3/2}}{N^{1/2}} \Big).$$

Lemma 5.9. *Let $\delta > 0$ and suppose that M, N satisfy (5.18) as well a $M/N < p^{-1-\delta}$. Then*

$$\mathcal{S}_{N,M,p,p} \ll_{\varepsilon,\delta} p^\varepsilon \Big(p^{3/2} \Big(\frac{M}{N}\Big)^{1/2} + p^{1/2} \Big).$$

Proposition 5.1. *Let M, N be as in (5.18) and also satisfy*

$$p^{3/2-\frac{9}{100}} \leq N \leq p^{3/2+\frac{9}{100}}. \tag{5.31}$$

Then

$$\mathcal{S}_{N,M,p,p} \ll_\varepsilon p^\varepsilon \Big(\frac{M^{1/2}}{N^{1/2}} p^{3/2-\frac{1}{108}} + \frac{M^{1/2}}{N^{1/4}} p^{1+\frac{25}{216}} + \frac{M^{1/4}}{N^{1/2}} p^{3/2+\frac{25}{216}} + \frac{M^{1/4}}{N^{1/4}} p^{1+\frac{13}{54}} + \frac{M^{1/4}}{N^{1/8}} p^{1+\frac{23}{432}} + M^{1/4} p^{1-\frac{29}{216}} + p^{1-\frac{1}{10}} \Big).$$

We defer the proof of Proposition 5.1 (assuming the truth of Theorem 1.2) to Section 5.3. The proof of Theorem 1.2 is given in Section 8.

The last estimate we require in order to obtain bounds for (5.29) uses spectral methods. For $\ell_1, \ell_2, h \in \mathbb{Z}_{\geq 1}$, define the shifted convolution sum

$$\mathcal{D}(\ell_1, \ell_2, h, N, M) := \sum_{\ell_1 n - \ell_2 m = h} a(m) a(n) V_1\Big(\frac{\ell_2 m}{M}\Big) V_2\Big(\frac{\ell_1 n}{N}\Big), \tag{5.32}$$

and for d a positive integer,

$$\mathcal{S}(\ell_1, \ell_2, d, N, M) := \sum_{r \geq 1} \mathcal{D}(\ell_1, \ell_2, rd, N, M). \tag{5.33}$$

Lemma 5.3 gives (for all $N \geq 20M$),

$$\mathcal{S}_{N,M,p,p} = \frac{p}{(MN)^{1/2}} \mathcal{S}(1,1,p,N,M) + O(p^{1/2+\varepsilon}). \tag{5.34}$$

Proposition 5.2. *Suppose M, N are as in (5.18) and $N \geq 20M$. Then*

$$\mathcal{S}(\ell_1,\ell_2,p,N,M) \ll_\varepsilon p^\varepsilon (\ell_1\ell_2,p)^{1/2}\Big(\frac{N}{p^{1/2}} + \frac{N^{5/4}M^{1/4}}{p} + \frac{N^{3/4}M^{1/4}}{p^{1/4}} + \frac{NM^{1/2}}{p^{3/4}}\Big).$$

Thus,

$$\mathcal{S}_{N,M,p,p} \ll_\varepsilon p^\varepsilon \Big(\frac{(Np)^{1/2}}{M^{1/2}} + \frac{N^{3/4}}{M^{1/4}} + \frac{p^{3/4}N^{1/4}}{M^{1/4}} + p^{1/4}N^{1/2}\Big) + p^{1/2+\varepsilon}. \tag{5.35}$$

We defer the proof of Proposition 5.2 to Section 7.

5.3. *Coefficients in residue classes*

Here we prove Lemmas 5.8 and 5.9, as well as Proposition 5.1. The starting point for these results is Voronoi summation in the long variable.

For $r, v \in \mathbb{N}$, let

$$\mathrm{c}_v(r) = \sum_{\substack{u=1\\(u,v)=1}}^{v} e\Big(\frac{ru}{v}\Big).$$

denote the usual Ramanujan sum. Ramanujan sums are multiplicative in the modulus variable,

$$\mathrm{c}_{st}(r) = \mathrm{c}_s(r)\mathrm{c}_t(r) \quad \text{for } (s,t) = 1. \tag{5.36}$$

Proof of Lemma 5.9. Applying Lemma 5.3, we write the right side of (5.22) as the sum of three subsums

$$\mathcal{S}^\ell_{N,M,p,p} = \frac{p}{(MN)^{1/2}} \sum_{\substack{m\equiv n \,(\mathrm{mod}\, p)\\ m-n\equiv \ell \,(\mathrm{mod}\, 4)}} a(m)a(n)V_1\Big(\frac{m}{M}\Big)V_2\Big(\frac{n}{N}\Big), \quad \ell\in\{0,-1,1\}, \tag{5.37}$$

incurring an error of $O(p^{1/2+\varepsilon})$. Kohnen's plus space condition in (1.5) explains why we need only consider $\ell \in \{0,-1,1\}$ in (5.37).

Here, the condition $m \neq n$ is moot. In order to apply Lemma 4.1 we will need the moduli occurring in additive characters to be divisible by 4. Hence, we will use orthogonality in the form

$$\frac{1}{2p}\mathrm{c}_4(r)(1+\mathrm{c}_p(r)) = \begin{cases} 0 & \text{if } 2 \nmid r \\ -1 & \text{if } 2 \mid r \text{ and } 4 \nmid r \\ 1 & \text{if } r \equiv 0 \ (\mathrm{mod}\ 4) \end{cases} \times \begin{cases} 0 & \text{if } r \not\equiv 0 \ (\mathrm{mod}\ p) \\ 1 & \text{if } r \equiv 0 \ (\mathrm{mod}\ p). \end{cases} \tag{5.38}$$

By (5.36), the left side (5.38) is

$$\frac{1}{2p}\mathrm{c}_4(r) + \frac{1}{2p}\mathrm{c}_{4p}(r). \tag{5.39}$$

Without loss of generality we now focus on the case $\ell = 0$ in (5.37) and use (5.39) to detect both congruence conditions in (5.37). The $\ell \in \{\pm 1\}$ cases follow from similar arguments (replace $\mathrm{c}_4(r)$ with $1 - \mathrm{c}_4(r)$ above).

We apply (5.39) to $\mathcal{S}^0_{N,M,p,p}$ to remove both congruences. Thus $\mathcal{S}^0_{N,M,p,p}$ is the sum of the following two expressions:

$$\frac{1}{2(MN)^{1/2}} \sideset{}{^*}\sum_{j \ (\mathrm{mod}\, 4p)} \Bigg(\sum_m a(m) e\Big(\frac{jm}{4p}\Big) V_1\Big(\frac{m}{M}\Big)\Bigg)\Bigg(\sum_n a(n) e\Big(-\frac{jn}{4p}\Big) V_2\Big(\frac{n}{N}\Big)\Bigg), \tag{5.40}$$

and

$$\frac{1}{2(MN)^{1/2}} \sideset{}{^*}\sum_{j \ (\mathrm{mod}\, 4)} \Bigg(\sum_m a(m) e\Big(\frac{jm}{4}\Big) V_1\Big(\frac{m}{M}\Big)\Bigg)\Bigg(\sum_n a(n) e\Big(-\frac{jn}{4}\Big) V_2\Big(\frac{n}{N}\Big)\Bigg). \tag{5.41}$$

Partial summation and (4.5) guarantees that (5.41) is negligible. Let u be any integer with $u \equiv -j \pmod{4p}$. Applying Lemma 4.1 to the n summation in (5.40) we obtain

$$\sum_n a(n) e\Big(-\frac{jn}{4p}\Big) V_2\Big(\frac{n}{N}\Big) = \frac{N}{4p}\nu_\theta(\gamma) \sum_n a(n) e\Big(\frac{n\bar{j}}{4p}\Big) \mathring{V}_2\Big(\frac{nN}{16p^2}\Big),$$

where

$$\gamma := \begin{pmatrix} u & * \\ 4p & * \end{pmatrix} \in \Gamma_0(4p).$$

Observe that

$$\nu_\theta(\gamma) = \varepsilon_{-\bar{j}}^{-1}\Big(\frac{4p}{-\bar{j}}\Big) = \varepsilon_{-j}^{-1}\Big(\frac{4p}{j}\Big).$$

Thus (5.40) becomes

$$\frac{N^{1/2}}{8pM^{1/2}} \sum_m a(m) V_1\Big(\frac{m}{M}\Big) \sum_n a(n) \mathring{V}_2\Big(\frac{nN}{16p^2}\Big) K_1(-m,-n;4p). \tag{5.42}$$

Given the rapid decay of the Hankel transform we can truncate the n summation $1 \le n \le p^{2+\varepsilon}/N$ in (5.42) up to negligible error. By hypothesis we have $N > Mp^{1+\delta}$, and thus any $m \asymp M$ and $1 \le n \le p^{2+\varepsilon}/N$ satisfies $(mn, p) = 1$. Then by Lemma 4.2 and (4.24) we have

$$|K_1(-m,-n;4p)| \le 4p^{1/2}. \tag{5.43}$$

Inserting (5.43) into (5.42), we see that (5.42) is

$$\ll \frac{N^{1/2}}{p^{1/2}M^{1/2}} \Bigg(\sum_m \Big|a(m) V_1\Big(\frac{m}{M}\Big)\Big|\Bigg)\Bigg(\sum_n \Big|a(n) \mathring{V}_2\Big(\frac{nN}{16p^2}\Big)\Big|\Bigg).$$

Both the m and n summations can be estimated trivially using Cauchy–Schwarz and (4.4).

■

Proof of Lemma 5.8. Repeat the proof of Lemma 5.9 to the display (5.42). Note this incurs an error of $O(p^{1/2+\varepsilon})$. As in the last proof, we give details of the argument when $\ell = 0$. The $\ell \in \{\pm 1\}$ cases follow from similar arguments.

Interchanging the m and n summation in (5.42) gives

$$\frac{N^{1/2}}{8pM^{1/2}} \sum_{n} a(n) \mathring{V}_2\left(\frac{nN}{16p^2}\right) \sum_{m} a(m) V_1\left(\frac{m}{M}\right) K_1(-m,-n;4p). \tag{5.44}$$

We apply Cauchy–Schwarz (now in the n variable) and Lemma 4.2, and we see that (5.44) is

$$\ll \frac{1}{M^{1/2}} \max_{e,f,g \ (\mathrm{mod}\ 4)} \Bigg(\sum_{\substack{M \le m_1, m_2 \le 2M \\ m_1 \equiv e \ (\mathrm{mod}\ 4) \\ m_2 \equiv f \ (\mathrm{mod}\ 4)}} |a(m_1)a(m_2)| \\ \times \Bigg| \sum_{\substack{1 \le n \ll p^{2+\varepsilon}/N \\ n \equiv g \ (\mathrm{mod}\ 4)}} S(m_1, \overline{16}n; p) \overline{S(m_2, \overline{16}n; p)} \Bigg| \Bigg)^{1/2} \tag{5.45}$$

Using (4.24) the summation over n becomes

$$p \sum_{\substack{1 \le n \ll p^{2+\varepsilon}/N \\ n \equiv g \ (\mathrm{mod}\ 4)}} \sum_{\substack{x,y \ (\mathrm{mod}\ p) \\ x^2 \equiv \overline{16} m_1 n \ (\mathrm{mod}\ p) \\ y^2 \equiv \overline{16} m_2 n \ (\mathrm{mod}\ p)}} e\left(\frac{2(x-y)}{p}\right). \tag{5.46}$$

When $m_1 = m_2$, we estimate (5.46) trivially by $p^{3+\varepsilon}/N$. Then using (4.4) yields a contribution of $p^{3/2+\varepsilon}/N^{1/2}$ to (5.45).

Since $m_1, m_2 \asymp M$ and $M \ll p^{1/2-\delta_0}$ by hypothesis, $m_1 \neq m_2$ implies $m_1 \not\equiv m_2$ (mod p). Thus we can write (5.46) in terms of complete sums

$$p \sum_{n=0}^{p-1} \sum_{\substack{x,y \ (\mathrm{mod}\ p) \\ x^2 \equiv \overline{16} m_1 n \ (\mathrm{mod}\ p) \\ y^2 \equiv \overline{16} m_2 n \ (\mathrm{mod}\ p)}} e\left(\frac{2(x-y)}{p}\right) \sum_{\substack{1 \le w \le p^{2+\varepsilon}/N \\ w \equiv g \ (\mathrm{mod}\ 4)}} \frac{1}{p} \sum_{t=0}^{p-1} e\left(\frac{t(n-w)}{p}\right).$$

After interchanging the t and w summation, and estimating the sum over w in the usual way, it suffices to estimate the maximum of

$$p \log p \max_{t \in \mathbb{Z} \cap [0,p-1]} \left| \sum_{n=0}^{p-1} \sum_{\substack{x,y \ (\mathrm{mod}\ p) \\ x^2 \equiv \overline{16} m_1 n \ (\mathrm{mod}\ p) \\ y^2 \equiv \overline{16} m_2 n \ (\mathrm{mod}\ p)}} e\left(\frac{2(x-y)}{p}\right) e\left(\frac{tn}{p}\right) \right|. \tag{5.47}$$

The $n = 0$ term contributes $O(p \log p)$. Let $2 \le a \le p-2$ (since $m_1 \not\equiv m_2$ (mod p)) be such that $a^2 \equiv m_1 \overline{m_2}$ (mod p). We have $(x\overline{y})^2 \equiv m_1 \overline{m_2}$ (mod p). This implies that if $y \equiv v$ (mod p) for some $1 \le v \le p-1$, then $x \equiv \pm av$ (mod p) and $n \equiv 16\overline{m_2}v^2 \equiv$

$16\overline{m_1}a^2v^2 \pmod p$. Conversely, given $1 \le v \le p-1$, let $y \equiv v \pmod p$ and $x \equiv \pm av \pmod p$. Then $(x\overline{y})^2 \equiv m_1\overline{m_2} \pmod p$. Moreover, $16\overline{m_2}y^2 \equiv 16\overline{m_1}x^2 \pmod p$. Thus (5.47) becomes (with $n = 0$ excluded)

$$p \log p \max_{t \in \mathbb{Z} \cap [0,p-1]} \left| \sum_{v=1}^{p-1} e\left(\frac{16t\overline{m_2}v^2 + (\pm a - 1)v}{p} \right) \right|.$$

If $t \neq 0$, then the above line is bounded above by $p^{3/2} \log p$ since it is a Gauss sum. If $t = 0$, then it is a Ramanujan sum and is $p \log p$ since $2 \le a \le p-2$. Applying Cauchy–Schwarz and (4.4) we obtain a contribution of $p^{3/4+\varepsilon} M^{1/2}$ to (5.45) from this case. ■

Now we prove Proposition 5.1 assuming the validity of Theorem 1.2, whose proof is deferred to Section 8.

Proof of Proposition 5.1 (*assuming Theorem* 1.2). Repeat the proof of Lemma 5.9 to the display (5.42). Note this incurs an error $O(p^{1/2+\varepsilon})$. As in the previous two proofs, we give details of the argument when $\ell = 0$. The $\ell \in \{\pm 1\}$ cases follow from similar arguments.

We may restrict the n summation in (5.42) to $n \ge p^{2/5}$ using the same argument following (5.42) with error $O(p^{9/10+\varepsilon})$. Define the sets

$$\begin{aligned} \mathcal{N}_0(f) &:= \{n \in \mathbb{N} : 0 \le |a(n)| \le 1\}, \\ \mathcal{N}_r(f) &:= \{n \in \mathbb{N} : 2^{r-1} < |a(n)| \le 2^r\} \quad \text{for all } r \ge 1. \end{aligned}$$

We decompose (5.42) into $O(\log^2 p)$ subsums

$$\frac{N^{1/2}}{8pM^{1/2}} \sum_m a(m) V_1\left(\frac{m}{M}\right) \sum_{\substack{n \asymp A \\ n \in \mathcal{N}_r(f)}} a(n) \mathring{V}_2\left(\frac{nN}{16p^2}\right) K_1(-m,-n;4p), \tag{5.48}$$

where

$$p^{2/5} \le A \le p^{2+\varepsilon}/N. \tag{5.49}$$

Observe that (4.4) implies that for any $X \ge 1$ we have

$$|\mathcal{N}_r(f) \cap [0,X]| \ll X^{1+\varepsilon}/2^{2r}. \tag{5.50}$$

Applying Cauchy–Schwarz to the m summation in (5.48) and then using (4.4) we see that (5.48) is

$$\ll \frac{N^{1/2}}{p^{1-\varepsilon}} \left(\sum_{M \le m \le 2M} \left| \sum_{\substack{n \asymp A \\ n \in \mathcal{N}_r(f)}} a(n) K_1(-m,-n;4p) \mathring{V}_2\left(\frac{nN}{16p^2}\right) \right|^2 \right)^{1/2}. \tag{5.51}$$

Expanding the square and interchanging the summations, the expression inside the square root in (5.51) becomes

$$\sum_{\substack{n_1,n_2 \asymp A \\ n_1,n_2 \in \mathcal{N}_r(f)}} a(n_1)\overline{a(n_2)} \mathring{V}_2\left(\frac{n_1N}{16p^2}\right) \overline{\mathring{V}_2\left(\frac{n_2N}{16p^2}\right)} \times \sum_{M \le m \le 2M} K_1(-m,-n_1;4p)\overline{K_1(-m,-n_2;4p)}. \tag{5.52}$$

We use the bound $|K_1(-m,-n,p)| \leq 4p^{1/2}$ (we have $(mn,p)=1$ in the relevant ranges), Cauchy–Schwarz, and (5.50) to estimate (5.52) by

$$\ll \frac{p^5}{N^2}\frac{M}{2^{2r}},$$

for any A satisfying (5.49). Inserting this into (5.51), we see that (5.48) is

$$\ll \frac{1}{2^r}p^{3/2}\left(\frac{M}{N}\right)^{1/2}. \tag{5.53}$$

We now estimate (5.52) non-trivially to obtain another upper bound for (5.48). Applying the triangle inequality to (5.52), we see that (5.52) is

$$\ll 2^{2r} \sum_{\substack{n_1,n_2 \asymp A \\ n_1,n_2 \in \mathcal{N}_r(f)}} \left| \sum_{M \leq m \leq 2M} K_1(-m,-n_1;4p)\overline{K_1(-m,-n_2;4p)} \right|. \tag{5.54}$$

By positivity we can extend the summation over all $n_1, n_2 \asymp A$ in (5.54). Applying Lemma 4.2, (4.24) and the fact that $p \equiv 1 \pmod 4$ we see that the summand in the m summation is

$$K_1(-m,-n_1;4)\overline{K_1(-m,-n_2;4)}S(m,\overline{16}n_1;p)\overline{S(m,\overline{16}n_2;p)}. \tag{5.55}$$

Note that $K_1(-m,-n_1;4)\overline{K_1(-m,-n_2;4)}$ is an absolute constant depending only on m, n_1, n_2 modulo 4. Thus we rewrite (5.52) so that each summation variable runs in a fixed congruence class modulo 4. Thus it suffices to bound

$$2^{2r} \sum_{\substack{n_1,n_2 \asymp A \\ n_1 \equiv e \ (\mathrm{mod}\ 4) \\ n_2 \equiv f \ (\mathrm{mod}\ 4)}} \left| \sum_{\substack{M \leq m \leq 2M \\ m \equiv g \ (\mathrm{mod}\ 4)}} S(m,\overline{16}n_1;p)\overline{S(m,\overline{16}n_2;p)} \right| \tag{5.56}$$

for e, f, g modulo 4. The hypothesis on N ensures that

$$p^{\frac{41}{100}} \leq p^2/N \leq p^{\frac{59}{100}}.$$

Recall (5.49). The bound in Theorem 1.2 applied to (5.56) is increasing in A, thus by Theorem 1.2 and Remark 8.2 we can bound (5.56) by

$$\ll 2^{2r}\left(\frac{M}{N^2}p^{\frac{134}{27}} + \frac{M}{N}p^{\frac{187}{54}} + \frac{p^{\frac{295}{54}}}{N^2} + \frac{p^{\frac{107}{27}}}{N} + \frac{p^{\frac{347}{108}}}{N^{1/2}} + p^{\frac{133}{54}}\right), \tag{5.57}$$

uniformly in A satisfying (5.49). Inserting (5.57) into (5.51), we see that (5.48) is

$$\ll 2^r p^{\varepsilon}\left(\frac{M^{1/2}}{N^{1/2}}p^{\frac{40}{27}} + M^{1/2}p^{\frac{79}{108}} + \frac{p^{\frac{187}{108}}}{N^{1/2}} + p^{\frac{53}{54}} + N^{1/4}p^{\frac{131}{216}} + N^{1/2}p^{\frac{25}{108}}\right). \tag{5.58}$$

Denoting the right side of (5.58) as $2^r X$, one can choose

$$2^r = \frac{p^{3/4}(\frac{M}{N})^{1/4}}{X^{1/2}}$$

to balance (5.53) and (5.58). Substituting this quantity back into (5.58) and noting the error $O(p^{9/10+\varepsilon})$ inherited at the start of the argument yields Proposition 5.1. ∎

5.4. *Proof of Theorem* 1.1

Proof of Theorem 1.1 (*assuming Proposition* 5.2 *and Theorem* 1.2). The main terms in Theorem 1.1 were computed in Section 5.1, incurring a cost (cf. (5.7))

$$\ll p^{3/4+\theta/2+\varepsilon}. \tag{5.59}$$

Without loss of generality we have $0 \le \theta \le 7/64$ by Kim–Sarnak [32]. It suffices to bound $\mathcal{S}_{N,M,p,d}$ and $\widetilde{\mathcal{S}}_{N,M,p,d}$ for $d = 1$ and p and all M, N satisfying (5.18).

