THE SECOND MOMENT OF TWISTED MODULAR L-FUNCTIONS

VALENTIN BLOMER AND DJORDJE MILIĆEVIĆ

Abstract. We prove an asymptotic formula with a power saving error term for the (pure or mixed) second moment

$$\sum_{\chi \bmod q}^{*} L(1/2, f_1 \otimes \chi)\overline{L(1/2, f_2 \otimes \chi)}$$

of central values of L-functions of any two (possibly equal) fixed cusp forms f_1, f_2 twisted by all primitive characters modulo q, valid for all sufficiently factorable q including 99.9% of all admissible moduli. The two key ingredients are a careful spectral analysis of a potentially highly unbalanced shifted convolution problem in Hecke eigenvalues and power-saving bounds for sums of products of Kloosterman sums where the length of the sum is below the square-root threshold of the modulus. Applications are given to simultaneous non-vanishing and lower bounds on higher moments of twisted L-functions.

1. Introduction

1.1. **The main result.** Most L-functions come in families, and often their moments encode some deep properties about the family. The complexity of an L-function is measured by its analytic conductor $\mathcal{C}$ (which is typically essentially constant within a family $\mathcal{F}$), and a measure for the complexity of a moment calculation is the ratio $r = \log \mathcal{C}/\log|\mathcal{F}|$ (the family may not be discrete in which case an obvious modification is necessary). The edge of current technology where one can hope to obtain an asymptotic formula with a *power saving error term* is $r = 4$. The stock of asymptotic formulas of this kind, however, is very small, and experience has shown that quite often in the case $r = 4$ the current methods of analytic number theory fail "by an ε"; if they don't, then typically some very deep input is required.

The most classical example is the fourth moment of the Riemann zeta-function, where one has the asymptotic formula

$$\int_0^T |\zeta(1/2 + it)|^4 dt = TP_4(\log T) + \mathrm{O}(T^{2/3+\varepsilon}) \tag{1.1}$$

for a certain polynomial P_4 (see [Za, IM, Mot]), which is one of the prime applications of the Kuznetsov formula. This formula can be seen as the second moment of the L-function attached to a (derivative of an) Eisenstein series, and the corresponding cuspidal analogue, proved by Good [Go], states that

$$\int_0^T |L(1/2 + it, f)|^2 dt = TP_1(\log T) + \mathrm{O}(T^{2/3+\varepsilon}) \tag{1.2}$$

for a certain polynomial P_1 depending on the holomorphic Hecke cusp form f. In addition to spectral analysis of automorphic forms, this result also required an optimal bound for the decay rate of triple products.

Other results on moments with power saving error terms in the case $r = 4$ have been established by Kowalski-Michel-VanderKam [KMV], Iwaniec-Sarnak [IS], Blomer [Bl2], and with a slightly broader interpretation of the notion of a "moment" by Li [Li] and Khan [Kh].

2010 *Mathematics Subject Classification.* Primary 11F66; Secondary 11L07, 11F72.

Key words and phrases. Asymptotic formula, L-functions, character twists, summation formulae, p-adic methods.

The first author acknowledges the support by the Volkswagen Foundation and a Starting Grant of the European Research Council. The second author acknowledges the support by the National Security Agency. Project is sponsored by the NSA under Grant Number H98230-14-1-0139. The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation herein.

From an adelic point of view, it is natural to replace the archimedean twist by $|\det|^{it}$ with a non-archimedean twist by a Dirichlet character χ, and to consider the moments

$$\text{(A)} \quad \sum_{\chi \bmod q}^{*} |L(1/2,\chi)|^4 \qquad \text{and} \qquad \text{(B)} \quad \sum_{\chi \bmod q}^{*} |L(1/2, f\otimes\chi)|^2, \tag{1.3}$$

where the sum runs over all primitive Dirichlet characters χ modulo q and f is a fixed Hecke cusp form in the second sum. Equally interesting and related in spirit are the moments over quadratic characters only:

$$\text{(C)} \quad \sum_{\substack{d\leqslant X\\ d \text{ squarefree}}}^{*} |L(1/2,\chi_d)|^4 \qquad \text{and} \qquad \text{(D)} \quad \sum_{\substack{d\leqslant X\\ d \text{ squarefree}}}^{*} |L(1/2, f\otimes\chi_d)|^2.$$

It was a major breakthrough when M. Young [Y] established an asymptotic formula with power saving for (A) for prime numbers q:

$$\sum_{\chi \bmod q}^{*} |L(1/2,\chi)|^4 = q\sum_{i=1}^{4} c_i (\log q)^i + \mathrm{O}\left(q^{1-\frac{1}{80}+\frac{\theta}{40}+\varepsilon}\right), \tag{1.4}$$

where c_i are effectively computable constants and $\theta \leqslant 7/64$ is an admissible exponent towards the Ramanujan-Petersson conjecture.

The (harder) cases (B), (C) and (D) have remained unsolved up until now. This is perhaps a bit surprising, but it is important to notice that all 4 moments (A)–(D) single out the point 1/2, and therefore carry some *intrinsic arithmetic information*. This is in contrast to the true adelic analogues of (1.1) and (1.2) (with a test function expanding in the non-archimedean direction), which are

$$\int_{-\infty}^{\infty} \sum_{\chi \bmod q}^{*} |\Lambda(1/2+it,\chi)|^4 dt \qquad \text{and} \qquad \int_{-\infty}^{\infty} \sum_{\chi \bmod q}^{*} |\Lambda(1/2+it, f\otimes\chi)|^2 dt, \tag{1.5}$$

where Λ denotes the completed L-function. It is an interesting phenomenon that, comparing (1.5) to (1.3), an additional average of essentially bounded length in the t-aspect makes the problem incomparably easier, and indeed good asymptotic formulas for both quantities in (1.5) are fairly routine.