Applying Lemma 5.4 we obtain

$$\mathcal{S}_{N,M,p,1} \ll p^{1/2+\varepsilon} \quad \text{for all } M, N \text{ satisfying (5.18)}. \tag{5.60}$$

Applying Lemma 5.6 we have

$$\widetilde{\mathcal{S}}_{N,M,p,1} \ll p^{7/8+\theta/4+\varepsilon} + p^{15/16+\theta/8+\varepsilon} \ll p^{15/16+\theta/8+\varepsilon} \text{ for all } M, N \text{ satisfying (5.18)}, \tag{5.61}$$

where the last inequality follows by the above fact about θ.

Recalling (5.29), it suffices to consider $\mathcal{S}_{N,M,p,p}$. Applying Lemma 5.7 we have

$$\mathcal{S}_{N,M,p,p} \ll p^{1/2+\theta+\varepsilon} \quad \text{for all } M \le N \le 20M. \tag{5.62}$$

We can now assume $N \ge 20M$. Let α and β be such that

$$M := p^{\alpha}, \quad N := p^{\beta} \quad \text{such that} \quad \alpha + \beta \le 2, \quad \alpha \ge 0, \quad \beta \ge 0.$$

Let

$$\eta_0 := \tfrac{1}{600}.$$

We will now prove that

$$\mathcal{S}_{M,N,p,p} \ll p^{1-\eta_0}. \tag{5.63}$$

Lemma 5.2 guarantees that (5.63) holds when

$$\beta < \tfrac{4}{3} - \tfrac{4}{3}\eta_0 - \tfrac{2}{3}\alpha. \tag{5.64}$$

Lemma 5.9 guarantees that (5.63) holds when

$$\beta > 1 + \alpha + 2\eta_0. \tag{5.65}$$

Proposition 5.2 guarantees that (5.63) holds when

$$\begin{aligned} 0 \le \alpha \le \tfrac{151}{300} \quad &\text{and} \quad \beta < \tfrac{149}{150} + \alpha, \\ \alpha > \tfrac{151}{300} \quad &\text{and} \quad \beta < \tfrac{449}{300}. \end{aligned} \tag{5.66}$$

Lemma 5.8 guarantees that (5.63) holds when

$$\alpha < 1/2 - 2\eta_0 \quad \text{and} \quad \beta > 1 + 2\eta_0. \tag{5.67}$$

Plotting the inequalities (5.64)–(5.67) shows that now only the solid trapezoid in the $\alpha\beta$-plane with vertices

$$\left(\tfrac{299}{600}, \tfrac{901}{600}\right), \quad \left(\tfrac{151}{300}, \tfrac{449}{300}\right), \quad \left(\tfrac{149}{300}, \tfrac{149}{100}\right), \quad \left(\tfrac{149}{300}, \tfrac{3}{2}\right)$$

has to be considered. Writing out the exponents of Proposition 5.1 (which was established under the assumption of Theorem 1.2) we have

$$\begin{aligned} &\tfrac{\alpha}{2} - \tfrac{\beta}{2} + \tfrac{161}{108}, && \tfrac{\alpha}{2} - \tfrac{\beta}{4} + \tfrac{241}{216}, && \tfrac{\alpha}{4} - \tfrac{\beta}{2} + \tfrac{349}{216}, \\ &\tfrac{\alpha}{4} - \tfrac{\beta}{4} + \tfrac{67}{54}, && \tfrac{\alpha}{4} - \tfrac{\beta}{8} + \tfrac{455}{432}, && \tfrac{\alpha}{4} + \tfrac{187}{216}, \quad \tfrac{9}{10}. \end{aligned}$$

A computation shows that each linear function evaluated at each of the four vertices is less than $1 - \eta_0$. Thus (5.59)–(5.62) are subsumed by (5.63). The rest of the paper is dedicated to proving Proposition 5.2 and Theorem 1.2, and that will complete the proof of Theorem 1.1. ∎

6. Automorphic preliminaries II (integral weight)

6.1. *Maass forms*

We give a brief background on Maass forms relevant to our setting. One can see [16, Section 4], [22, Section 2] and [9, Sections 2 and 5] for supplementary material.

Throughout $\kappa = 0$ or 1. For $\gamma \in \mathrm{SL}_2(\mathbb{R})$, define the weight κ slash operator for real analytic forms by

$$g|_\kappa \gamma := j(\gamma, \tau)^{-\kappa} g(\gamma\tau), \quad j(\gamma, \tau) := \frac{c\tau + d}{|c\tau + d|} = e^{i \arg(c\tau + d)},$$

where the argument is always chosen in $(-\pi, \pi]$. The weight κ Laplacian is defined by

$$\Delta_\kappa := y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) - i\kappa y \frac{\partial}{\partial x}.$$

A smooth function $g : \mathbb{H} \to \mathbb{C}$ is an eigenfunction of Δ_κ with eigenvalue $\lambda \in \mathbb{C}$ if

$$(\Delta_\kappa + \lambda) g = 0.$$

All eigenfunctions of Δ_κ are real analytic since it is an elliptic operator.

Let χ be a character modulo D with $\chi(-1) = (-1)^\kappa$. A function $g : \mathbb{H} \to \mathbb{C}$ is *automorphic of weight κ and nebentypus χ* for $\Gamma_0(D)$ if

$$g|_\kappa \gamma = \chi(d) g \quad \text{for all } \gamma := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(D).$$

Let $\mathcal{A}_\kappa(D, \chi)$ denote the space of such functions. If $g \in \mathcal{A}_\kappa(D, \chi)$ is a smooth eigenfunction of Δ_κ that also satisfies the growth condition

$$g(\tau) \ll y^\sigma + y^{1-\sigma} \quad \text{for all } \tau := x + iy \in \mathbb{H} \text{ and some } \sigma > 0,$$

then it is called a *Maass form.* Let

$$\mathcal{L}_\kappa(D,\chi) := \{g \in \mathcal{A}_\kappa(D,\chi) : \|g\| < \infty\},$$

where the norm is induced by the Petersson inner product

$$\langle g_1, g_2\rangle := \int_{\Gamma_0(D)\backslash\mathbb{H}} g_1(\tau)\overline{g_2(\tau)}\, d\mu(\tau), \quad d\mu := \frac{dxdy}{y^2}.$$

Let $\mathcal{R}_\kappa(D,\chi)$ denote the subspace of $\mathcal{L}_\kappa(D,\chi)$ consisting of smooth functions g such that g and $\Delta_\kappa g$ are bounded on $\mathbb{H}$. One can show that $\mathcal{R}_\kappa(D,\chi)$ is dense in $\mathcal{L}_\kappa(D,\chi)$. For all $g_1, g_2 \in \mathcal{R}_\kappa(D,\chi)$ we have

$$\langle \Delta_\kappa g_1, g_2\rangle = \langle g_1, \Delta_\kappa g_2\rangle.$$

Furthermore, for any $g \in \mathcal{R}_\kappa(D,\chi)$ we have

$$\langle g, -\Delta_\kappa g\rangle \geq \frac{|\kappa|}{2}\left(1 - \frac{|\kappa|}{2}\right) \geq 0.$$

Thus by a theorem of Friedrichs, the operator $-\Delta_\kappa$ has a unique self-adjoint extension to $\mathcal{L}_\kappa(D,\chi)$ (which we also denote by $-\Delta_\kappa$). Then by a theorem of von Neumann, the space $\mathcal{L}_\kappa(D,\chi)$ has a complete spectral resolution with respect to $-\Delta_\kappa$. There is a continuous, discrete, and residual spectrum, worked out in detail by Maass and Selberg.

Let $\mathfrak{a}$ be a cusp of $\Gamma_0(D)$ and

$$\Gamma_0(D)_\mathfrak{a} := \{\gamma \in \Gamma_0(D) : \gamma\mathfrak{a} = \mathfrak{a}\},$$

denote its stability group. Let $\sigma_\mathfrak{a}$ denote the unique (up to translation on the right) matrix in $\mathrm{SL}_2(\mathbb{R})$ satisfying $\sigma_\mathfrak{a}\infty = \mathfrak{a}$ and $\sigma_\mathfrak{a}^{-1}\Gamma_0(D)_\mathfrak{a}\sigma_\mathfrak{a} = \Gamma_0(D)_\infty$. We say $\mathfrak{a}$ is *singular* when

$$\chi\left(\sigma_\mathfrak{a}\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\sigma_\mathfrak{a}^{-1}\right) = 1.$$

For each singular cusp $\mathfrak{a}$ (and only at such cusps), the Eisenstein series is defined by

$$E_\mathfrak{a}(\tau,s,\chi) = \sum_{\gamma\in\Gamma_0(D)_\mathfrak{a}\backslash\Gamma_0(D)} \overline{\chi}(\gamma) j(\sigma_\mathfrak{a}^{-1}\gamma,\tau)^{-\kappa}(\operatorname{Im}\sigma_\mathfrak{a}^{-1}\gamma\tau)^s, \quad \operatorname{Re} s > 1 \text{ and } \tau \in \mathbb{H}.$$

One can check that each $E_\mathfrak{a}$ is independent of the choice of the scaling matrix $\sigma_\mathfrak{a}$. Moreover, if $\mathfrak{b} = \gamma\mathfrak{a}$ are $\Gamma_0(D)$-equivalent cusps, then (cf. [66, (3.2)])

$$E_{\gamma\mathfrak{a}}(\tau,s,\chi) = \overline{\chi}(\gamma)E_\mathfrak{a}(\tau,s,\chi). \tag{6.1}$$

Selberg [57] proved that $E_\mathfrak{a}(\tau,s,\chi)$ has an analytic continuation to the whole complex plane with only finitely many simples poles s with $1/2 < s \leq 1$. In particular, when χ is non-principal there are no poles in the region $\operatorname{Re} s \geq 1/2$. When χ is principal, there is only one simple pole at $s = 1$ in this region with constant (but automorphic) residue

$$\operatorname{Res}_{s=1} E_\mathfrak{a}(\tau,s,\chi) = \frac{1}{\operatorname{Vol}(\Gamma_0(D)\backslash\mathbb{H})}.$$

If s is not a pole of $E_{\mathfrak{a}}(\tau, s, \chi)$, then $E_{\mathfrak{a}}(\tau, s, \chi)$ is a Maass form with eigenvalue $\lambda(s)$, but is not in $\mathcal{L}_\kappa(D, \chi)$. The continuous spectrum is composed of all the Eisenstein series on the critical line $s = 1/2 + it$.

The remainder of the spectrum is discrete and is spanned by *Maass cusp forms*. It is countable and of finite multiplicity (with ∞ being the only limit point). We denote it by

$$\lambda_1 \leq \lambda_2 \leq \cdots .$$

To summarise, every $g \in \mathcal{L}_\kappa(D, \chi)$ decomposes as

$$\begin{aligned} g(\tau) &= \sum_{j \geq 0} \langle g, u_j \rangle u_j(\tau) \\ &\quad + \sum_{\mathfrak{a}} \frac{1}{4\pi} \int_{\operatorname{Re} s = 1/2} \langle g, E_{\mathfrak{a}}(\star, 1/2 + it, \chi) \rangle E_{\mathfrak{a}}(\tau, 1/2 + it, \chi)\, dt, \end{aligned} \tag{6.2}$$

where $u_0(\tau)$ is the constant function of Petersson norm 1 (if $\kappa = 0$), $\mathcal{C}_\kappa(D, \chi) = \{u_j\}_{j \geq 1}$ denotes an orthonormal basis of Maass cusp forms, and $\{\mathfrak{a}\}$ runs over all singular cusps of $\Gamma_0(D)$ relative to χ. The convergence in (6.2) is with respect to the norm topology. If $g \in \mathcal{B}_\kappa(D, \chi)$, then (6.2) converges pointwise absolutely and uniformly on compacta.

A Maass cusp form decays exponentially at the cusps and admits a Fourier expansion with the zeroth Fourier coefficient vanishing. At ∞, such an expansion is given by

$$g(\tau) = \sum_{n=-\infty}^{\infty} \rho_g(n) W_{\frac{\kappa n}{2|n|}, it}(4\pi |n| y) e(nx), \tag{6.3}$$

where $W_{\alpha,\beta}(y)$ is the usual Whittaker function and $\lambda_g := 1/4 + t_g^2$ is the Laplace eigenvalue of g. We call t_g the *spectral parameter* of g. When $\kappa = 0$, note that $t_g \in [-i\theta, i\theta] \cup [0, \infty)$, where $\theta = 7/64$ is the best currently known [32].

The Eisenstein series has the expansion

$$\begin{aligned} E_{\mathfrak{a}}(\tau, 1/2 + it, \chi) &= \delta_{\mathfrak{a}=\infty} y^{1/2+it} + \phi_{\mathfrak{a}}(1/2 + it) y^{1/2-it} \\ &\quad + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \rho_{\mathfrak{a}}(n, t) W_{\frac{\kappa n}{2|n|}, it}(4\pi |n| y) e(nx), \end{aligned} \tag{6.4}$$

where $\phi_{\mathfrak{a}}(1/2 + it)$ is the $(\mathfrak{a}, \infty)$ entry of the relevant scattering matrix.

6.2. *Holomorphic forms*

Let χ be a Dirichlet character modulo D with $\chi(-1) = (-1)^\kappa$ ($\kappa = 0$ or 1). For $\ell \in \mathbb{N}$ with $\ell \equiv \kappa \pmod 2$, let $\mathcal{S}_\ell(D, \chi)$ denote the space of holomorphic cusp forms of level D, weight ℓ and nebentypus χ. This space is equipped with the Petersson inner product

$$\langle g_1, g_2 \rangle = \int_{\Gamma_0(D) \backslash \mathbb{H}} y^\ell g_1(\tau) \overline{g_2(\tau)}\, d\mu(\tau).$$

We also have the Fourier expansion (at ∞)

$$g(\tau) = \sum_{n\geq 1} \rho_g(n)(4\pi n)^{\ell/2} e(n\tau). \tag{6.5}$$

6.3. *Hecke operators and newform theory*

Recall that $\mathcal{L}_\kappa(D, \chi)$ (and the subspace generated by Maass cusp forms) is acted on by an algebra $\mathbf{T}$ generated by Hecke operators $\{T_n\}_{n\geq 1}$. Each operator is defined by

$$(T_n g)(\tau) = \frac{1}{\sqrt{n}} \sum_{ad=n} \chi(a) \sum_{b \pmod d} g\left(\frac{a\tau + b}{d}\right).$$

These operators are commutative and multiplicative. They also satisfy the relation

$$T_m T_n = \sum_{d|(m,n)} \chi(d) T_{mn/d^2}. \tag{6.6}$$

Let $\mathbf{T}^{(D)}$ denote the subalgebra generated by $\{T_n\}_{(n,D)=1}$. We call a Maass cusp form which is an eigenform for $\mathbf{T}^{(D)}$ a *Hecke–Maass* cusp form. The elements of $\mathbf{T}^{(D)}$ are normal with respect to the Petersson inner product, so the cuspidal subspace of $\mathcal{L}_\kappa(D, \chi)$ admits an orthonormal basis of Hecke–Maass cusp forms. For a Hecke–Maass cusp form g, the following relations hold:

$$\sqrt{n}\rho_g(\pm n) = \rho_g(\pm 1)\lambda_g(n) \quad \text{for } (n, D) = 1, \tag{6.7}$$

where $\lambda_g(n)$ denotes the eigenvalue of T_n, and

$$\sqrt{m}\rho_g(m)\lambda_g(n) = \sum_{d|(m,n)} \chi(d)\rho_g\left(\frac{mn}{d^2}\right)\sqrt{\frac{mn}{d^2}}, \tag{6.8}$$

$$\sqrt{mn}\rho_g(mn) = \sum_{d|(m,n)} \chi(d)\mu(d)\rho_g\left(\frac{m}{d}\right)\sqrt{\frac{m}{d}}\lambda_g\left(\frac{n}{d}\right). \tag{6.9}$$

The space of newforms is defined to be the space spanned by the Hecke–Maass cusp forms orthogonal to the subspace spanned by the oldforms. If g is a Hecke form and in the new subspace, then g is a Hecke eigenform of all Hecke operators by Atkin–Lehner theory and the above relations are satisfied for all n.

For a Hecke–Maass cusp form g, we have the pointwise bound

$$|\lambda_g(n)| \leq n^{\theta+\varepsilon}$$

and the Rankin–Selberg bound

$$\sum_{n\leq x} |\lambda_g(n)|^2 \ll_\varepsilon (D(1+|t|)x)^\varepsilon x.$$

If g is an L^2-normalised newform of weight $\kappa \in \{0, 1\}$ and level D, then by [22, (30)] we have

$$(D(1+|t_g|))^{-\varepsilon}\left(\frac{\cosh(\pi t_g)}{D(1+|t_g|)^{\kappa}}\right)^{1/2} \ll_{\varepsilon} |\rho_g(1)| \ll_{\varepsilon} (D(1+|t_g|))^{\varepsilon}\left(\frac{\cosh(\pi t_g)}{D(1+|t_g|)^{\kappa}}\right)^{1/2}. \tag{6.10}$$

The upper bound (resp. lower bound) is a consequence of the seminal works of Hoffstein–Lockhart [24] (resp. Duke–Friedlander–Iwaniec [16]).

We now give a more explicit treatment of bases and newforms due to Blomer and Milićević ([9, Section 5] and [29]). Let $\mathcal{B}_\kappa(D, t, \chi)$ (resp. $\mathcal{H}_\ell(D, \chi)$) denote an L^2-basis for $\mathcal{A}_\kappa(D, t, \chi)$ (resp. $\mathcal{S}_\ell(D, \chi)$). In general, both of these bases will include oldforms. We will focus on Maass forms since the holomorphic case will be the same, only requiring small notational changes. Suppose χ has conductor $D_\chi^\star$ and underlying primitive character $\chi^\star$. For $u \mid D$, let $\widetilde{\chi}$ modulo u be the character induced by $\chi^\star$ and $\mathcal{B}_\kappa^\star(u, D, t, \widetilde{\chi}) \subseteq \mathcal{B}_\kappa(D, t, \chi)$ denote the set of all $L^2(\Gamma_0(D)\backslash\mathbb{H})$-normalised newforms of level u and spectral parameter t. We write $g|_d(\tau) := g(d\tau)$. By Atkin–Lehner theory we have

$$\mathcal{A}_\kappa(D, t, \chi) = \mathop{\textcircled{$\perp$}}_{\substack{D_\chi^\star | u\\ u|D}} \ \mathop{\textcircled{$\perp$}}_{g\in\mathcal{B}_\kappa^\star(u,D,t,\widetilde{\chi})} \ \bigoplus_{d|D/u} g|_d \cdot \mathbb{C}. \tag{6.11}$$

The first two sums in (6.11) are orthogonal, but the last is not orthogonal in general. Gram–Schmidt is required to make this sum orthogonal.

An orthogonal basis $\mathcal{B}_\kappa(D, \chi)$ for $\mathcal{A}_\kappa(D, \chi)$ is produced by collecting all spectral parameters,

$$\mathcal{B}_\kappa(D, \chi) := \coprod_t \mathcal{B}_\kappa(D, t, \chi).$$

Correspondingly,

$$\mathcal{B}_\kappa(u, D, \widetilde{\chi}) := \coprod_t \mathcal{B}_\kappa^\star(u, D, t, \widetilde{\chi}).$$

For a newform $g \in \mathcal{B}_\kappa(u, D, \widetilde{\chi})$, define the arithmetic functions

$$r_g(c) := \sum_{b|c} \frac{\mu(b)\lambda_g(b)^2}{b}\left(\sum_{d|b}\frac{\chi(b)}{b}\right)^{-2}, \quad \alpha(c) := \sum_{b|c}\frac{\mu(b)\chi(b)^2}{b^2},$$

$$\beta(c) := \sum_{b|c}\frac{\mu^2(b)\chi(b)}{b}, \quad L(g,s)^{-1} := \sum_c \frac{\mu_g(c)}{c^s},$$

where

$$\mu_g(p) = -\lambda_g(p), \quad \mu_g(p^2) = \chi(p), \quad \mu_g(p^\nu) = 0 \quad \text{for } \nu > 2.$$

For $d \mid e$, define

$$\xi_e'(d) := \frac{\mu(e/d)\lambda_g(e/d)}{r_g(e)^{1/2}(e/d)^{1/2}\beta(e/d)}, \quad \xi_e''(d) := \frac{\mu_g(e/d)}{(e/d)^{1/2}(r_g(e)\alpha(e))^{1/2}}$$

Now write $e = e_1 e_2$ uniquely with e_1 squarefree, e_2 squarefull and $(e_1, e_2) = 1$. Then for $d \mid e$ define

$$\xi_e(d) := \xi'_{e_1}((e_1, d))\xi''_{e_2}((e_2, d)) \ll e^{\varepsilon}(e/d)^{\theta-\frac{1}{2}}.$$

Lemma 6.1 ([9, Lemma 9]). *Let $u \mid D$ and $g^\star \in \mathcal{B}_\kappa(u, D, \widetilde{\chi}) \subset \mathcal{B}_\kappa(D, \chi)$ be an $L^2(\Gamma_0(D)\backslash\mathbb{H})$-normalised newform of level u. Then the set of functions*

$$\Big\{g^{(e)} := \sum_{d|e} \xi_e(d) g^\star|_d(\tau) : e \mid D/u\Big\}$$

is an orthonormal basis for the space $\bigoplus_{d|D/u} g^\star|_d \cdot \mathbb{C}$. If g is any member of this basis, then its Fourier coefficients satisfy the bound

$$\sqrt{n}\rho_g(n) \ll (nD)^{\varepsilon} n^{\theta}(D, n)^{1/2-\theta}|\rho_{g^\star}(1)|. \tag{6.12}$$

Note that one can also see [56, Theorem 3.2] for the above lemma.

Proof of Lemma 6.1. The proof is verbatim that of [9, Lemma 9] with χ replacing the principal character modulo D with χ. ∎

Elements $g \in \mathcal{B}_\kappa(D, \chi)$ (or $g \in \mathcal{H}_\ell(D, \chi)$) have multiplicative-like properties. Given $r \in \mathbb{N}$, if $m = rm' \in \mathbb{N}$ with $(m', r) = 1$, then by [7, p. 74] we have

$$\sqrt{m}\rho_g(m) = \sum_{d|(D,(r/(r,D))} \mu(d)\chi(d)\lambda_{g^\star}\left(\frac{r}{d(r, D)}\right)\left(\frac{(r, D)m'}{d}\right)^{1/2} \rho_g\left(\frac{(r, D)m'}{d}\right), \tag{6.13}$$

where $g^\star$ is the underlying newform. In particular, if $(r, D) = 1$ we have

$$\sqrt{m}\rho_g(m) = \lambda_{g^\star}(r)\sqrt{m'}\rho_g(m') \tag{6.14}$$

Moreover, if $g^\star$ satisfies the Ramanujan–Petersson conjecture (i.e. an integral weight holomorphic eigenform) and a_m is any finite sequence of complex numbers then

$$\Big|\sum_m a_m \sqrt{m}\rho_g(m)\Big|^2 \leq \tau(r)^2 \sum_{d|(r,D)} \Big|\sum_{m'} a_{rm'}\sqrt{dm'}\rho_g(dm')\Big|^2. \tag{6.15}$$

6.4. *Kuznetsov–Proskurin formula and spectral inequalities*

Let χ be a character modulo D with $\chi(-1) = 1$. Let $\phi : [0, \infty) \to \mathbb{C}$ have continuous derivatives up to third order and satisfy

$$\phi(0) = \phi'(0) = 0, \quad \phi^j(x) \ll (1 + x)^{-3} \quad \text{for } j = 1, 2, 3.$$

Define the transforms

$$\dot{\phi}(\ell) := 4i^{\ell} \int_0^{\infty} \phi(x) J_{\ell-1}(x) \frac{dx}{x},$$
$$\widetilde{\phi}(t) := 2\pi i \int_0^{\infty} \phi(x) \frac{J_{2it}(x) - J_{-2it}(x)}{\sinh(\pi t)} \frac{dx}{x},$$

$$\check{\phi}(t) := 8\int_0^\infty \phi(x)\cosh(\pi t)K_{2it}(x)\,\frac{dx}{x}.$$

These transforms are normalised like those occurring in [9, Section 5] and [14]. For $D \mid c$ and $m, n \in \mathbb{Z}$, define the Kloosterman sum (at $\infty\infty$) by

$$K(m,n,c,\chi) := \sideset{}{^*}\sum_{d \,(\mathrm{mod}\, c)} \overline{\chi}(d)e\left(\frac{md + n\overline{d}}{c}\right),$$

where the superscript $*$ denotes the condition $(d,c) = 1$.

Lemma 6.2 ([14, Lemma 4.5] *and* [9, Lemma 10]). *Let ϕ be as above and χ be a character modulo D with $\chi(-1) = 1$. For $\ell \geq 2$ and $\ell \equiv 0 \pmod 2$, let $\mathcal{B}_0(D,\chi)$ (resp. $\mathcal{H}_\ell(D,\chi)$) denote the orthonormal basis of Maass cusp forms (resp. holomorphic cusp forms) given above. Recalling the notations (6.3)–(6.5), for $m,n \in \mathbb{N}$ we have*

$$\begin{aligned}\sum_{D\mid c} \frac{1}{c}K(m,n,c,\chi)\phi\left(\frac{4\pi\sqrt{mn}}{c}\right) = &\sum_{\substack{\ell\geq 2\\ \ell\equiv 0 \,(\mathrm{mod}\, 2)}} \sum_{g\in\mathcal{H}_\ell(D,\chi)} \Gamma(\ell)\dot{\phi}(\ell)\sqrt{mn}\,\overline{\rho_g(m)}\rho_g(n)\\ &+ \sum_{g\in\mathcal{B}_0(D,\chi)} \tilde{\phi}(t_g)\frac{\sqrt{mn}}{\cosh(\pi t_g)}\overline{\rho_g(m)}\rho_g(n)\\ &+ \frac{1}{4\pi}\sum_{\mathfrak{a}\text{ sing.}} \sqrt{mn}\int_{-\infty}^{\infty} \frac{\tilde{\phi}(t)}{\cosh(\pi t)}\overline{\rho_{\mathfrak{a}}(m,t)}\rho_{\mathfrak{a}}(n,t)\,dt\end{aligned}$$

and

$$\begin{aligned}\sum_{D\mid c} \frac{1}{c}K(m,-n,c,\chi)\phi\left(\frac{4\pi\sqrt{mn}}{c}\right) = &\sum_{g\in\mathcal{B}_0(D,\chi)} \check{\phi}(t_g)\frac{\sqrt{mn}}{\cosh(\pi t_g)}\overline{\rho_g(m)}\rho_g(-n)\\ &+ \frac{1}{4\pi}\sum_{\mathfrak{a}\text{ sing.}} \sqrt{mn}\int_{-\infty}^{\infty} \frac{\check{\phi}(t)}{\cosh(\pi t)}\overline{\rho_{\mathfrak{a}}(m,t)}\rho_{\mathfrak{a}}(-n,t)\,dt.\end{aligned}$$

6.5. Spectral large sieve and multiplicative sequences I

In this section we record a spectral large sieve inequality for coefficients of Maass forms supported on sequences with multiplicative structure. This approach is originally due to Blomer and Milićević [9, Theorem 13], and is crucial to avoiding the Ramanujan–Petersson conjecture in their treatment of shifted convolution sums. We also record some standard spectral tools in enough generality for our purpose.

Lemma 6.3. *Suppose χ is a character modulo D with $\chi(-1) = (-1)^\kappa$ and conductor $D_\chi^\star := N$ or $4N$, where N is odd and squarefree. Suppose $c \in \mathbb{N}$ with $D \mid c$. Then for $a, b \in \mathbb{N}$ we have*

$$|K(a,b,c,\chi)| \leq 16\tau(c)(a,b,c)^{1/2}c^{1/2}. \tag{6.16}$$

Proof. This bound follows from [35, Corollary 9.14, Propositions 9.4, 9.7, 9.8]. ∎

We now present the standard spectral large sieve.

Lemma 6.4. *Let χ be as in Lemma 6.3 with $\kappa = 0$ and $\{a_m\}$ be a sequence of complex numbers. Suppose $T \geq 1$ and $M \geq 1/2$. Then each of the three quantities*

$$\sum_{\substack{\kappa<\ell\leq T\\ \ell\equiv\kappa \pmod 2}} \Gamma(\ell) \sum_{g\in\mathcal{H}_\ell(D,\chi)} \Big| \sum_{M<m\leq 2M} a_m\sqrt{m}\rho_g(m)\Big|^2,$$

$$\sum_{\substack{g\in\mathcal{B}_0(D,\chi)\\ |t_g|\leq T}} \frac{1}{\cosh(\pi t_g)}\Big| \sum_{M<m\leq 2M} a_m\sqrt{m}\rho_g(\pm m)\Big|^2,$$

$$\sum_{\mathfrak{a}\text{ sing.}} \int_{-T}^{T} \frac{1}{\cosh(\pi t)}\Big| \sum_{M<m\leq 2M} a_m\sqrt{m}\rho_{\mathfrak{a}}(\pm m, t)\Big|^2$$

is bounded, up to a constant depending on ε, by

$$\left(T^2 + \frac{M^{1+\varepsilon}}{D}\right) \sum_{M<m\leq 2M} |a_m|^2.$$

Proof. Observe that this result has been proved in [14, Proposition 4.7], except that the bound that appears there is

$$\left(T^2 + (D_\chi^\star)^{1/2}\frac{M^{1+\varepsilon}}{D}\right) \sum_{M<m\leq 2M} |a_m|^2.$$

The appearance of the conductor in [14] is due to the general estimate for Kloosterman sums in [35, Theorem 9.2]. Since the conductor is either $D_\chi^\star = N$ or $4N$ with N odd and squarefree, we can apply Lemma 6.3 to remove the factor of $(D_\chi^\star)^{1/2}$ in [14, (4.20)]. The rest of the proof is verbatim the same as that of [14, Proposition 4.7]. ■

Lemma 6.5. *Let χ be as in Lemma 6.3 with $\kappa = 0$ and $m \in \mathbb{N}$. Then*

$$\sum_{\substack{|t_g|\leq T\\ g\in\mathcal{B}_0(D,\chi)}} \frac{1}{\cosh(\pi t_g)}|\sqrt{m}\rho_g(m)|^2 \ll_\varepsilon \left(T^2 + \frac{(D,m)^{1/2}m^{1/2}}{D}\right)(Tm)^\varepsilon.$$

Proof. One starts with the “pre-Kuznetsov formula” of [51, Lemma 3]. The proof is then verbatim that of [49, Lemma 2.4], except that in [49, (2.7.3), (2.7.10)], the Kloosterman sum is $S(m, m, c, \chi)$, and an extra divisibility condition $D \mid c$ is added to the summation. One then appeals to Lemma 6.3 and modifies the last two displays in the proof accordingly. ■

Blomer and Milićević prove the following result for Maass forms using a fourth moment approach. The main feature is that it allows one to avoid invoking the Ramanujan –Petersson conjecture (see Remark 6.1 below).

Theorem 6.1 ([9, Theorem 13]). *Let χ be as in Lemma 6.3 with $\kappa = 0$. Let $r \in \mathbb{N}$, $M, T \geq 1$, and let $\{\alpha_{m'}\}_{M \leq m' \leq 2M}$ be any sequence of complex numbers with $|\alpha_{m'}| \leq 1$. Then*

$$\sum_{\substack{|t_g| \leq T \\ g \in \mathcal{B}_0(D,\chi)}} \frac{1}{\cosh(\pi t_g)} \Bigg| \sum_{\substack{M \leq m' \leq 2M \\ (m', rD)=1}} \alpha_{m'} \sqrt{rm'} \rho_g(rm') \Bigg|^2 \ll_\varepsilon (DrMT)^\varepsilon (r, D) \left(T + \frac{r^{1/2}}{D^{1/2}}\right)\left(T + \frac{M}{D^{1/2}}\right) M. \tag{6.17}$$

Proof. This is verbatim the proof of [9, Theorem 13] using (6.6) (for eigenvalues), (6.7), (6.10)–(6.14), and Lemmas 6.4 and 6.5 whenever their principal character analogues are used in the proof. ■

Remark 6.1. For the sake of argument, suppose that $(r, D) = 1$. One could naively apply the Hecke relation (6.14) to the left side of (6.17), then estimate $\sqrt{r}\rho_g(r)$ by (6.12), and finally apply the usual spectral large sieve in Lemma 6.4 to obtain

$$\sum_{\substack{|t_g| \leq T \\ g \in \mathcal{B}_0(D,\chi)}} \frac{1}{\cosh(\pi t_g)} \Bigg| \sum_{\substack{M \leq m' \leq 2M \\ (m', rD)=1}} \alpha_{m'} \sqrt{rm'} \rho_g(rm') \Bigg|^2 \ll_\varepsilon (DrMT)^\varepsilon r^{2\theta} \left(T^2 + \frac{M}{D}\right) M, \tag{6.18}$$

which however is insufficient for our purposes. For the analogous case when g runs over holomorphic forms this naive approach is sufficient because Deligne's bound for $\sqrt{r}\rho_g(r)$ is available.

6.6. *Spectral large sieve and multiplicative sequences II*

We now prove a version of the spectral large sieve inequality for coefficients of Eisenstein series supported on a sequences with multiplicative structure. This will be a generalisation of [9, (5.5)] for a more general nebentypus. The proof uses ideas from the explicit computations in [22, 47] and [7, pp. 76–80]. The machinery set out for Eisenstein series in [34, 66] will be useful throughout the proof.

We set up the notation and preliminary results required for the proof of Lemma 6.7. Let $D \in \mathbb{N}$. A full set of inequivalent cusps of $\Gamma_0(D)$ is given by

$$\left\{ \mathfrak{a} := \frac{1}{w} = \frac{1}{uf} : f \mid D,\ u \in \mathcal{U}_f \right\}, \tag{6.19}$$

where for each $f \mid D$, $\mathcal{U}_f$ is a set of integers coprime to f representing each reduced residue class modulo $\widetilde{f} := (f, D/f)$ exactly once. Moreover, one may always further choose a representative $u \pmod{\widetilde{f}}$ such that $(u, D) = 1$ by adding a suitable multiple of $\widetilde{f}$ (cf. [34, Corollary 3.2]). Note that with these choices of representatives u we have $\frac{u}{f} \sim_\Gamma \frac{1}{uf}$.

Set

$$D' = \frac{D}{f}, \quad D'' = \frac{D'}{(f, D')}, \quad w' = u. \tag{6.20}$$

The stabiliser group of an arbitrary cusp $\mathfrak{a} = \frac{1}{w}$ [34, Proposition 3.3] is given by

$$\Gamma_{\mathfrak{a}} = \{\pm \tau_{\mathfrak{a}}^t : t \in \mathbb{Z}\}, \quad \text{where} \quad \tau_{\mathfrak{a}}^t = \begin{pmatrix} 1 - wD''t & D''t \\ -w^2 D''t & 1 + wD''t \end{pmatrix},$$

and one can take the choice of scaling matrix

$$\sigma_{\mathfrak{a}} = \begin{pmatrix} \sqrt{D''} & 0 \\ w\sqrt{D''} & 1/\sqrt{D''} \end{pmatrix}. \tag{6.21}$$

Observe that [34, Lemma 3.5] (expanding around the cusp at $1/D \sim \infty$) asserts that

$$\sigma_{\mathfrak{a}}^{-1} \Gamma \sigma_{1/D} := \left\{ \begin{pmatrix} a/\sqrt{D''} & b/\sqrt{D''} \\ c\sqrt{D''} & d\sqrt{D''} \end{pmatrix} : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \text{ and } c \equiv -wa \ (\mathrm{mod}\ D) \right\}. \tag{6.22}$$

In particular, [66, (6.4), (6.5)] asserts

$$\Gamma_\infty \backslash \sigma_{\mathfrak{a}}^{-1} \Gamma := \delta_{f=D} \Gamma_\infty \cup \left\{ \begin{pmatrix} * & * \\ c\sqrt{D''} & d\sqrt{D''} \end{pmatrix} : \right.$$
$$\left. c > 0,\ (c, d) = 1,\ c = f\gamma,\ (\gamma, D') = 1,\ d \equiv -u\overline{\gamma} \ (\mathrm{mod}\ (f, D')) \right\}, \tag{6.23}$$

where the union is disjoint.

We recall a useful lemma giving necessary and sufficient conditions for a cusp to be singular.

Lemma 6.6 ([66, Lemma 5.4]). *Let χ be a Dirichlet character modulo D and let $\mathfrak{a} = \frac{1}{uf}$ with $u \mid D$ and $(u, D) = 1$. Then $\mathfrak{a}$ is singular relative to χ if and only if χ is periodic modulo $D/(f, D/f) = [f, D/f]$, equivalently, the primitive character inducing χ has modulus dividing $D/(f, D/f)$.*

We recall a decomposition for χ given in [66, p. 17]. There exist integers f_0 and D_0' such that

$$f_0 \mid f \quad \text{and} \quad D_0' \mid D' \quad \text{with} \quad [f, D'] := f_0 D_0' \text{ and } (f_0, D_0') = 1,$$

such that

$$\chi := \chi^{(D_0')} \chi^{(f_0)}, \tag{6.24}$$

where $\chi^{(D_0')}$ and $\chi^{(f_0)}$ are characters modulo D_0' and f_0 respectively. The choices for f_0 and D_0' may not be unique. This decomposition will be useful in the proof of Lemma 6.7, but will not feature in the final statement.

Given $\mathfrak{a} = 1/w$ and $\mu \in \sigma_{\mathfrak{a}}^{-1}\Gamma$ (written as in (6.22)), the argument in [66, p. 19] proves that $\overline{\chi}(\sigma_{\mathfrak{a}}\mu)$ depends only on the coset $\Gamma_\infty \mu$. Thus $\overline{\chi}(\sigma_{\mathfrak{a}}\mu)$ depends only on the data contained in (6.23). In particular, [66, (6.6)] asserts that

$$\overline{\chi}(\sigma_{\mathfrak{a}}\mu) = \chi(a) = \chi^{(D_0')}(-\overline{u}\gamma)\chi^{(f_0)}(\overline{d}). \tag{6.25}$$

We recall the definition of the Gauss sum in (4.15). Let Ψ be a Dirichlet character modulo c and $c \mid s$. Then for $n \in \mathbb{N}$ we have the Gauss sum

$$\mathcal{G}_{\Psi}(n;s) := \sum_{d \ (\mathrm{mod}\, s)}^{*} \Psi(d) e\left(\frac{nd}{s}\right), \tag{6.26}$$

where the superscript $*$ denotes the condition $(d,s) = 1$. Without loss of generality one can replace Ψ with its underlying primitive character $\Psi^{\star}$ in (6.26).

Given a cusp $\mathfrak{a} = \frac{1}{uf}$, $m \in \mathbb{N}$, $t \in \mathbb{R}$, and characters Ψ_1 and Ψ_2 whose moduli divide D/f and f respectively, consider the series

$$\mathcal{S}(t,m;\Psi_1,\Psi_2,f) := \sum_{\substack{\gamma>0 \\ (\gamma,D/f)=1}} \frac{\Psi_1(\gamma)}{\gamma^{1+2it}} \mathcal{G}_{\Psi_2}(m;\gamma f), \tag{6.27}$$

Note that (6.27) also satisfies

$$\mathcal{S}(t,m;\Psi_1,\Psi_2,f) = \mathcal{S}(t,m;\Psi_1^{\star},\Psi_2^{\star},f). \tag{6.28}$$

For technical convenience we restrict our attention to even characters in the next result.

Lemma 6.7. *Suppose $D \in \mathbb{N}$ and χ is an even character modulo D such that all cusps $\mathfrak{a}$ of $\Gamma_0(D)$ are singular with respect to χ, and let $\rho_{\mathfrak{a}}(m,t)$ be as in (6.4). For a given $r \in \mathbb{N}$, let $\{a_m\}$ be a finite sequence of complex numbers supported only on integers $m = rm'$ with $(r,m') = 1$. Then*

$$\sum_{\mathfrak{a}} \Big| \sum_{m} a_m \sqrt{|m|}\rho_{\mathfrak{a}}(m,t)\Big|^2 \leq 16\tau(D)^4\tau(r)^4 \sum_{d \mid (r,D)} \sum_{\mathfrak{a}} \Big| \sum_{\substack{m' \\ (m',r)=1}} a_{rm'}\sqrt{dm'}\rho_{\mathfrak{a}}(dm',t)\Big|^2, \tag{6.29}$$

Proof. Without loss of generality we will work with a complete set of inequivalent cusps given in (6.19) where each $(u,D) = 1$ (cf. (6.1)). Combining [51, (12)–(14)], (6.20)–(6.27) and [50, (13.14.31)] we obtain

$$\sqrt{|m|}\rho_{\mathfrak{a}}(m,t) = \frac{|m|^{it}\pi^{1/2+it}}{\Gamma(1/2+it)} \left(\frac{(f,D')}{Df}\right)^{1/2+it} \sum_{\substack{\gamma>0 \\ (\gamma,D/f)=1}} \frac{\chi^{(D_0')}(-\overline{u}\gamma)}{\gamma^{1+2it}} \times \sum_{\substack{0\leq d<\gamma f \\ (d,\gamma f)=1 \\ d\gamma\equiv -u \ (\mathrm{mod}\, \widetilde{f})}} \overline{\chi^{(f_0)}}(d) e\left(\frac{md}{\gamma f}\right), \tag{6.30}$$

where $\widetilde{f} := (f, D/f)$. Detecting the congruence in (6.30) with multiplicative characters we obtain

$$\sqrt{|m|}\rho_{\mathfrak{a}}(m,t) = \frac{|m|^{it}\pi^{1/2+it}}{\Gamma(1/2+it)}\left(\frac{(f,D')}{Df}\right)^{1/2+it}$$
$$\times \frac{1}{\phi(\tilde{f})}\sum_{\psi \ (\mathrm{mod}\ \tilde{f})} \overline{\chi^{(D_0')}}\psi(-u)\mathcal{S}(t,m;\psi\chi^{(D_0')},\psi\overline{\chi^{(f_0)}},f). \tag{6.31}$$

All cusps are singular by hypothesis. Summing (6.31) over all m, squaring, and then summing over all the cusps we obtain

$$\sum_{\mathfrak{a}}\Big|\sum_m a_m\sqrt{|m|}\rho_{\mathfrak{a}}(m,t)\Big|^2 = \frac{\pi}{|\Gamma(1/2+it)|^2}\sum_{f|D}\frac{\tilde{f}}{Df\phi(\tilde{f})^2}$$
$$\times \sum_{u\in\mathcal{U}_f}|\overline{\chi^{(D_0')}}(-u)|^2\Big|\sum_{\psi \ (\mathrm{mod}\ \tilde{f})}\overline{\psi}(-u)\sum_m a_m|m|^{it}\mathcal{S}(t,m;\psi\chi^{(D_0')},\psi\overline{\chi^{(f_0)}},f)\Big|^2,$$

where $\mathcal{U}_f$ represents a complete system of residues coprime to $\tilde{f}$ and also coprime to D. We may remove the $|\overline{\chi^{(D_0')}}(-u)|^2$ factor since $|\overline{\chi^{(D_0')}}(-u)| = 1$. Applying Parseval's identity above yields

$$\sum_{\mathfrak{a}}\Big|\sum_m a_m\sqrt{|m|}\rho_{\mathfrak{a}}(m,t)\Big|^2 = \frac{\pi}{|\Gamma(1/2+it)|^2}\sum_{f|D}\frac{\tilde{f}}{Df\phi(\tilde{f})}$$
$$\times \sum_{\psi \ (\mathrm{mod}\ \tilde{f})}\Big|\sum_m a_m|m|^{it}\mathcal{S}(t,m;\psi\chi^{(D_0')},\psi\overline{\chi^{(f_0)}},f)\Big|^2. \tag{6.32}$$

Next we set

$$(\psi\chi^{(D_0')})^* =: \chi_1 \quad \text{and} \quad (\psi\overline{\chi^{(f_0)}})^* =: \chi_2,$$

where χ_i is primitive of modulus q_i. A necessary condition on χ_1 and χ_2 is that $\chi_1\overline{\chi_2}\sim\chi$, where $\sim$ means that both sides are induced by the same primitive character. Recalling (6.28) and moving the sum on ψ to the inside of (6.32) gives

$$\sum_{\mathfrak{a}}\Big|\sum_m a_m\sqrt{|m|}\rho_{\mathfrak{a}}(m,t)\Big|^2$$
$$= \frac{\pi}{|\Gamma(1/2+it)|^2}\sum_{f|D}\frac{\tilde{f}}{Df\phi(\tilde{f})}\sum_{q_1|D/f}\sum_{q_2|f}\sideset{}{'}\sum_{\substack{\chi_1 \ (\mathrm{mod}\, q_1)\\ \chi_2 \ (\mathrm{mod}\, q_2)\\ \chi_1\overline{\chi_2}\sim\chi}}\Big|\sum_m a_m|m|^{it}\mathcal{S}(t,m;\chi_1,\chi_2,f)\Big|^2$$
$$\times \sum_{\psi \ (\mathrm{mod}\ \tilde{f})}\delta(\chi,\chi_1,\chi_2,\psi), \tag{6.33}$$

where $'$ denotes summation over primitive characters only, and

$$\delta(\chi,\chi_1,\chi_2,\psi) := \begin{cases} 1 & \text{if } (\psi\chi^{(D_0')})^* = \chi_1 \text{ and } (\psi\overline{\chi^{(f_0)}})^* = \chi_2, \\ 0 & \text{otherwise.}\end{cases}$$

The argument in [66, pp. 21–22] proves that

$$\sum_{\psi \ (\mathrm{mod}\ \tilde{f})}\delta(\chi,\chi_1,\chi_2,\psi) = 1,$$

under the conditions in the second, third and fourth summations in (6.33). Thus

$$\sum_{\mathfrak{a}} \Big| \sum_m a_m \sqrt{|m|} \rho_{\mathfrak{a}}(m,t) \Big|^2 = \frac{\pi}{|\Gamma(1/2+it)|^2} \\ \times \sum_{f \mid D} \frac{\tilde{f}}{D f \phi(\tilde{f})} \sum_{q_1 \mid D/f} \sum_{q_2 \mid f} \sideset{}{'}\sum_{\substack{\chi_1 \ (\operatorname{mod} q_1) \\ \chi_2 \ (\operatorname{mod} q_2) \\ \chi_1 \overline{\chi_2} \sim \chi}} \Big| \sum_m a_m |m|^{it} \mathcal{S}(t,m;\chi_1,\chi_2,f) \Big|^2. \tag{6.34}$$

For a given $r \in \mathbb{N}$, recall that $\{a_m\}$ is supported on integers $m = rm'$ such that $(r,m') = 1$. Our goal is now to write $\mathcal{S}(t,m;\chi_1,\chi_2;f)$ in terms of $\mathcal{S}(dm';\chi_1,\chi_2;\bullet)$ for $d \mid (r,D)$. For a given f and γ we write

$$f = q_2 f' f'', \quad f' \mid q_2^\infty, \quad (f'',q_2) = 1, \tag{6.35}$$

and

$$\gamma = \gamma'\gamma'', \quad \gamma' \mid q_2^\infty, \quad (\gamma'',q_2) = 1. \tag{6.36}$$

The Chinese remainder theorem, orthogonality, [1, p. 165 and Theorem 8.19], and the fact that χ_2 is primitive are used in the following computation:

$$\begin{aligned} \mathcal{G}_{\chi_2}(m;\gamma f) &= \mathcal{G}_{\chi_2}(m;\gamma'\gamma'' q_2 f' f'') \\ &= \chi_2(\gamma'' f'') \mathcal{G}_{\chi_2}(m; q_2 f' \gamma') r(m;\gamma'' f'') \\ &= \delta_{f'\gamma' \mid m} f'\gamma' \chi_2(\gamma'' f'') \mathcal{G}_{\chi_2}\Big(\frac{m}{f'\gamma'}; q_2\Big) r(m;\gamma'' f''), \\ &= \delta_{f'\gamma' \mid m} f'\gamma' \chi_2(\gamma'' f'') \overline{\chi_2}\Big(\frac{m}{f'\gamma'}\Big) \mathcal{G}_{\chi_2}(1;q_2) r(m;\gamma'' f''). \end{aligned} \tag{6.37}$$

where $r(m;c) := G_{1_c}(m,c)$ (Ramanujan sum) and $\delta_{f'\gamma' \mid m} = 1$ is the indicator function for $f'\gamma' \mid m$. Let

$$r = r'_{q_2} r'^{(q_2)} \quad \text{and} \quad m' = m'_{q_2} m'^{(q_2)}$$

be such that

$$r'_{q_2}, m'_{q_2} \mid q_2^\infty \quad \text{and} \quad (r'^{(q_2)} m'^{(q_2)}, q_2) = 1.$$

In this notation we see that (6.37) is 0 unless

$$f'\gamma' = r'_{q_2} m'_{q_2}. \tag{6.38}$$

Recalling (6.27) and combining (6.35)–(6.38) we obtain (after relabelling γ'' as γ),

$$\mathcal{S}(t,rm';\chi_1,\chi_2,f) = \delta_{f' \mid m} f' \delta_{(r'_{q_2} m'_{q_2}/f', D/f)=1} \Big(\frac{f'}{r'_{q_2} m'_{q_2}}\Big)^{2it} \chi_1\Big(\frac{r'_{q_2} m'_{q_2}}{f'}\Big) \chi_2(f'') \\ \times \overline{\chi_2}(r'^{(q_2)} m'^{(q_2)}) \mathcal{G}_{\chi_2}(1;q_2) \sum_{\substack{\gamma > 0 \\ (\gamma, D/f)=1 \\ (\gamma, q_2)=1}} \frac{\chi_1\chi_2(\gamma)}{\gamma^{1+2it}} r(m;\gamma f''). \tag{6.39}$$

A computation using (6.39) and [1, Theorem 8.6] gives

$$\mathcal{S}(t,m,\chi_1,\chi_2;f) = \frac{\delta_{f'|m} f' \delta_{(r'_{q_2}m'_{q_2}/f',D/f)=1} \chi_2(f'') \overline{\chi_2}(r'^{(q_2)} m'^{(q_2)}) \mathcal{G}_{\chi_2}(1,q_2)}{L^{(D)}(\chi_1\chi_2, 1+2it)} \\ \times \mathcal{R}(t,m;\chi_1\chi_2,f'') \eta_{\chi_1\chi_2}(m) \left(\frac{f'}{r'_{q_2}m'_{q_2}}\right)^{2it} \chi_1\left(\frac{r'_{q_2}m'_{q_2}}{f'}\right), \qquad (6.40)$$

where the superscript (D) denotes that the local factors at the primes dividing D have been removed and

$$\mathcal{R}(t,m;\chi_1\chi_2,f'') := \sum_{\substack{\gamma|D^\infty \\ (\gamma,D/f)=1 \\ (\gamma,q_2)=1}} \frac{\chi_1\chi_2(\gamma)}{\gamma^{1+2it}} r(m;\gamma f'') \qquad (6.41)$$

and

$$\eta_{\chi_1\chi_2}(m) := \sum_{\substack{a|m \\ (a,D)=1}} \frac{\chi_1\chi_2(a)}{a^{2it}}.$$

We now consider (6.41). Since $\gamma \mid D^\infty$ and $(\gamma, D/f) = (\gamma, q_2) = 1$, we must have $\gamma \mid (f'')^\infty$. We apply [1, Theorem 8.7] and write (6.41) as

$$\mathcal{R}(t,m;\chi_1\chi_2,f'') := \prod_{\substack{p^\alpha || f'' \\ p^\alpha || D}} \sum_{\beta \geq 0} \frac{\chi_1\chi_2(p^\beta)}{p^{\beta(1+2it)}} r(p^{v_p(m)}; p^{\alpha+\beta}). \qquad (6.42)$$

We refine the factorisation in (6.35) to

$$f' = f'_r f'^{(r)}, \quad f'' = f''_r f''^{(r)}, \quad \text{where} \quad f'_r, f''_r \mid r^\infty, \quad (f'^{(r)} f''^{(r)}, r) = 1.$$

Since $(r, m') = 1$ and $f' \mid rm'$, it follows that $f'_r = (f', r)$. Recalling (6.40), we obtain

$$\mathcal{S}(t,rm',\chi_1,\chi_2;f) \\ = \Bigg(\delta_{f'_r|r} f'_r \delta_{(r'_{q_2}/f'_r,D/f)=1} \overline{\chi_2}(r'^{(q_2)}) \chi_2(f''_r) \mathcal{R}(t,r;\chi_1\chi_2,f''_r) \\ \times \eta_{\chi_1\chi_2}(r) \left(\frac{f'_r}{r'_{q_2}}\right)^{2it} \chi_1\left(\frac{r'_{q_2}}{f'_r}\right)\Bigg) \\ \times \Bigg(\delta_{f'^{(r)}|m'} f'^{(r)} \delta_{(m'_{q_2}/f'^{(r)},D/f)=1} \overline{\chi_2}(m'^{(q_2)}) \chi_2(f''^{(r)}) \mathcal{R}(t,m';\chi_1\chi_2,f''^{(r)}) \\ \times \eta_{\chi_1\chi_2}(m') \left(\frac{f'^{(r)}}{m'_{q_2}}\right)^{2it} \chi_1\left(\frac{m'_{q_2}}{f'^{(r)}}\right)\Bigg) \frac{\mathcal{G}_{\chi_2}(1,q_2)}{L^{(D)}(\chi_1\chi_2,1+2it)}. \qquad (6.43)$$

We now define

$$y_r := \left(\frac{f''}{(f'',2)}, r\right) \quad \text{and} \quad \tilde{y}_r := \prod_{\substack{p^\alpha || f''_r \\ \alpha \leq v_p(r)+1}} p^\alpha.$$

An explicit computation using (6.42) and [1, Theorems 8.6 and 8.7] shows that $R(y_r;\chi_1\chi_2,\widetilde{y}_r)\neq 0$. In particular,

$$\left|\frac{\mathcal{R}(t,r;\chi_1\chi_2,f_r'')}{\mathcal{R}(t,y_r;\chi_1\chi_2,\widetilde{y}_r)}\right|\leq\prod_{p|(r,D)}(v_p(r)+1)\frac{1+1/p}{\kappa(p)}\leq 4\tau(D)\tau(r),\tag{6.44}$$

where

$$\kappa(p):=\begin{cases}1/2 & \text{if } p=2,\\ 1-2/p & \text{if } p>2.\end{cases}$$

Observe that $(f_r',y_r)=1$ and $f_r',y_r\,|\,r$ and $f_r',y_r\,|\,D$. Thus $f_r'y_r\,|\,(r,D)$. Furthermore, $\widetilde{y}_r\,|\,f_r''$, $(f_r'y_r,m')=1$ and $\eta_{\chi_1\chi_2}(f_r'y_r)=1$. Using (6.43) and (6.42) we obtain

$$\begin{aligned}\mathcal{S}(t,rm';\chi_1,\chi_2,f)&=\delta_{f_r'|r}\delta_{(r_{q_2}'/f_r',D/f)=1}\overline{\chi_2}\left(\frac{r'^{(q_2)}}{y_r}\right)\chi_2\left(\frac{f_r''}{\widetilde{y}_r}\right)\frac{\mathcal{R}(t,r;\chi_1\chi_2,f_r'')}{\mathcal{R}(t,f_r'y_r;\chi_1\chi_2,\widetilde{y}_r)}\\&\quad\times\eta_{\chi_1\chi_2}(r)\left(\frac{f_r'}{r_{q_2}'}\right)^{2it}\chi_1\left(\frac{r_{q_2}'}{f_r'}\right)\mathcal{S}(t,f_r'y_rm';\chi_1,\chi_2,q_2f'\widetilde{y}_rf''^{(r)}).\end{aligned}\tag{6.45}$$

Combining (6.44), (6.45) and $|\eta_{\chi_1\chi_2}(r)|\leq\tau(r)$ we obtain

$$\begin{aligned}&\left|\sum_m a_m|m|^{it}\mathcal{S}(t,m;\chi_1,\chi_2,f)\right|^2\\&\leq 16\tau(D)^2\tau(r)^4\sum_{d|(r,Q)}\left|\sum_{\substack{m'\\(m',r)=1}}a_{rm'}|m'|^{it}\mathcal{S}(t,dm';\chi_1,\chi_2,q_2f'\widetilde{y}_rf''^{(r)})\right|^2,\end{aligned}\tag{6.46}$$

where the summation over d was introduced by positivity.

Given f and q_2 (and r as above) such that $q_2\,|\,f\,|\,D$, the integer $\mathcal{F}(q_2,f,r):=q_2f'\widetilde{y}_rf''^{(r)}$ satisfies

$$q_2\,|\,\mathcal{F}(q_2,f,r)\,|\,f\,|\,D.\tag{6.47}$$

Applying (6.46) to each summand of the right side of (6.34) yields

$$\begin{aligned}\sum_{\mathfrak{a}}\left|\sum_m a_m\sqrt{|m|}\rho_{\mathfrak{a}}(m,t)\right|^2&\leq 16\tau(D)^2\tau(r)^4\sum_{d|(r,D)}\frac{\pi}{|\Gamma(1/2+it)|^2}\\&\times\sum_{f|D}\sum_{q_2|f}\sum_{q_1|D/f}\frac{\widetilde{f}}{Df\phi(\widetilde{f})}\sideset{}{'}\sum_{\substack{\chi_1\ (\mathrm{mod}\,q_1)\\ \chi_2\ (\mathrm{mod}\,q_2)\\ \chi_1\overline{\chi_2}\sim\chi}}\left|\sum_{\substack{m'\\(m',r)=1}}a_{rm'}|m'|^{it}\mathcal{S}(t,dm';\chi_1,\chi_2,\mathcal{F}(q_2,f,r))\right|^2.\end{aligned}\tag{6.48}$$

We now relate the right side of (6.48) to a sum over all cusps. Observe that

$$\frac{\widetilde{f}}{Df\phi(\widetilde{f})}\leq\tau(D)\frac{\widetilde{\mathcal{F}(q_2,f,r)}}{D\mathcal{F}(q_2,f,r)\phi(\widetilde{\mathcal{F}(q_2,f,r)})},\tag{6.49}$$

so (6.48) becomes

$$\sum_{\mathfrak{a}}\Big|\sum_{m} a_m\sqrt{|m|}\rho_{\mathfrak{a}}(m,t)\Big|^2$$
$$\leq 16\tau(D)^3\tau(r)^4\sum_{d\,|(r,D)}\frac{\pi}{|\Gamma(1/2+it)|^2}\sum_{f\,|D}\sum_{q_2|f}\sum_{q_1|D/f}\frac{\widetilde{\mathcal{F}(q_2,f,r)}}{D\mathcal{F}(q_2,f,r)\phi(\widetilde{\mathcal{F}(q_2,f,r)})}$$
$$\times\sideset{}{'}\sum_{\substack{\chi_1\ (\mathrm{mod}\,q_1)\\ \chi_2\ (\mathrm{mod}\,q_2)\\ \chi_1\overline{\chi_2}\sim\chi}}\Big|\sum_{\substack{m'\\ (m',r)=1}} a_{rm'}|m'|^{it}\mathcal{S}(t,dm';\chi_1,\chi_2,\mathcal{F}(q_2,f,r))\Big|^2. \tag{6.50}$$

Recalling (6.47), by positivity we obtain

$$\sum_{\mathfrak{a}}\Big|\sum_{m} a_m\sqrt{|m|}\rho_{\mathfrak{a}}(m,t)\Big|^2$$
$$\leq 16\tau(D)^3\tau(r)^4\sum_{d\,|(r,D)}\frac{\pi}{|\Gamma(1/2+it)|^2}\sum_{f\,|D}\sum_{q_2|f}\sum_{\substack{\mathcal{F}\\ q_2|\mathcal{F}|f}}\sum_{q_1|D/f}\frac{\widetilde{\mathcal{F}}}{D\mathcal{F}\phi(\widetilde{\mathcal{F}})}$$
$$\times\sideset{}{'}\sum_{\substack{\chi_1\ (\mathrm{mod}\,q_1)\\ \chi_2\ (\mathrm{mod}\,q_2)\\ \chi_1\overline{\chi_2}\sim\chi}}\Big|\sum_{\substack{m'\\ (m',r)=1}} a_{rm'}|m'|^{it}\mathcal{S}(t,dm';\chi_1,\chi_2,\mathcal{F})\Big|^2, \tag{6.51}$$

where $\widetilde{\mathcal{F}}=(\mathcal{F},D/\mathcal{F})$. By positivity we may extend the sum over q_1 in (6.51) to all q_1 satisfying $q_1\mid D/\mathcal{F}$. Then interchanging the summations over f,q_2 and $\mathcal{F}$ we obtain

$$\sum_{\mathfrak{a}}\Big|\sum_{m} a_m\sqrt{|m|}\rho_{\mathfrak{a}}(m,t)\Big|^2 \leq 16\tau(D)^3\tau(r)^4\sum_{d\,|(r,D)}\frac{\pi}{|\Gamma(1/2+it)|^2}$$
$$\times\sum_{\mathcal{F}|D}\sum_{q_2|\mathcal{F}}\sum_{\mathcal{F}|f|D}\frac{\widetilde{\mathcal{F}}}{D\mathcal{F}\phi(\widetilde{\mathcal{F}})}\sum_{q_1|D/\mathcal{F}}\sideset{}{'}\sum_{\substack{\chi_1\ (\mathrm{mod}\,q_1)\\ \chi_2\ (\mathrm{mod}\,q_2)\\ \chi_1\overline{\chi_2}\sim\chi}}\Big|\sum_{\substack{m'\\ (m',r)=1}} a_{rm'}|m'|^{it}\mathcal{S}(t,dm';\chi_1,\chi_2,\mathcal{F})\Big|^2. \tag{6.52}$$

Now observe that all summands on the right side of (6.52) are completely independent of f. Thus we can remove the summation on f at the cost of an extra $\tau(D)$ factor. Then applying (6.34) to each summand of the d summation yields (6.29). ■

7. Spectral methods and shifted convolution sums

We choose a large parameter

$$C := N^{1000}, \tag{7.1}$$

and throughout this section make the general assumption that

$$h \asymp N \geq 20M.$$

For $\ell_1, \ell_2, h \in \mathbb{Z}_{\geq 1}$, recall the definitions (5.32) and (5.33):

$$\mathcal{D}(\ell_1, \ell_2, h, N, M) = \sum_{\ell_1 n - \ell_2 m = h} a(m)a(n)V_1\left(\frac{\ell_2 m}{M}\right)V_2\left(\frac{\ell_1 n}{N}\right),$$
$$\mathcal{S}(\ell_1, \ell_2, d, N, M) = \sum_{r \geq 1} \mathcal{D}(\ell_1, \ell_2, rd, N, M),$$

where $a(n)$ are the Fourier coefficients (normalised as in (1.3)) of $f \in \mathcal{S}_k(4)$ and $k := 1/2 + 2j$, $j \in \mathbb{N}$. Without loss of generality we can assume that

$$1 \leq \ell_1, \ell_2 \leq 2N, \tag{7.2}$$

otherwise $\mathcal{D}(\ell_1, \ell_2, h, N, M)$ vanishes trivially. Slightly more generally than in Section 5.2 we assume

$$V_{1,2} \text{ are supported in } [1,2] \text{ and } V_{1,2}^{(j)} \ll_{j,\varepsilon} C^{j\varepsilon}.$$

Proof of Proposition 5.2. The argument here is similar to the proof of [9, Proposition 8]. Our exposition will be sparse, sketching only the details unique to our situation. We refer the reader to [9, Sections 7 and 8] for more details. Let W be a smooth function with bounded derivatives such that

$$W(x) \equiv 1 \quad \text{for } 1 \leq x \leq 2 \quad \text{and} \quad \operatorname{supp}(W) \subseteq [1/2, 3].$$

After attaching a smooth redundant weight function W we obtain

$$\begin{aligned}\mathcal{D}(\ell_1, \ell_2, h, N, M) &= \sum_{\substack{m,n \\ \ell_1 n - \ell_2 m = h}} a(m)a(n)V_1\left(\frac{\ell_2 m}{M}\right)V_2\left(\frac{\ell_2 m + h}{N}\right)W\left(\frac{\ell_1 n - h}{M}\right) \\ &= \int_{-\infty}^{\infty} V_2^{\dagger}(z)e\left(\frac{zh}{N}\right)\mathcal{D}_z(\ell_1, \ell_2, h, N, M)\, dz,\end{aligned}$$

where $V_2^{\dagger}$ denotes the Fourier transform,

$$\mathcal{D}_z(\ell_1, \ell_2, h, N, M) := \sum_{\substack{m,n \\ \ell_1 n - \ell_2 m = h}} a(m)a(n)V_z\left(\frac{\ell_2 m}{M}\right)W\left(\frac{\ell_1 n - h}{M}\right),$$

and

$$V_z(x) := V_1(x)e(zxM/N).$$

We truncate the z-integral at $|z| \leq C^{\varepsilon}$ with a small error, say $O(C^{-100})$.

With the the notation as in [9, Lemma 19], we make the choice of parameters

$$Q := C \quad \text{and} \quad \delta := C^{-1}.$$

Let w_0 be a fixed smooth function with support in $[1,2]$ and let

$$w(c) = \begin{cases} w_0(c/C) & \text{if } 16\ell_1\ell_2 \mid c, \\ 0 & \text{otherwise.} \end{cases}$$

We see that

$$\Lambda \gg C^2(\ell_1\ell_2)^{-1-\varepsilon} \tag{7.3}$$

(cf. [9, Lemma 19]). Applying Jutila's circle method [9, Lemma 19], (4.4) and arguing as in [9, pp. 484–485] we obtain

$$\mathcal{D}_z(\ell_1, \ell_2, h, N, M) = \frac{1}{2\delta}\int_{-\delta}^{\delta} \mathcal{D}_{z,\eta}(\ell_1, \ell_2, h, N, M)\, d\eta + O(C^{-2/5}), \tag{7.4}$$

where

$$\begin{aligned} &D_{z,\eta}(\ell_1, \ell_2, h, N, M) \\ &\quad := \frac{1}{\Lambda} \sum_{16\ell_1\ell_2 | c} w_0\left(\frac{c}{C}\right) \sum_{\substack{d \;(\mathrm{mod}\, c)\\ (c,d)=1}} \sum_{m,n} a(m)a(n)e\left(\frac{d}{c}(\ell_1 n - \ell_2 m - h)\right) \\ &\qquad\qquad \times W_{\eta M}\left(\frac{\ell_1 n - h}{M}\right) V_{z,\eta M}\left(\frac{\ell_2 m}{M}\right), \end{aligned} \tag{7.5}$$

where

$$V_{z,\eta}(x) := V_z(x)e(-\eta x) = V_1(x)e\left(x\left(z\frac{M}{N} - \eta\right)\right) \quad \text{and} \quad W_\eta(x) := W(x)e(\eta x).$$

We stress that only Cauchy–Schwarz and (4.4) were used to obtain the error term of $C^{-2/5}$ in (7.4), not (4.9). Since $|\eta| \leq C^{-1} = N^{-1000}$ (in particular $\eta \ll M^{-1}$), the functions $V_{z,\eta M}$ and $W_{\eta M}$ are well behaved. In particular,

$$W^{(j)}_{\eta M} \ll 1 \quad \text{and} \quad V^{(j)}_{z,\eta M} \ll C^{j\varepsilon} \quad \text{uniformly in } |z| \ll C^{j\varepsilon}.$$

Observe that $V_{z,\eta M}$ and $W_{\eta M}$ have support in $[1, 2]$ and $[1/2, 3]$ respectively.

Here we will see that a Voronoi summation in m, n variables of (7.5) leads to a twist by a quadratic character depending on ℓ_1 and ℓ_2. Applying Lemma 4.1 to the m summation in (7.5) we obtain

$$\begin{aligned} &\sum_m a(m)e\left(-\frac{dm}{c/\ell_2}\right) V_{z,\eta M}\left(\frac{\ell_2 m}{M}\right) \\ &\qquad = \frac{M}{c}\sum_m a(m)e\left(\frac{\ell_2 m\bar{d}}{c}\right)\nu_\theta(\gamma_2)\mathring{V}_{z,\eta M}\left(\frac{\ell_2 m M}{c^2}\right), \end{aligned} \tag{7.6}$$

where $d\bar{d} \equiv 1 \pmod{c}$, $\tilde{d}$ is any integer such that $\tilde{d} \equiv d \pmod{c}$ and

$$\gamma_2 = \begin{pmatrix} -\tilde{d} & b_2 \\ c/\ell_2 & X_2 \end{pmatrix} \in \Gamma_0(c/\ell_2). \tag{7.7}$$

Applying Lemma 4.1 to the n summation in (7.5) we obtain

$$\begin{aligned} &\sum_n a(n)e\left(\frac{dn}{c/\ell_1}\right) W_{\eta M}\left(\frac{\ell_1 n - h}{M}\right) \\ &\qquad = \frac{M}{c}\sum_n a(n)\nu_\theta(\gamma_1)e\left(-\frac{\ell_1 n\bar{d}}{c}\right) W^*_{\eta M}\left(\frac{h\ell_1 n}{c^2}, \frac{M\ell_1 n}{c^2}\right), \end{aligned} \tag{7.8}$$

where

$$\gamma_1 = \begin{pmatrix} \widetilde{d} & b_1 \\ c/\ell_1 & X_1 \end{pmatrix} \in \Gamma_0(c/\ell_1), \tag{7.9}$$

and

$$W^*_{\eta M}(z,w) := 2\pi i^k \int_0^\infty W_{\eta M}(y) J_{k-1}(4\pi\sqrt{yw+z})\, dy.$$

It follows from (4.1), (4.2), (7.7), and (7.9) that

$$\nu_\theta(\gamma_1)\nu_\theta(\gamma_2) = \left(\frac{c/\ell_1}{X_1}\right)\overline{\varepsilon}_{X_1}\left(\frac{c/\ell_2}{X_2}\right)\overline{\varepsilon}_{X_2} = -i\left(\frac{4\ell_1\ell_2}{d}\right) \tag{7.10}$$

for d such that $(d,c) = 1$. Note that in (7.10) we have used the feature that the X_i are defined via the determinants of the γ_i for $i = 1, 2$. Let

$$\chi := \chi_{\ell_1\ell_2} = \left(\frac{4\ell_1\ell_2}{\bullet}\right).$$

Observe that χ is an even character modulo $4\ell_1\ell_2$ (in particular modulo $16\ell_1\ell_2$). Combining (7.5)–(7.10) and [9, Lemma 17] we obtain

$$\begin{aligned} D_{z,\eta}(\ell_1,\ell_2,h,N,M) &:= -\frac{M^2 i}{\Lambda C} \sum_{16\ell_1\ell_2 \mid c} w_1\left(\frac{c}{C}\right)\frac{1}{c}\sum_{m,n} a(m)a(n)K(\ell_1 n - \ell_2 m, h, c, \chi) \\ &\times \sum_{\pm} W_{\pm}\left(\frac{h\ell_1 n}{c^2}, \frac{M\ell_1 n}{c^2}\right) e\left(\pm 2\frac{\sqrt{h\ell_1 n}}{c}\right)\mathring{V}_{z,\eta M}\left(\frac{\ell_2 m}{c^2/M}\right) + O(C^{-A}), \end{aligned} \tag{7.11}$$

where

$$w_1(x) = w_0(x)/x,$$

and $W_{\pm}$ are as in [9, Lemma 17]. By [9, (6.15)] and the fact that $\mathring{V}_{z,\eta M}$ is a Schwartz class function (cf. (4.21)) we can restrict to

$$\ell_1 n \le \mathcal{N}_0 := C^{2+\varepsilon}N/M^2 \quad \text{and} \quad \ell_2 m \le \mathcal{M}_0 := C^{2+\varepsilon}/M, \tag{7.12}$$

with negligible error. For $\mathcal{N} \le \mathcal{N}_0$, $\mathcal{M} \le \mathcal{M}_0$ and $\mathcal{K} > 0$, we will restrict the right side of (7.11) to subsums

$$n \asymp \mathcal{N}, \quad m \asymp \mathcal{M}, \quad |\ell_1 n - \ell_2 m| \asymp \mathcal{K}.$$

Here $x \asymp X$ denotes $X \le x \le 2X$. The arising subsums are then split into three sums Σ_+, Σ_0 and Σ_- according to

$$\Sigma_+ : \ell_1 n > \ell_2 m, \quad \Sigma_0 : \ell_1 n = \ell_2 m, \quad \Sigma_- : \ell_1 n < \ell_2 m.$$

7.1. *Treatment of* Σ_0

Let $\chi^\star$ be the primitive character of conductor $C^\star_\chi \mid 4\ell_1\ell_2$ inducing $\chi\mathbf{1}_c$ modulo c. Then by [48, Lemma 3.1.3] we have

$$S(0,h,c,\chi) = \mathcal{G}_{\chi^\star}(1; C^\star_\chi) \sum_{\substack{d>0 \\ d \mid (h, c/C^\star_\chi)}} d\mu\left(\frac{c}{dC^\star_\chi}\right)\chi^\star\left(\frac{c}{dC^\star_\chi}\right)\overline{\chi^\star}\left(\frac{h}{d}\right).$$

Thus

$$|S(0,h,c,\chi)| \leq |C_\chi^\star|^{1/2}\tau(h)(h,c) \ll (\ell_1\ell_2)^{1/2}\tau(h)(h,c). \tag{7.13}$$

A trivial estimate using (7.13), Cauchy–Schwarz, (4.4), (7.3), and (7.12) yields

$$\begin{aligned}
\Sigma_0 &\ll \frac{M^2\tau(h)(\ell_1\ell_2)^{1/2}}{\Lambda C^{1-\varepsilon}} \sum_{C\leq c\leq 2C} \frac{(h,c)}{c} \sum_{\substack{\ell_1 n\asymp\mathcal{N},\, \ell_2 m\asymp\mathcal{M}\\ \ell_1 n=\ell_2 m}} |a(m)a(n)| \\
&\ll \frac{M^2\tau(h)^2(\ell_1\ell_2)^{1/2}}{\Lambda C^{1-\varepsilon}}\Big(\sum_{m\ll\mathcal{M}}|a(m)|^2\Big)^{1/2}\Big(\sum_{n\ll\mathcal{N}}|a(n)|^2\Big)^{1/2} \\
&\ll \frac{M^2\tau(h)^2(\ell_1\ell_2)^{1/2}(\mathcal{N}_0\mathcal{M}_0)^{1/2}}{\Lambda C^{1-\varepsilon}} \ll \frac{(\ell_1\ell_2)^{3/2+\varepsilon}(NM)^{1/2}}{C^{1-\varepsilon}} \ll C^{-1/2},
\end{aligned}$$

where the last equality follows from (7.1) and (7.2).

7.2. *Spectral treatment of* Σ_+

Now we consider

$$\Sigma_+ = -\frac{iM^2}{\Lambda C}\sum_{\substack{b>0\\ |b|\asymp\mathcal{K}}}\ \sum_{\substack{\ell_1 n-\ell_2 m=b\\ \ell_1 n\asymp\mathcal{N},\, \ell_2 m\asymp\mathcal{M}}} a(m)a(n) \sum_{16\ell_1\ell_2|c} \frac{K(b,h,c,\chi)}{c}\Phi\bigg(4\pi\frac{\sqrt{|b|h}}{c}\bigg), \tag{7.14}$$

where Φ is defined in [9, p. 487] (with a $\sum_\pm$ inserted into their definition). In view of the transforms occurring in the Kuznetsov formula, define

$$\mathcal{J}^+_{2it}(x) := \pi i\,\frac{J_{2it}(x)-J_{-2it}(x)}{\sinh(\pi t)} \quad\text{and}\quad \mathcal{J}^-_{2it}(x) := 4\cosh(\pi t)K_{2it}(x).$$

Let $\widetilde{\Phi}$, $\dot{\Phi}$, Ω be defined as in [9, p. 487]. Also define

$$\mathcal{T}_+ := C^\varepsilon\bigg(1+\Big(\frac{\mathcal{K}N}{C^2}\Big)^{1/4}+\Big(\frac{\mathcal{M}N}{C^2}\Big)^{1/2}\bigg) \quad\text{and}\quad \mathcal{T}_h := C^\varepsilon\bigg(1+\Big(\frac{\mathcal{K}N}{C^2}\Big)^{1/4}\bigg).$$

By the argument in [9, p. 487], the transforms $\widetilde{\Phi}(t)$ and $\dot{\Phi}(\ell)$ are negligible (cf. [9, Lemma 16]) unless

$$|t| \ll \mathcal{T}_+ \quad\text{and}\quad \ell \ll \mathcal{T}_h$$

respectively. Applying Lemma 6.2 (recalling that χ is an even Dirichlet character) to the summation over c in (7.14) and then truncating the appropriate summations and integrations using the above remarks we obtain

$$\Sigma_+ = \mathcal{H}_+(h) + \mathcal{M}_+(h) + \mathcal{E}_+(h) + O(C^{-A}),$$

where the terms on the right side correspond to the holomorphic, Maass and Eisenstein components of the spectrum. They are

$$\mathcal{H}_+(h) := -\frac{iM^2}{\Lambda C}\int_0^\infty \sum_{\substack{2\le \ell\le \mathcal{T}_h\\ \ell\equiv 0\,(\mathrm{mod}\,2)}} \sum_{g\in\mathcal{H}_\ell(16\ell_1\ell_2,\chi)} 4i^\ell\Gamma(\ell)J_{\ell-1}(x)\sqrt{h}\rho_g(h)$$
$$\times \sum_{\substack{b>0\\ |b|\asymp\mathcal{K}}} w_1\biggl(\frac{4\pi\sqrt{|b|h}}{Cx}\biggr)\sqrt{|b|}\rho_g(b)\gamma_+(b,h,x)\,\frac{dx}{x}, \tag{7.15}$$

$$\mathcal{M}_+(h) := -\frac{2iM^2}{\Lambda C}\int_0^\infty \sum_{\substack{g\in\mathcal{B}_0(16\ell_1\ell_2,\chi)\\ |t_g|\le\mathcal{T}_+}} \frac{\mathcal{J}^+_{2it_g}(x)}{\cosh(\pi t_g)}\sqrt{h}\rho_g(h)$$
$$\times \sum_{\substack{b>0\\ |b|\asymp\mathcal{K}}} w_1\biggl(\frac{4\pi\sqrt{|b|h}}{Cx}\biggr)\sqrt{|b|}\rho_g(b)\gamma_+(b,h,x)\,\frac{dx}{x}, \tag{7.16}$$

$$\mathcal{E}_+(h) := -\frac{2iM^2}{\Lambda C}\int_0^\infty \frac{1}{4\pi}\sum_{\mathfrak{a}}\int_{-\mathcal{T}_+}^{\mathcal{T}_+} \frac{\mathcal{J}^+_{2it}(x)}{\cosh(\pi t)}\sqrt{h}\rho_{\mathfrak{a}}(h,t)$$
$$\times \sum_{\substack{b>0\\ |b|\asymp\mathcal{K}}} w_1\biggl(\frac{4\pi\sqrt{|b|h}}{Cx}\biggr)\sqrt{|b|}\rho_{\mathfrak{a}}(b,t)dt\,\gamma_+(b,h,x)\,\frac{dx}{x}, \tag{7.17}$$

where

$$\gamma_+(b,h,x) := \sum_{\substack{\ell_1 n-\ell_2 m=b\\ \ell_1 n\asymp\mathcal{N},\,\ell_2 m\asymp\mathcal{M}}} a(m)a(n)\mathring{V}_{z,\eta M}\biggl(\frac{x^2\ell_2 mM}{(4\pi)^2|b|h}\biggr)$$
$$\times\sum_{\pm} W_\pm\biggl(\frac{x^2\ell_1 n}{(4\pi)^2|b|},\frac{x^2\ell_1 nM}{(4\pi)^2|b|h}\biggr)\theta_x^\pm\biggl(\frac{\ell_2 m}{|b|}\biggr),$$

and

$$\theta_x^\pm(y) := \exp\bigl(\pm ix\sqrt{1+y}\bigr)v\biggl(\frac{y}{\mathcal{M}/\mathcal{K}}\biggr),$$

and v a redundant smooth weight function of compact support on $[1/4,3]$ that is constantly 1 on $[1/2,2]$.

7.3. Spectral treatment of Σ_-

Now consider

$$\Sigma_- = -\frac{iM^2}{\Lambda C}\sum_{\substack{b<0\\ |b|\asymp\mathcal{K}}}\sum_{\substack{\ell_1 n-\ell_2 m=b\\ \ell_1 n\asymp\mathcal{N},\,\ell_2 m\asymp\mathcal{M}}} a(m)a(n)\sum_{16\ell_1\ell_2|c}\frac{K(b,h,c,\chi)}{c}\Phi\biggl(4\pi\frac{\sqrt{|b|h}}{c}\biggr).$$

Define

$$\mathcal{T}_- := C^\varepsilon\biggl(1+\Bigl(\frac{\mathcal{M}N}{C^2}\Bigr)^{1/2}\biggr).$$

Applying an argument similar to [9, pp. 488–489] using Lemma 6.2 we obtain

$$\Sigma_- = \mathcal{M}_-(h) + \mathcal{E}_-(h) + O(C^{-A}),$$

where

$$\begin{aligned} \mathcal{M}_-(h) = -\frac{2iM^2}{\Lambda C} \int_0^\infty \int_{\sigma - iC^\varepsilon \mathcal{T}_-}^{\sigma + iC^\varepsilon \mathcal{T}_-} \sum_{\substack{g \in \mathcal{B}_0(16\ell_1\ell_2,\chi) \\ |t_g| \leq \mathcal{T}_-}} \widehat{\mathcal{J}}_{2it_g}^{-}(s) \frac{\sqrt{h}\rho_g(h)}{\cosh(\pi t_g)} w_1\left(\frac{\sqrt{h}}{Cx}\right) \\ \times \sum_{\substack{b<0 \\ |b| \asymp \mathcal{K}}} (4\pi\sqrt{|b|}x)^{-s} \sqrt{|b|}\rho_g(b)\gamma_-(b,h,x) \frac{ds}{2\pi i} \frac{dx}{x}, \end{aligned}$$

$$\begin{aligned} \gamma_-(b,h,x) = \sum_{\substack{\ell_1 n - \ell_2 m = b \\ \ell_1 n \asymp \mathcal{N}, \ell_2 m \asymp \mathcal{M}}} a(m)a(n)\mathring{V}_{z,\eta M}\left(\frac{x^2\ell_2 m M}{h}\right) \\ \times \sum_{\pm} W_\pm\left(x^2\ell_1 n, \frac{x^2\ell_1 n M}{h}\right) e\left(\pm 2x\sqrt{\ell_1 n}\right), \end{aligned}$$

and $\sigma = 7/32 + \varepsilon$. An analogous formula holds for the Eisenstein contribution $\mathcal{E}_-(h)$.

7.4. *Summary of setup*

By the above discussion it suffices to estimate the right side of

$$\begin{aligned} &\mathcal{D}(\ell_1, \ell_2, h, N, M) \\ &= \frac{1}{2\delta} \int_{-\delta}^{\delta} \int_{-C^\varepsilon}^{C^\varepsilon} V_2^\dagger(z) e\left(\frac{zh}{N}\right) \sum_{\mathcal{N} \leq \mathcal{N}_0} \sum_{\substack{\mathcal{M} \leq \mathcal{M}_0 \\ \mathcal{M} \leq \mathcal{N}}} \sum_{\substack{\mathcal{K} \leq \mathcal{N}_0 \\ \mathcal{K} \leq \mathcal{N}}} (\mathcal{H}_+(h) + \mathcal{M}_+(h) + \mathcal{E}_+(h))\, dz\, d\eta \\ &+ \frac{1}{2\delta} \int_{-\delta}^{\delta} \int_{-C^\varepsilon}^{C^\varepsilon} V_2^\dagger(z) e\left(\frac{zh}{N}\right) \sum_{\mathcal{N} \leq \mathcal{M}_0} \sum_{\mathcal{M} \leq \mathcal{M}_0} \sum_{\mathcal{K} \leq \mathcal{M}_0} (\mathcal{M}_-(h) + \mathcal{E}_-(h))\, dz\, d\eta + O(C^{-1/3}), \end{aligned} \tag{7.18}$$

where $\mathcal{N}$, $\mathcal{M}$ and $\mathcal{K}$ run over numbers ≥ 1 of the form $\mathcal{N}_0 2^{-\nu}$ or $\mathcal{M}_0 2^{-\nu}$, $\nu \in \mathbb{N}$.

7.5. *Shifted convolution sums on average*

We now turn our attention to the averaged convolution sum

$$\mathcal{S}(\ell_1, \ell_2, d, N, M) = \sum_r \mathcal{D}(\ell_1, \ell_2, rd, N, M).$$

Let

$$\beta := \operatorname{lcm}(16, \ell_1, \ell_2, d) \quad \text{and} \quad B := \{n \in \mathbb{N} : p \mid n \Rightarrow p \mid \beta \text{ for all primes } p\}.$$

Observe that $\mathcal{D}(\ell_1, \ell_2, rd, N, M)$ vanishes unless $r \asymp N/d$. Recall (7.18). To begin, consider the decomposition

$$\sum_{r \asymp N/d} e\left(\frac{zrd}{N}\right)\mathcal{H}_+(rd) = \sum_{\substack{r_2 \ll N/d \\ r_2 \in B}} \sum_{\substack{r_1 \asymp N/(dr_2) \\ (r_1,\beta)=1}} e\left(\frac{zr_1r_2d}{N}\right)\mathcal{H}_+(r_1r_2d),$$

where $\mathcal{H}_+$ was defined in (7.15). Observe that the range of integration for x in (7.15) is

$$x \asymp X_+ := \sqrt{\mathcal{K}N}/C.$$

We follow verbatim the Mellin inversion argument of [9, pp. 490–491] that separates the variables scattered throughout the various weight functions. In that argument,

$$S := C^{\varepsilon}\left(1 + \frac{X_+\mathcal{M}}{\sqrt{\mathcal{K}\mathcal{N}}}\right).$$

This yields

$$\sum_{r_1 \asymp N/(dr_2)} e\left(\frac{zr_1r_2d}{N}\right)\mathcal{H}_+(r_1r_2d) \ll \frac{C^{\varepsilon}M^2}{\Lambda C}\frac{\mathcal{K}^{1/4}}{X_+^{1/2}\mathcal{N}^{1/4}} S(\Xi^{\mathcal{H}}_{1,+}\Xi^{\mathcal{H}}_{2,+})^{1/2}, \tag{7.19}$$

where

$$\Xi^{\mathcal{H}}_{1,+} := \max_{|u_4| \le C^{\varepsilon}} \sum_{\substack{0<\ell\le\mathcal{T}_h \\ \ell\equiv 0 \ (\mathrm{mod}\ 2)}} \Gamma(\ell)$$
$$\times \sum_{g \in \mathcal{H}_\ell(16\ell_1\ell_2,\chi)} \left| \sum_{\substack{r_1 \asymp N/(dr_2) \\ (r_1,\beta)=1}} e\left(\frac{zr_1r_2d}{N}\right) r_1^{2\varepsilon+iu_4}\sqrt{r_1r_2d}\,\rho_g(r_1r_2d)\right|^2,$$

and

$$\Xi^{\mathcal{H}}_{2,+} = \max_{\substack{|u_2|\le C^{\varepsilon} \\ |u_1|,|u_3|\le S \\ x \asymp X_+}} \sum_{\substack{0<\ell\le\mathcal{T}_h \\ \ell\equiv 0 \ (\mathrm{mod}\ 2)}} |J_{\ell-1}(x)|^2\Gamma(\ell) \sum_{g\in\mathcal{H}_\ell(4\ell_1\ell_2,\chi)} \left|\sum_{|b|\asymp\mathcal{K}} \sqrt{|b|}\rho_g(b)\gamma^*(b)\right|^2,$$

with

$$\gamma^*(b) = \left(\frac{|b|}{\mathcal{K}}\right)^{1/4+\varepsilon+iu_3} \sum_{\substack{\ell_1 n - \ell_2 m = b \\ \ell_1 n \asymp \mathcal{N}, \ell_2 m \asymp \mathcal{M}}} \left(\frac{\ell_1 n}{\mathcal{N}}\right)^{-1/4+iu_2}\left(\frac{\ell_2 m}{\mathcal{M}}\right)^{-\varepsilon+iu_1} a(m)a(n).$$

The same analysis works mutatis mutandis for the Eisenstein and Maass spectrum, giving analogous expressions for $\Xi^{\mathcal{E}}_{i,+}$ and $\Xi^{\mathcal{M}}_{i,+}$ for $i = 1, 2$. Note that the breakdown of the Archimedean weight functions in the Cauchy–Schwarz inequality in x and g is analogous: $\Xi^{\mathcal{M}}_{1,+}$ has a factor of $1/\cosh(\pi t_g)$ and $\Xi^{\mathcal{M}}_{2,+}$ has a factor $|\mathcal{J}^+_{2it_g}(x)|^2/\cosh(\pi t_g)$.

We now bound the various $\Xi^{\star}_{i,+}$ for $i = 1, 2$. Applying the argument at the beginning of [9, p. 492] using (4.4) and (4.5) we obtain

$$\sum_{b} |\gamma^{*}(b)|^2 \ll C^{\varepsilon} \frac{\mathcal{N}\mathcal{M}}{\ell_1 \ell_2}, \tag{7.20}$$

uniformly in u_1, u_2 and u_3. Also by [9, p. 492] we have

$$J_{\ell-1}(x) \ll C^{\varepsilon} x^{-1/2}, \tag{7.21}$$

uniformly for all $x > 0$ and $1 \leq \ell \leq \mathcal{T}_h$. Thus by Lemma 6.4, (7.20) and (7.21) we have

$$\Xi^{\mathcal{H}}_{2,+} \ll \frac{C^{\varepsilon}}{X_{+}} \left(\mathcal{T}_h^2 + \frac{\mathcal{K}}{\ell_1 \ell_2} \right) \frac{\mathcal{N}\mathcal{M}}{\ell_1 \ell_2}. \tag{7.22}$$

Similarly,

$$|\Xi^{\mathcal{E}}_{2,+}| + |\Xi^{\mathcal{M}}_{2,+}| \ll \frac{C^{\varepsilon}}{X_{+}} \left(\mathcal{T}_{+}^2 + \frac{\mathcal{K}}{\ell_1 \ell_2} \right) \frac{\mathcal{N}\mathcal{M}}{\ell_1 \ell_2}. \tag{7.23}$$

By (6.15) and the fact that integral weight Hecke cusp forms satisfy Deligne's bound (see also Remark 6.1) we have

$$\begin{aligned} \Xi^{\mathcal{H}}_{1,+} \ll \max_{|u_4| \leq C^{\varepsilon}} C^{\varepsilon} \sum_{\delta | 16 \ell_1 \ell_2} \sum_{\substack{2 \leq \ell \leq \mathcal{T}_h \\ \ell \equiv 0 \, (\mathrm{mod}\, 2)}} \Gamma(\ell) \\ \times \sum_{g \in \mathcal{H}_{\ell}(16 \ell_1 \ell_2, \chi)} \Bigg| \sum_{\substack{r_1 \asymp N/(dr_2) \\ (r_1, \beta) = 1}} \alpha(r_1) \sqrt{r_1 \delta} \rho_g(r_1 \delta) \Bigg|^2, \end{aligned} \tag{7.24}$$

where

$$\alpha(r_1) = \alpha_{r_2 d, u_4}(r_1) = e\left(\frac{z r_1 r_2 d}{N} \right) r_1^{2\varepsilon + i u_4}.$$

Applying Lemma 6.4 to the right side of (7.24) yields

$$\Xi^{\mathcal{H}}_{1,+} \ll C^{\varepsilon} \sum_{\delta | 16 \ell_1 \ell_2} \left(\mathcal{T}_h^2 + \frac{N\delta}{dr_2 \ell_1 \ell_2} \right) \frac{N}{dr_2} \ll C^{\varepsilon} \left(\mathcal{T}_h^2 + \frac{N}{dr_2} \right) \frac{N}{dr_2}. \tag{7.25}$$

Observe that all cusps of $\Gamma_0(16\ell_1\ell_2)$ are singular relative to χ by Lemma 6.6 (i.e. the conductor of χ is of the form Z or $4Z$ where Z is odd and squarefree since it is a quadratic character). Thus we can apply Lemmas 6.7 and 6.4 to similarly obtain

$$\Xi^{\mathcal{E}}_{1,+} \ll C^{\varepsilon} \left(\mathcal{T}_{+}^2 + \frac{N}{dr_2} \right) \frac{N}{dr_2}. \tag{7.26}$$

Applying Theorem 6.1 we obtain

$$\begin{aligned} \Xi^{\mathcal{M}}_{1,+} &= \max_{|u_4| \leq C^{\varepsilon}} \sum_{\substack{|t_g| \leq \mathcal{T}_{+} \\ g \in \mathcal{B}_0(16 \ell_1 \ell_2, \chi)}} \frac{1}{\cosh(\pi t_g)} \Bigg| \sum_{\substack{r_1 \asymp N/(dr_2) \\ (r_1, \beta) = 1}} \alpha(r_1) \sqrt{r_1 r_2 d} \rho_g(r_1 r_2 d) \Bigg|^2 \\ &\ll C^{\varepsilon} (\ell_1 \ell_2, r_2 d) \left(\mathcal{T}_{+} + \frac{(r_2 d)^{1/2}}{(\ell_1 \ell_2)^{1/2}} \right) \left(\mathcal{T}_{+} + \frac{N}{dr_2 (\ell_1 \ell_2)^{1/2}} \right) \frac{N}{dr_2}. \end{aligned} \tag{7.27}$$

Combining (7.22), (7.23), (7.25), (7.26), (7.27), we obtain

$$(|\Xi^{\mathcal{H}}_{1,+}|+|\Xi^{\mathcal{M}}_{1,+}|+|\Xi^{\mathcal{E}}_{1,+}|)(|\Xi^{\mathcal{H}}_{2,+}|+|\Xi^{\mathcal{M}}_{2,+}|+|\Xi^{\mathcal{E}}_{2,+}|) \ll \frac{N}{dr_2}\,\frac{(\ell_1\ell_2, r_2 d)\mathcal{N}\mathcal{M}}{\ell_1\ell_2}\,\frac{C^{\varepsilon}}{X_+} \times \biggl(\biggl(\mathcal{T}_+ + \frac{(r_2 d)^{1/2}}{(\ell_1\ell_2)^{1/2}}\biggr)\biggl(\mathcal{T}_+ + \frac{N}{dr_2(\ell_1\ell_2)^{1/2}}\biggr) + \frac{N}{dr_2}\biggr)\biggl(\mathcal{T}_+^2 + \frac{\mathcal{K}}{\ell_1\ell_2}\biggr). \tag{7.28}$$

Inserting (7.28) into (7.19) (recalling that (7.19) has both $\mathcal{M}$ and $\mathcal{E}$ analogues), the brute force computation in [9, pp. 493–494] guarantees that

$$\sum_{\substack{r_1 \asymp N/(dr_2)\\ (r_1,\beta)=1}} e\biggl(\frac{zr_1r_2d}{N}\biggr)(\mathcal{H}_+(r_1r_2d)+\mathcal{M}_+(r_1r_2d)+\mathcal{E}_+(r_1r_2d)) \ll C^{\varepsilon}(\ell_1\ell_2, d)^{1/2}\biggl(\frac{N}{d^{1/2}} + \frac{N^{5/4}M^{1/4}}{d(\ell_1\ell_2)^{1/4}} + \frac{N^{3/4}M^{1/4}}{d^{1/4}} + \frac{NM^{1/2}}{d^{3/4}(\ell_1\ell_2)^{1/2}} + \frac{NM^{1/2}}{d}\biggr). \tag{7.29}$$

One can follow the argument of [9, pp. 494–495] appealing to Lemmas 6.4 and 6.7, as well as (6.18), whenever their principal character analogues are used, to conclude that

$$\sum_{\substack{r_1 \asymp N/(dr_2)\\ (r_1,\beta)=1}} e\biggl(\frac{zr_1r_2d}{N}\biggr)(\mathcal{M}_-(r_1r_2d)+\mathcal{E}_-(r_1r_2d)) \ll C^{\varepsilon}d^{\theta}(\ell_1\ell_2, d)^{1/2}\biggl(\frac{M^{1/4}N^{3/4}}{d^{1/2}} + \frac{M^{3/4}N^{3/4}}{d}\biggr). \tag{7.30}$$

Summing (7.29) and (7.30) over $r_2 \in B$ using Rankin's trick, and using $\theta \le 1/4$, we obtain Proposition 5.2. ■

8. Critical range and αn^2 modulo 1

Proof of Theorem 1.2. Recall that

$$1 \le M \le p/2, \quad p^{1/2-1/10} \le N \le p^{1/2+1/10}, \tag{8.1}$$

and consider

$$\sum_{N \le n_1, n_2 \le 2N} \biggl|\sum_{M \le m \le 2M} S(m, cn_1, p)\overline{S(m, cn_2, p)}\biggr|$$

for $c \in \mathbb{F}_p^{\times}$. The estimates we obtain will not depend on c. Substituting (4.24) we obtain

$$p \sum_{N \le n_1, n_2 \le 2N} \biggl|\sum_{M \le m \le 2M} \sum_{\substack{u,v \ (\mathrm{mod}\ p)\\ u^2 \equiv cmn_1 \ (\mathrm{mod}\ p)\\ v^2 \equiv cmn_2 \ (\mathrm{mod}\ p)}} e\biggl(\frac{2(u+v)}{p}\biggr)\biggr|. \tag{8.2}$$

Let $R := R(M, N, p)$ denote the multiple summation in (8.2), excluding the factor of p. We write

$$R := R_1 + R_{-1}$$

where R_i restricts the summation variables in the definition of R to

$$\left(\frac{n_1}{p}\right) = \left(\frac{n_2}{p}\right) = \left(\frac{cm}{p}\right) = i. \tag{8.3}$$

We first consider R_1. For $\ell \in \mathbb{F}_p$, define

$$A_{\ell,c} := \sum_{M \le m \le 2M} \sum_{t^2 \equiv cm \ (\mathrm{mod}\ p)} e\left(\frac{2t\ell}{p}\right),$$

and

$$\mathbb{S}_\ell := \{(u, v) \in (\mathbb{F}_p^\times)^2 : (u^2, v^2) \ (\mathrm{mod}\ p) \in [N, 2N] \times [N, 2N] \text{ and } u + v \equiv \ell \ (\mathrm{mod}\ p)\}. \tag{8.4}$$

Applying the triangle inequality we obtain

$$\begin{aligned} R_1 &= \frac{1}{2} \sum_{N \le n_1, n_2 \le 2N} \left| \sum_{M \le m \le 2M} \sum_{\substack{t \ (\mathrm{mod}\ p) \\ t^2 \equiv cm}} \sum_{\substack{u,v \ (\mathrm{mod}\ p) \\ u^2 \equiv n_1 \ (\mathrm{mod}\ p) \\ v^2 \equiv n_2 \ (\mathrm{mod}\ p)}} e\left(\frac{2t(u+v)}{p}\right) \right| \\ &\le \frac{1}{2} \sum_{\ell \ (\mathrm{mod}\ p)} |A_{\ell,c}| \, |\mathbb{S}_\ell|. \end{aligned} \tag{8.5}$$

Observe that the contribution from $\ell \equiv 0 \ (\mathrm{mod}\ p)$ to the right hand side of (8.5) is

$$|A_{0,c}| \, |\mathbb{S}_0| \ll MN. \tag{8.6}$$

Thus it suffices to consider the right side of (8.5) for $\ell \not\equiv 0 \ (\mathrm{mod}\ p)$. For each $\ell \not\equiv 0 \ (\mathrm{mod}\ p)$, we argue that the elements in $\mathbb{S}_\ell$ satisfy a strong diophantine property. Recall that $(u, v) \in \mathbb{S}_\ell$ means that

$$u + v \equiv \ell \ (\mathrm{mod}\ p),$$

and $u^2, v^2 \ (\mathrm{mod}\ p)$ lie in the interval $[N, 2N]$. After an algebraic manipulation we see that $(u, v) \in \mathbb{S}_\ell$ must satisfy

$$\overline{\ell}^2 (u^2 - v^2)^2 + \ell^2 \equiv 2(u^2 + v^2) \ (\mathrm{mod}\ p). \tag{8.7}$$

We set

$$\alpha_\ell := \overline{\ell}^2 / p \in \mathbb{Q}/\mathbb{Z} \quad \text{and} \quad \beta_\ell := \ell^2 / p \in \mathbb{Q}/\mathbb{Z}.$$

Thus (8.7) implies

$$\| \alpha_\ell (u^2 - v^2)^2 + \beta_\ell \| \le 8N/p, \tag{8.8}$$

where $\| \bullet \|$ denotes the distance to the closest integer. Therefore the pairs $(u, v) \in \mathbb{S}_\ell$ produce elements of the sequence $\{\alpha_\ell n^2\}_{0 \le n \le N}$ modulo 1 and lie in a cluster around $-\beta_\ell$.

It is now sufficient to bound the right side of (8.5). We fix a tuple

$$\boldsymbol{\delta} := (\delta_j) \in (0,1)^6$$

to be chosen later. Let

$$\mathcal{L}(c,\boldsymbol{\delta}) := \{\ell \in \mathbb{F}_p^\times : |A_{\ell,c}| \geq Mp^{-\delta_1}\}. \tag{8.9}$$

Thus (8.6) and (8.9) imply

$$R_1 \ll MN + MN^2 p^{-\delta_1} + \sum_{\ell \in \mathcal{L}(c,\boldsymbol{\delta})} |A_{\ell,c}|\,|\mathbb{S}_\ell|. \tag{8.10}$$

As explained in Section 2, the distribution of the $\alpha_\ell n^2$ (which governs the size of $\mathbb{S}_\ell$) is sensitive to the convergents of the continued fraction expansion of α_ℓ. Thus we will consider a partition of $\mathcal{L}(c,\boldsymbol{\delta})$,

$$\mathcal{L}(c,\boldsymbol{\delta}) = \mathcal{H}_1(c,\boldsymbol{\delta}) \cup \mathcal{H}_2(c,\boldsymbol{\delta}) \cup \mathcal{H}_3(c,\boldsymbol{\delta}), \tag{8.11}$$

defined below. By convention, all convergents in the following definitions and arguments are denoted by irreducible fractions. The sets in (8.11) are given by

$$\begin{aligned}\mathcal{H}_1(c,\boldsymbol{\delta}) := \{\ell \in \mathcal{L}(c,\boldsymbol{\delta}) : \text{for all } 1 \leq h \leq p^{\delta_5},&\\ h\alpha_\ell \text{ has a convergent } a_{\ell,h}/b_{\ell,h} \text{ such that } b_{\ell,h} \in [p^{\delta_2}, p^{\delta_3}]\},& \qquad (8.12)\end{aligned}$$

$$\begin{aligned}\mathcal{H}_2(c,\boldsymbol{\delta}) := \{\ell \in \mathcal{L}(c,\boldsymbol{\delta}) : \ \text{there exists } 1 \leq h_\ell \leq p^{\delta_5}&\\ \text{for which } h_\ell\alpha_\ell \text{ has no convergent } a/b \text{ with } b \in [p^{\delta_2}, p^{\delta_4}]\},& \qquad (8.13)\end{aligned}$$

$$\mathcal{H}_3(c,\boldsymbol{\delta}) := \mathcal{L}(c,\boldsymbol{\delta}) \setminus (\mathcal{H}_1(c,\boldsymbol{\delta}) \cup \mathcal{H}_2(c,\boldsymbol{\delta})). \tag{8.14}$$

Our argument will require $\boldsymbol{\delta} \in (0,1)^6$ to satisfy some constraints. We record them here for convenience:

$$\delta_2 < \delta_3 < \delta_4, \tag{8.15}$$

$$\delta_6 < 1/5, \tag{8.16}$$

$$16Np^{\delta_2+\delta_5} < p/2. \tag{8.17}$$

Remark 8.1. The constraint (8.15) implies that the sets (8.12) and (8.13) are well-defined. The constraints (8.15)–(8.17) will be used in the treatment of $\mathcal{H}_3(c,\boldsymbol{\delta})$ in Section 8.3.

Also note that the elements in $\mathcal{L}(c,\boldsymbol{\delta})$ depend on c by definition. However, the criterion for an element $\ell \in \mathcal{L}(c,\boldsymbol{\delta})$ to belong to $\mathcal{H}_j(c,\boldsymbol{\delta})$ is independent of c.

First we bound $|\mathcal{L}(c,\boldsymbol{\delta})|$ via a second moment estimate of the $A_{\ell,c}$. This will be useful in some of the following arguments. We have

$$\sum_{\ell \,(\mathrm{mod}\, p)} |A_{\ell,c}|^2 = \sum_{M \leq m,m' \leq 2M} \sum_{\substack{t^2 \equiv cm \,(\mathrm{mod}\, p)\\ t'^2 \equiv cm' \,(\mathrm{mod}\, p)}} \sum_{\ell \,(\mathrm{mod}\, p)} e\left(\frac{2\ell(t-t')}{p}\right) \ll pM. \tag{8.18}$$

Using (8.18) we obtain

$$|\mathcal{L}(c,\boldsymbol{\delta})| \leq \sum_{\ell \,(\mathrm{mod}\, p)} \left(\frac{|A_{\ell,c}|p^{\delta_1}}{M}\right)^2 \ll \frac{p^{1+2\delta_1}}{M}, \tag{8.19}$$

uniformly in c.

8.1. *Treatment of $\mathcal{H}_1(c,\boldsymbol{\delta})$*

We remark that for $\ell \in \mathcal{H}_1(c,\boldsymbol{\delta})$, the sequence

$$\mathcal{N}_\ell := \{\alpha_\ell n^2\}_{1\leq n\leq N} \tag{8.20}$$

has small discrepancy. Thus to bound its contribution to (8.10), we obtain an upper bound for $|\mathbb{S}_\ell|$. Let $\mathcal{D}(\mathcal{N}_\ell)$ denote the discrepancy of $\mathcal{N}_\ell$. The number of $n \in [1,N]$ such that

$$\|\alpha_\ell n^2 + \beta_\ell\| \leq 8N/p \tag{8.21}$$

is

$$\ll N^2/p + N\mathcal{D}(\mathcal{N}_\ell). \tag{8.22}$$

In order to bound $\mathcal{D}(\mathcal{N}_\ell)$ we consider, for each $1 \leq h \leq p^{\delta_5}$,

$$E_{\ell,h} := \sum_{1\leq n\leq N} e(h\alpha_\ell n^2).$$

By definition, the continued fraction expansion of $h\alpha_\ell$ has a convergent

$$a_{\ell,h}/b_{\ell,h} \quad \text{with} \quad b_{\ell,h} \in [p^{\delta_2}, p^{\delta_3}].$$

Moreover,

$$\left|h\alpha_\ell - \frac{a_{\ell,h}}{b_{\ell,h}}\right| \leq \frac{1}{b_{\ell,h}^2}.$$

Applying Weyl's inequality [64, Lemma 2.4] we obtain

$$E_{\ell,h} \ll N^{1+\varepsilon}\left(p^{-\delta_2} + N^{-1} + \frac{p^{\delta_3}}{N^2}\right)^{1/2} \ll N^{\varepsilon}(Np^{-\delta_2/2} + N^{1/2} + p^{\delta_3/2}), \tag{8.23}$$

which is uniform in h, ℓ and c. Next, applying the Erdős–Turán inequality [41, (2.42), p. 114] and (8.23) we obtain, uniformly in ℓ and c,

$$\begin{aligned} N\mathcal{D}(\mathcal{N}_\ell) &\ll Np^{-\delta_5} + \sum_{1\leq h\leq p^{\delta_5}} |E_{\ell,h}|/h \\ &\ll (Np)^{\varepsilon}(Np^{-\delta_5} + Np^{-\delta_2/2} + N^{1/2} + p^{\delta_3/2}). \end{aligned} \tag{8.24}$$

Thus the right side (8.22) is

$$\ll (Np)^{\varepsilon}(N^2/p + Np^{-\delta_5} + Np^{-\delta_2/2} + N^{1/2} + p^{\delta_3/2}), \tag{8.25}$$

uniformly in ℓ and c. Observe that for each $\ell \in \mathcal{L}(c, \boldsymbol{\delta})$ (and in particular $\ell \in \mathcal{H}_1(c, \boldsymbol{\delta})$) and $n \in [1, N]$ satisfying (8.21), there is at most one element $(u, v) \in \mathbb{S}_\ell$ such that (cf. (8.8))

$$u^2 - v^2 \equiv \pm n \pmod{p}.$$

The same statement holds when $n = 0$. Observing that $|A_{\ell,c}| \ll M$ and using (8.19) and (8.25) we obtain

$$\sum_{\ell \in \mathcal{H}_1(c,\boldsymbol{\delta})} |A_{\ell,c}| \, |\mathbb{S}_\ell| \\ \ll (Np)^\varepsilon (p^{2\delta_1} N^2 + Np^{1+2\delta_1-\delta_5} + Np^{1+2\delta_1-\delta_2/2} + N^{1/2} p^{1+2\delta_1} + p^{1+2\delta_1+\delta_3/2}). \tag{8.26}$$

8.2. *Treatment of $\mathcal{H}_2(c, \boldsymbol{\delta})$*

We draw on the intuition that membership of $\mathcal{H}_2(c, \boldsymbol{\delta})$ is a rare event. Thus we give an upper bound for $|\mathcal{H}_2(c, \boldsymbol{\delta})|$ that is stronger than that implied by (8.19).

For each $\ell \in \mathcal{H}_2(c, \boldsymbol{\delta})$, fix $1 \le h_\ell \le p^{\delta_5}$ such that $h_\ell \alpha_\ell$ has no convergent

$$a/b \quad \text{with} \quad b \in [p^{\delta_2}, p^{\delta_4}].$$

Let a_ℓ / b_ℓ be the convergent to $h_\ell \alpha_\ell$ with $b_\ell \in [1, p^{\delta_2})$ maximal and let a_ℓ^* / b_ℓ^* denote the next convergent. Both such convergents exist. Then we must have $b_\ell^* > p^{\delta_4}$ and we know that

$$\left| h_\ell \alpha_\ell - \frac{a_\ell}{b_\ell} \right| \le \frac{1}{b_\ell b_\ell^*} < \frac{1}{b_\ell p^{\delta_4}}.$$

Therefore

$$|b_\ell h_\ell \overline{\ell}^2 - p a_\ell| < p^{1-\delta_4}. \tag{8.27}$$

Let $\mu_\ell \in \mathbb{Z} \cap (-p/2, p/2]$ be such that

$$\mu_\ell \equiv b_\ell h_\ell \overline{\ell}^2 \pmod{p}. \tag{8.28}$$

Thus (8.27) guarantees

$$|\mu_\ell| < p^{1-\delta_4}.$$

Conversely, consider the congruence

$$\mu \equiv b h \overline{\ell}^2 \pmod{p}. \tag{8.29}$$

Any given

$$\mu \in (-p^{1-\delta_4}, p^{1-\delta_4}), \quad b \in [1, p^{\delta_2}), \quad h \in [1, p^{\delta_5})$$

determine ℓ in (8.29) up to sign. Thus

$$|\mathcal{H}_2(c, \boldsymbol{\delta})| \ll p^{1-\delta_4+\delta_2+\delta_5},$$

uniformly in c. Since $|A_{\ell,c}| \ll M$ and $|\mathbb{S}_\ell| \ll N$ we obtain

$$\sum_{\ell \in \mathcal{H}_2(c,\boldsymbol{\delta})} |A_{\ell,c}| \, |\mathbb{S}_\ell| \ll MNp^{1-\delta_4+\delta_2+\delta_5}. \tag{8.30}$$

8.3. Treatment of $\mathcal{H}_3(c, \boldsymbol{\delta})$

Recall (8.15). We unpack the definition of $\mathcal{H}_3(c, \boldsymbol{\delta})$. Let $\ell \in \mathcal{H}_3(c, \boldsymbol{\delta})$. Since $\ell \notin \mathcal{H}_1(c, \boldsymbol{\delta})$, there exists $1 \leq h_\ell \leq p^{\delta_5}$ such that $h_\ell \alpha_\ell$ does not have a convergent

$$a/b \quad \text{with} \quad b \in [p^{\delta_2}, p^{\delta_3}].$$

For each ℓ, fix such a choice h_ℓ. Furthermore, since $\ell \notin \mathcal{H}_2(c, \boldsymbol{\delta})$, $h_\ell \alpha_\ell$ is guaranteed to have a convergent

$$a_\ell^* / b_\ell^* \quad \text{such that} \quad b_\ell^* \in (p^{\delta_3}, p^{\delta_4}].$$

Take such a convergent with b_ℓ^* minimal.

For each ℓ, denote

$$\mathbb{V}_\ell := \{0 \leq n \leq N : \|\alpha_\ell n^2 + \beta_\ell\| \leq 8N/p\}.$$

For each $p^{\delta_3} \leq U \leq p^{\delta_4}$ and $0 \leq V \leq N$ we define

$$\mathcal{E}_c(U, V, \boldsymbol{\delta}) := \{\ell \in \mathcal{H}_3(c, \boldsymbol{\delta}) : b_\ell^* \in [U, 2U] \text{ and } |\mathbb{V}_\ell| \in [V, 2V]\}.$$

Uniformly in U, V, c and $\boldsymbol{\delta}$ (satisfying (8.15)–(8.17)) we have, by (8.19),

$$|\mathcal{E}_c(U, V, \boldsymbol{\delta})| \ll p^{1+2\delta_1}/M. \tag{8.31}$$

We prove that the contribution to (8.10) from all

$$\ell \in \bigcup_{0 \leq V \leq Np^{-\delta_6}} \mathcal{E}_c(U, V, \boldsymbol{\delta}) =: \mathcal{B}_{\delta_6} \tag{8.32}$$

is small. Applying (8.31), $|A_{\ell,c}| \ll M$ and the remark following (8.25) we see that the contribution to (8.10) from $\ell \in \mathcal{B}_{\delta_6}$ is

$$\ll Np^{1+2\delta_1-\delta_6}. \tag{8.33}$$

We now consider the case when V is large. Observe that $\mathcal{H}_3(c, \boldsymbol{\delta}) \setminus \mathcal{B}_{\delta_6}$ can be covered by $O(\log^2 p)$ sets $\mathcal{E}(U, V, \boldsymbol{\delta})$ with

$$p^{\delta_3} \leq U \leq p^{\delta_4}, \quad Np^{-\delta_6} \leq V \leq N. \tag{8.34}$$

From (8.33) and the remark following (8.25) we obtain

$$\sum_{\ell \in \mathcal{H}_3(c, \boldsymbol{\delta})} |A_{\ell,c}| |\mathbb{S}_\ell| \ll p^{1+2\delta_1-\delta_6} N + M \log^2 p \max_{p^{\delta_3} \leq U \leq p^{\delta_4}} \max_{Np^{-\delta_6} \leq V \leq N} V \cdot |\mathcal{E}_c(U, V, \boldsymbol{\delta})|. \tag{8.35}$$

Thus we need to bound $V \cdot |\mathcal{E}_c(U, V, \boldsymbol{\delta})|$. For each $\ell \in \mathcal{E}_c(U, V, \boldsymbol{\delta})$, we now construct an algebraic set $\mathfrak{C}_\ell \subseteq \mathbb{F}_p^3$ with restricted variables. Arrange the numbers $n_{\ell,j} \in \mathbb{V}_\ell$ as

$$0 \leq n_{\ell,1} < n_{\ell,2} < \cdots < n_{\ell,|\mathbb{V}_\ell|} \leq N. \tag{8.36}$$

The average consecutive gap between these numbers is

$$\frac{N}{|\mathbb{V}_\ell|} \asymp \frac{N}{V} \ll p^{\delta_6}.$$

More than $|\mathbb{V}_\ell|/2$ consecutive gaps are less than or equal to $2N/|\mathbb{V}_\ell|$. By the pigeonhole principle there exists an integer $1 \leq d_\ell \leq 2N/|\mathbb{V}_\ell|$ that is repeated as a consecutive gap at least $|\mathbb{V}_\ell|^2/(4N) \gg 1$ times (note that (8.1), (8.16), and (8.34) guarantee that $|\mathbb{V}_\ell|^2/(4N) \gg 1$). Thus we define

$$\begin{aligned}\mathfrak{C}_\ell := \{(n, A, B) \in [1, N] \times [-8N, 8N]^2 : \overline{\ell}^2 n^2 + \ell^2 \equiv A \pmod{p} \\ \text{and } \overline{\ell}^2(n + d_\ell)^2 + \ell^2 \equiv B \pmod{p)}\}. \end{aligned} \tag{8.37}$$

We form

$$\mathfrak{U}_c(U, V, \boldsymbol{\delta}) := \bigcup_{\ell \in \mathcal{E}_c(U,V,\boldsymbol{\delta})} \{\ell\} \times \mathfrak{C}_\ell \subseteq \mathbb{F}_p^4,$$

and study this object now.

The above discussion implies the pointwise bound $|\mathfrak{C}_\ell| \gg V^2/N$, so

$$|\mathfrak{U}_c(U, V, \boldsymbol{\delta})| \gg V^2 |\mathcal{E}_c(U, V, \boldsymbol{\delta})|/N. \tag{8.38}$$

Thus it suffices to establish an upper bound for $|\mathfrak{U}_c(U, V, \boldsymbol{\delta})|$. We count the number of $Q := (\ell; n, A, B) \in \mathfrak{U}_c(U, V, \boldsymbol{\delta})$ with $A \equiv B \pmod{p}$ and $A \not\equiv B \pmod{p}$ separately.

Given $\ell \in \mathcal{E}_c(U, V, \boldsymbol{\delta})$ and $A \equiv B \pmod{p}$, an algebraic manipulation determines at most one possible Q. Thus (8.31) implies that there are

$$\ll p^{1+2\delta_1}/M \tag{8.39}$$

such Q.

The rest of the argument treats the case $A \not\equiv B \pmod{p}$. Recall the constraint (8.17). Let

$$\mathfrak{T}_c(U, V, \boldsymbol{\delta}) := \left\{ g + pr \in \mathbb{Z} : |r| \leq \frac{36N^2}{UV} + 1,\ |g| \leq 16 p^{\delta_2 + \delta_5} N \text{ and } g \neq 0 \right\}$$

be a set containing a union of short arithmetic progressions. We will construct a map

$$t_\bullet : Q \in \mathfrak{U}_c(U, V, \boldsymbol{\delta})\ (A \not\equiv B \pmod{p}) \mapsto t_Q \in \mathfrak{T}_c(U, V, \boldsymbol{\delta}),$$

whose fibers have size $O(p^\varepsilon)$ for any fixed $\varepsilon > 0$. These facts will imply

$$|\mathfrak{U}_c(U, V, \boldsymbol{\delta})| \ll p^\varepsilon |\mathfrak{T}_c(U, V, \boldsymbol{\delta})| + p^{1+2\delta_1}/M. \tag{8.40}$$

Starting with $Q \in \mathfrak{U}_c(U, V, \boldsymbol{\delta})$, subtracting the congruences in (8.37) yields

$$\overline{\ell}^2(2n d_\ell + d_\ell^2) \equiv B - A \not\equiv 0 \pmod{p}. \tag{8.41}$$

Recall that for each $\ell \in \mathcal{H}_3(c, \boldsymbol{\delta})$, we fixed a choice $h_\ell \in [1, p^{\delta_5}]$ such that $h_\ell \alpha_\ell = h_\ell \overline{\ell}^2/p$ has no convergent

$$a/b \quad \text{with} \quad b \in [p^{\delta_2}, p^{\delta_3}],$$

and has a convergent

$$a_\ell^*/b_\ell^* \quad \text{such that} \quad b_\ell^* \in (p^{\delta_3}, p^{\delta_4}],$$

with b_ℓ^* minimal. Moreover, $\ell \in \mathcal{E}_c(U, V, \boldsymbol{\delta})$ restricts $b_\ell^* \in [U, 2U]$. Let a_ℓ/b_ℓ denote the convergent to $h_\ell \alpha_\ell$ with $b_\ell \in [1, p^{\delta_2})$ maximal. Thus a_ℓ/b_ℓ and a_ℓ^*/b_ℓ^* are consecutive convergents. Let $\mu_\ell \in \mathbb{Z} \cap (-p/2, p/2]$ be such that

$$\mu_\ell \equiv b_\ell h_\ell \overline{\ell}^2 \pmod{p}.$$

Note that $\mu_\ell \not\equiv 0 \pmod{p}$. By a similar argument to the one in Section 8.2 we have

$$|\mu_\ell| \leq p/U.$$

Multiplying (8.41) by $b_\ell h_\ell$ we obtain

$$\mu_\ell(2nd_\ell + d_\ell^2) \equiv b_\ell h_\ell (B - A) \pmod{p}. \tag{8.42}$$

Writing (8.42) as an equation of integers we have

$$\mu_\ell(2nd_\ell + d_\ell^2) = pr + b_\ell h_\ell (B - A) \quad \text{for some } r \in \mathbb{Z}.$$

Observe that

$$0 < |b_\ell h_\ell (B - A)| \leq 16 p^{\delta_2 + \delta_5} N \quad \text{and} \quad |\mu_\ell(2nd_\ell + d_\ell^2)| \leq \frac{36 N^2 p}{UV} + 1.$$

Thus

$$t_Q := \mu_\ell(2nd_\ell + d_\ell^2) \in \mathfrak{T}_c(U, V, \boldsymbol{\delta}).$$

Suppose we are given $t = pr + g \in \mathfrak{T}_c(U, V, \boldsymbol{\delta})$. Since $t \neq 0$ (by (8.17)), $g \neq 0$ and $A \not\equiv B \pmod{p}$, the number of tuples

$$(\mu, n, d, b, h, B - A) \in \mathbb{Z} \times [0, N] \times [1, 2N/V] \times [1, p^{\delta_2}] \times [1, p^{\delta_5}] \times [-16N, 16N] \tag{8.43}$$

satisfying

$$t = \mu d(2n + d) \quad \text{and} \quad g = bh(B - A)$$

is at most $O(p^\varepsilon)$ by divisor considerations. A tuple in (8.43) then determines two values of ℓ mod p using

$$\mu \equiv bh\overline{\ell}^2 \pmod{p}.$$

Thus there are at most $O(p^\varepsilon)$ valid (n, d, ℓ) for a given t. Each 3-tuple together with the equations defining $\mathfrak{C}_\ell$ in (8.37) determines at most one pair $(A, B) \in [-8N, 8N]^2$. Thus there are at most $O(p^\varepsilon)$ quadruples $(\ell; n, A, B) \in \mathfrak{U}_c(U, V, \boldsymbol{\delta})$ such that $t_Q = t$ and (8.40) holds.

Combining (8.38) and (8.40) gives

$$\max_{p^{\delta_3}\le U\le p^{\delta_4}}\ \max_{Np^{-\delta_6}\le V\le N} V\cdot|\mathcal{E}_c(U,V,\boldsymbol{\delta})| \\ \ll (Np)^{\varepsilon}(p^{1+2\delta_1+\delta_6}/M + N^2p^{\delta_2+\delta_5+2\delta_6-\delta_3} + Np^{\delta_2+\delta_5+\delta_6}).$$

Inserting this into (8.35) we obtain

$$\sum_{\ell\in\mathcal{H}_3(c,\boldsymbol{\delta})}|A_{\ell,c}|\,|\mathbb{S}_\ell| \\ \ll (Np)^{\varepsilon}(p^{1+2\delta_1-\delta_6}N + p^{1+2\delta_1+\delta_6} + N^2Mp^{\delta_2+\delta_5+2\delta_6-\delta_3} + MNp^{\delta_2+\delta_5+\delta_6}). \tag{8.44}$$

Inserting (8.26), (8.30) and (8.44) into (8.10) we obtain

$$\begin{aligned} R_1 \ll (Np)^{\varepsilon}\big(&MN + MN^2p^{-\delta_1} + p^{2\delta_1}N^2 + Np^{1+2\delta_1-\delta_5} + Np^{1+2\delta_1-\delta_2/2} \\ &+ N^{1/2}p^{1+2\delta_1} + p^{1+2\delta_1+\delta_3/2} + MNp^{1-\delta_4+\delta_2+\delta_5} + Np^{1+2\delta_1-\delta_6} \\ &+ p^{1+2\delta_1+\delta_6} + MN^2p^{\delta_2+\delta_5+2\delta_6-\delta_3} + MNp^{\delta_2+\delta_5+\delta_6}\big). \end{aligned} \tag{8.45}$$

The same argument can be applied to bound R_{-1} by the right hand side of (8.45). One fixes a non-zero non-quadratic residue j modulo p and sees that (8.3) is equivalent to

$$\left(\frac{\bar{j}n_1}{p}\right) = \left(\frac{\bar{j}n_2}{p}\right) = \left(\frac{jcm}{p}\right) = 1.$$

Thus the analogue of (8.5) is

$$\begin{aligned} R_{-1} &= \frac{1}{2}\sum_{N\le n_1,n_2\le 2N}\Bigg|\sum_{M\le m\le 2M}\ \sum_{\substack{t\ (\mathrm{mod}\ p)\\ t^2\equiv cjm}}\ \sum_{\substack{u,v\ (\mathrm{mod}\ p)\\ u^2\equiv \bar{j}n_1\ (\mathrm{mod}\ p)\\ v^2\equiv \bar{j}n_2\ (\mathrm{mod}\ p)}} e\left(\frac{2t(u+v)}{p}\right)\Bigg| \\ &\le \frac{1}{2}\sum_{\ell\ (\mathrm{mod}\ p)}|A_{\ell,cj}|\,|\mathbb{S}_{\ell,j}|, \end{aligned}$$

where

$$\mathbb{S}_{\ell,j} := \{(u,v)\in(\mathbb{F}_p^{\times})^2 : (ju^2, jv^2)\ (\mathrm{mod}\ p)\in[N,2N]\times[N,2N] \text{ and } u+v\equiv\ell\ (\mathrm{mod}\ p)\}.$$

Repeating the algebraic manipulation with the linear congruence in the definition of $\mathbb{S}_{\ell,j}$ we obtain

$$\|\alpha_{j,\ell}(u^2-v^2)^2 + \beta_{j,\ell}\| \le 8N/p,$$

where

$$\alpha_{\ell,j} := j\bar{\ell}^2/p \in \mathbb{Q}/\mathbb{Z} \quad\text{and}\quad \beta_{\ell,j} := \frac{j\ell^2}{p} \in \mathbb{Q}/\mathbb{Z}.$$

Then (8.10) becomes

$$R_{-1} \ll MN + MN^2p^{-\delta_1} + \sum_{\ell\in\mathcal{L}(cj,\boldsymbol{\delta})}|A_{\ell,cj}|\,|\mathbb{S}_{\ell,j}| \tag{8.46}$$

and we consider the partition

$$\mathcal{L}(cj,\boldsymbol{\delta}) = \mathcal{H}_1(cj,\boldsymbol{\delta}) \cup \mathcal{H}_2(cj,\boldsymbol{\delta}) \cup \mathcal{H}_3(cj,\boldsymbol{\delta}),$$

where one replaces α_ℓ (resp. β_ℓ) by $\alpha_{\ell,j}$ (resp. $\beta_{\ell,j}$) in the definitions of the $\mathcal{H}_i(c,\boldsymbol{\delta})$ occurring in (8.12)–(8.14). One can then repeat the arguments in Sections 8.1–8.3 making the necessary modifications.

For $p^{1/2-1/10} \leq N \leq p^{1/2+1/10}$, we see that

$$\boldsymbol{\delta} := \left(\tfrac{11}{288}, \tfrac{11}{48}, \tfrac{25}{36}, \tfrac{407}{432}, \tfrac{11}{96}, \tfrac{11}{96}\right)$$

satisfies (8.15)–(8.17), and is sufficient to obtain Theorem 1.2 (after multiplication by p, cf. (8.2)). For aesthetic reasons we take a larger estimate (i.e. all denominators multiples of 27).

Remark 8.2. Observe that the above argument can be modified so that the estimate in Theorem 1.2 holds when n_1, n_2 and m are each restricted to fixed congruence classes modulo 4.

This completes the proof of Theorem 1.2. ■

Acknowledgments. We thank Matthew Young and Djordje Milićević for their comments on the manuscript and Valentin Blomer and E. Mehmet Kıral for helpful correspondences. We also thank the referee for their meticulous comments on the manuscript.

References

[1] Apostol, T. M.: Introduction to analytic number theory. Undergrad. Texts Math., Springer, New York-Heidelberg (1976) Zbl 0335.10001 MR 0434929

[2] Baker, R. C.: Kloosterman sums and Maass forms. Vol. I. Kendrick Press, Heber City, UT (2003) Zbl 1028.11034 MR 1970058

[3] Blomer, V.: Non-vanishing of class group L-functions at the central point. Ann. Inst. Fourier (Grenoble) **54**, 831–847 (2004) Zbl 1063.11040 MR 2111013

[4] Blomer, V.: Subconvexity for a double Dirichlet series. Compos. Math. **147**, 355–374 (2011) Zbl 1228.11140 MR 2776608

[5] Blomer, V., Fouvry, É., Kowalski, E., Michel, P., Milićević, D.: On moments of twisted L-functions. Amer. J. Math. **139**, 707–768 (2017) Zbl 1476.11081 MR 3650231

[6] Blomer, V., Fouvry, É., Kowalski, E., Michel, P., Milićević, D., Sawin, W.: The second moment theory of families of L-functions—the case of twisted Hecke L-functions. Mem. Amer. Math. Soc. **282**, no. 1394, v+148 pp. (2023) Zbl 1519.11001 MR 4539366

[7] Blomer, V., Harcos, G., Michel, P.: A Burgess-like subconvex bound for twisted L-functions. Forum Math. **19**, 61–105 (2007) Zbl 1168.11014 MR 2296066

[8] Blomer, V., Milićević, D.: p-adic analytic twists and strong subconvexity. Ann. Sci. École Norm. Sup. (4) **48**, 561–605 (2015) Zbl 1401.11095 MR 3377053

[9] Blomer, V., Milićević, D.: The second moment of twisted modular L-functions. Geom. Funct. Anal. **25**, 453–516 (2015) Zbl 1400.11097 MR 3334233

[10] Casati, G., Guarneri, I., Izraĭlev, F. M.: Statistical properties of the quasi-energy spectrum of a simple integrable system. Phys. Lett. A **124**, 263–266 (1987) Zbl 1037.82525 MR 0910875

[11] Chinta, G.: Analytic ranks of elliptic curves over cyclotomic fields. J. Reine Angew. Math. **544**, 13–24 (2002) Zbl 1028.11040 MR 1887886

[12] Conrey, J. B., Ghosh, A.: Remarks on the generalized Lindelöf hypothesis. Funct. Approx. Comment. Math. **36**, 71–78 (2006) Zbl 1196.11121 MR 2296639

[13] 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

[14] Drappeau, S.: Sums of Kloosterman sums in arithmetic progressions, and the error term in the dispersion method. Proc. London Math. Soc. (3) **114**, 684–732 (2017) Zbl 1392.11059 MR 3653244

[15] Duke, W., Friedlander, J., Iwaniec, H.: Bounds for automorphic L-functions. Invent. Math. **112**, 1–8 (1993) Zbl 0765.11038 MR 1207474

[16] Duke, W., Friedlander, J. B., Iwaniec, H.: The subconvexity problem for Artin L-functions. Invent. Math. **149**, 489–577 (2002) Zbl 1056.11072 MR 1923476

[17] Duke, W., Iwaniec, H.: Bilinear forms in the Fourier coefficients of half-integral weight cusp forms and sums over primes. Math. Ann. **286**, 783–802 (1990) Zbl 0671.10021 MR 1045402

[18] Dunn, A., Kerr, B., Shparlinski, I. E., Zaharescu, A.: Bilinear forms in Weyl sums for modular square roots and applications. Adv. Math. **375**, article no. 107369, 58 pp. (2020) Zbl 1469.11270 MR 4137069

[19] Gao, P., Khan, R., Ricotta, G.: The second moment of Dirichlet twists of Hecke L-functions. Acta Arith. **140**, 57–65 (2009) Zbl 1242.11035 MR 2557853

[20] Good, A.: The square mean of Dirichlet series associated with cusp forms. Mathematika **29**, 278–295 (1983) (1982) Zbl 0497.10016 MR 0696884

[21] 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

[22] 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

[23] Hoffstein, J., Kontorovich, A.: The first non-vanishing quadratic twist of an automorphic L-series. arXiv:1008.0839 (2010)

[24] Hoffstein, J., Lockhart, P.: Coefficients of Maass forms and the Siegel zero. Ann. of Math. (2) **140**, 161–181 (1994) Zbl 0814.11032 MR 1289494

[25] Ivić, A., Motohashi, Y.: On the fourth power moment of the Riemann zeta-function. J. Number Theory **51**, 16–45 (1995) Zbl 0824.11048 MR 1321722

[26] Iwaniec, H.: Fourier coefficients of modular forms of half-integral weight. Invent. Math. **87**, 385–401 (1987) Zbl 0606.10017 MR 0870736

[27] Iwaniec, H.: Topics in classical automorphic forms. Grad. Stud. Math. 17, American Mathematical Society, Providence, RI (1997) Zbl 0905.11023 MR 1474964

[28] Iwaniec, H., Kowalski, E.: Analytic number theory. Amer. Math. Soc. Colloq. Publ. 53, American Mathematical Society, Providence, RI (2004) Zbl 1059.11001 MR 2061214

[29] Iwaniec, H., Luo, W., Sarnak, P.: Low lying zeros of families of L-functions. Inst. Hautes Études Sci. Publ. Math. **91**, 55–131 (2000) Zbl 1012.11041 MR 1828743

[30] Iwaniec, H., Sarnak, P.: The non-vanishing of central values of automorphic L-functions and Landau–Siegel zeros. Israel J. Math. **120**, 155–177 (2000) Zbl 0992.11037 MR 1815374

[31] Kerr, B., Shkredov, I. D., Shparlinski, I. E., Zaharescu, A.: Energy bounds for modular roots and their applications. J. Inst. Math. Jussieu (online, 2024)

[32] Kim, H. H.: Functoriality for the exterior square of GL_4 and the symmetric fourth of GL_2. J. Amer. Math. Soc. **16**, 139–183 (2003) Zbl 1018.11024 MR 1937203

[33] Kıral, E. M.: Subconvexity for half integral weight L-functions. Math. Z. **281**, 689–722 (2015) Zbl 1330.11057 MR 3421637

[34] Kıral, E. M., Young, M. P.: Kloosterman sums and Fourier coefficients of Eisenstein series. Ramanujan J. **49**, 391–409 (2019) Zbl 1436.11102 MR 3949076

[35] Knightly, A., Li, C.: Kuznetsov's trace formula and the Hecke eigenvalues of Maass forms. Mem. Amer. Math. Soc. **224**, no. 1055, vi+132 pp. (2013) Zbl 1314.11038 MR 3099744

[36] Kohnen, W.: Modular forms of half-integral weight on $\Gamma_0(4)$. Math. Ann. **248**, 249–266 (1980) Zbl 0416.10023 MR 0575942

[37] Kohnen, W., Zagier, D.: Values of L-series of modular forms at the center of the critical strip. Invent. Math. **64**, 175–198 (1981) Zbl 0468.10015 MR 0629468

[38] Kowalski, E., Michel, P., Sawin, W.: Bilinear forms with Kloosterman sums and applications. Ann. of Math. (2) **186**, 413–500 (2017) Zbl 1441.11194 MR 3702671

[39] Kowalski, E., Michel, P., Sawin, W.: Stratification and averaging for exponential sums: bilinear forms with generalized Kloosterman sums. Ann. Scuola Norm. Sup. Pisa Cl. Sci. (5) **21**, 1453–1530 (2020) Zbl 07373253 MR 4288639

[40] Kowalski, E., Michel, P., VanderKam, J.: Mollification of the fourth moment of automorphic L-functions and arithmetic applications. Invent. Math. **142**, 95–151 (2000) Zbl 1054.11026 MR 1784797

[41] Kuipers, L., Niederreiter, H.: Uniform distribution of sequences. Pure Appl. Math., Wiley-Interscience, New York (1974) Zbl 0281.10001 MR 0419394

[42] Lang, S.: Complex analysis. Addison-Wesley, Reading, MA (1977) Zbl 0366.30001 MR 0477000

[43] Lang, S.: Elliptic functions. 2nd ed., Grad. Texts in Math. 112, Springer, New York (1987) Zbl 0615.14018 MR 0890960

[44] Lester, S., Radziwiłł, M.: Quantum unique ergodicity for half-integral weight automorphic forms. Duke Math. J. **169**, 279–351 (2020) Zbl 1441.11107 MR 4057145

[45] Li, X.: Moments of quadratic twists of modular L-functions. Invent. Math. **237**, 697–733 Zbl 07887296 (2024) MR 4768632

[46] Hoffstein, J., Lee, M.: Second moments and simultaneous non-vanishing of GL_2-automorphic L-series. arXiv:1308.5980 (2013)

[47] Michel, P.: The subconvexity problem for Rankin–Selberg L-functions and equidistribution of Heegner points. Ann. of Math. (2) **160**, 185–236 (2004) Zbl 1068.11033 MR 2119720

[48] Miyake, T.: Modular forms. Springer Monogr. Math., Springer, Berlin (2006) Zbl 1159.11014 MR 2194815

[49] Motohashi, Y.: Spectral theory of the Riemann zeta-function. Cambridge Tracts in Math. 127, Cambridge University Press, Cambridge (1997) Zbl 0878.11001 MR 1489236

[50] NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.18 of 2018-03-27, visited on August 25, 2024

[51] Proskurin, N. V.: On general Kloosterman sums. Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. **302**, 107–134, 200 (2003) (in Russian) Zbl 1140.11340 MR 2023036

[52] Rudnick, Z., Sarnak, P.: The pair correlation function of fractional parts of polynomials. Comm. Math. Phys. **194**, 61–70 (1998) Zbl 0919.11052 MR 1628282

[53] Rudnick, Z., Sarnak, P., Zaharescu, A.: The distribution of spacings between the fractional parts of $n^2\alpha$. Invent. Math. **145**, 37–57 (2001) Zbl 1006.11041 MR 1839285

[54] Salié, H.: Über die Kloostermanschen Summen $S(u, v; q)$. Math. Z. **34**, 91–109 (1932) Zbl 57.0211.01 MR 1545243

[55] Sarnak, P.: Some applications of modular forms. Cambridge Tracts in Math. 99, Cambridge University Press, Cambridge (1990) Zbl 0721.11015 MR 1102679

[56] Schulze-Pillot, R., Yenirce, A.: Petersson products of bases of spaces of cusp forms and estimates for Fourier coefficients. Int. J. Number Theory **14**, 2277–2290 (2018) Zbl 1422.11099 MR 3846405

[57] Selberg, A.: Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. J. Indian Math. Soc. (N.S.) **20**, 47–87 (1956) Zbl 0072.08201 MR 0088511

[58] Shimura, G.: On modular forms of half integral weight. Ann. of Math. (2) **97**, 440–481 (1973) Zbl 0266.10022 MR 0332663

[59] Shkredov, I. D., Shparlinski, I. E., Zaharescu, A.: Bilinear forms with modular square roots and twisted second moments of half integral weight Dirichlet series. Int. Math. Res. Notices **2022**, 17431–17474 Zbl 1516.11039 MR 4514447

[60] Shkredov, I. D., Shparlinski, I. E., Zaharescu, A.: On the distribution of modular square roots of primes. Math. Z. **306**, article no. 43, 17 pp. (2024) Zbl 07807524 MR 4703506

[61] Soundararajan, K., Young, M. P.: The second moment of quadratic twists of modular L-functions. J. Eur. Math. Soc. **12**, 1097–1116 (2010) Zbl 1213.11165 MR 2677611

[62] Stefanicki, T.: Non-vanishing of L-functions attached to automorphic representations of GL(2) over $\mathbf{Q}$. J. Reine Angew. Math. **474**, 1–24 (1996) Zbl 0848.11023 MR 1390690

[63] Stevens, G.: Λ-adic modular forms of half-integral weight and a Λ-adic Shintani lifting. In: Arithmetic geometry (Tempe, AZ, 1993), Contemp. Math. 174, American Mathematical Society, Providence, RI, 129–151 (1994) Zbl 0869.11042 MR 1299739

[64] Vaughan, R. C.: The Hardy–Littlewood method. 2nd ed., Cambridge Tracts in Math. 125, Cambridge University Press, Cambridge (1997) Zbl 0868.11046 MR 1435742

[65] Young, M. P.: The fourth moment of Dirichlet L-functions. Ann. of Math. (2) **173**, 1–50 (2011) Zbl 1296.11112 MR 2753598

[66] Young, M. P.: Explicit calculations with Eisenstein series. J. Number Theory **199**, 1–48 (2019) Zbl 1454.11083 MR 3926186

[67] Zaharescu, A.: Correlation of fractional parts of $n^2\alpha$. Forum Math. **15**, 1–21 (2003) Zbl 1029.11039 MR 1957276

[68] Zavorotnyĭ, N. I.: On the fourth moment of the Riemann zeta function. In: Automorphic functions and number theory, Parts I, II, Akad. Nauk SSSR, Dal'nevostochn. Otdel., Vladivostok, 69–124a, 254 (1989) (in Russian) Zbl 0711.11028 MR 1683661