In this paper we couple spectral theory of automorphic forms with an algebro-arithmetic treatment of short sums of products of Kloosterman sums to solve the case (B) for 99.9% of all moduli q. Let

$$\psi(q) = \sum_{d\mid q} \phi(d)\mu\left(\frac{q}{d}\right) \tag{1.6}$$

denote the number of primitive characters modulo q. It is non-zero if and only if $q \not\equiv 2 \pmod 4$, and in this case $\psi(q) = q^{1+o(1)}$. We call a modulus $q \not\equiv 2 \pmod 4$ admissible.

Theorem 1. *For $j = 1, 2$, let f_j be (fixed) holomorphic cuspidal newforms of (even) weight κ_j for the group $\mathrm{SL}_2(\mathbb{Z})$ with Hecke eigenvalues $\lambda_j(n)$, normalized as in (2.3). Assume that $\kappa_1 \equiv \kappa_2 \pmod 4$. Let*

$$P(s) = \left(\frac{L_q(s, \mathrm{sym}^2 f_1)}{\zeta_q(2s)}\right)^{-1} = \prod_{p\mid q}\left(1 - \frac{\lambda_1(p^2)}{p^s} + \frac{\lambda_1(p^2)}{p^{2s}} - \frac{1}{p^{3s}}\right)\left(1-\frac{1}{p^{2s}}\right)^{-1}, \tag{1.7}$$

$$\begin{aligned} Q(s) &= \left(\frac{L_q(s, f_1\times f_2)}{\zeta_q(2s)}\right)^{-1} \\ &= \prod_{p\mid q}\left(1 - \frac{\lambda_1(p)\lambda_2(p)}{p^s} + \frac{\lambda_1(p^2)+\lambda_2(p^2)}{p^{2s}} - \frac{\lambda_1(p)\lambda_2(p)}{p^{3s}} + \frac{1}{p^{4s}}\right)\left(1-\frac{1}{p^{2s}}\right)^{-1}. \end{aligned} \tag{1.8}$$

Let $q \in \mathbb{N}$, and $q_1 \mid q$ be a divisor such that $(q, 6^\infty) \mid q_1$. Then,

$$\sum_{\chi \bmod q}^{*} L(1/2, f_1 \otimes \chi)\overline{L(1/2, f_2 \otimes \chi)} = \frac{2}{\zeta(2)}\psi(q) \cdot M(f_1, f_2, q) + \mathrm{O}_{f_1,f_2}\left(q^{1+\varepsilon}\left(q_1^{-\frac{1}{22}} + (q/q_1^2)^{-\frac{1}{22}}\right)\right), \tag{1.9}$$

where

$$M(f_1, f_2, q) = \begin{cases} P(1)L(1, \mathrm{sym}^2 f_1)\left(\log q + c + \dfrac{P'(1)}{P(1)}\right), & f_1 = f_2, \\ Q(1)L(1, f_1 \times f_2), & f_1 \neq f_2, \end{cases}$$

and c is a constant depending only on f_1 (not on q) given explicitly as

$$c = \gamma - \frac{1}{2}\log(2\pi) + \frac{\Gamma'(\kappa_1/2)}{\Gamma(\kappa_1/2)} + \frac{L'(1, \mathrm{sym}^2 f_1)}{L(1, \mathrm{sym}^2 f_1)} - \frac{2\zeta'(2)}{\zeta(2)}. \tag{1.10}$$

Note that $P(1), Q(1) = (\log\log q)^{\mathrm{O}(1)}$ and $P'(1)/P(1) = \mathrm{O}(\log\log q)$, and that the leading coefficients $L(1, \mathrm{sym}^2 f_1)$ and $L(1, f_1 \times f_2)$ do not vanish by the lower bounds of Hoffstein and Lockhart [HL] and Ramakrishnan and Wang [RW] (see also [Br]), so that the term $M(f_1, f_2, q)$ is not far from a linear polynomial in $\log q$ or a constant depending on f_1 and f_2.

The error term in Theorem 1 saves a power of q as soon as q has a divisor q_1 in the range

$$q^\eta \ll q_1 \ll q^{1/2-\eta}$$

for some fixed $\eta > 0$ and if in addition $2^{100} \nmid q$ and $3^{100} \nmid q$ (say) holds. We thus obtain a power saving for 99.9% of all admissible moduli q. In fact, it is not hard to see that these conditions are satisfied for all q except those that are highly divisible by 2 or 3 or are essentially a prime or the product of two primes of almost equal size, that is, those q for which there is a prime $p \geqslant q^{1-\eta}$ with $p \mid q$ or primes $p_1, p_2 \geqslant q^{1/2-\eta}$ with $p_1 p_2 \mid q$. We get the highest savings if q has a divisor of size $q_1 \asymp q^{1/3+o(1)}$, for example when $q = p^n$ is a high power of a fixed prime $p > 3$ or when q is essentially a cube, in which case our error term is $\mathrm{O}(q^{65/66+\varepsilon})$.

The condition that $(q, 6^\infty) \mid q_1$ is introduced for purely technical and notational reasons; it can be avoided without introducing any new ideas at the cost of increasing the length of the already rather long paper. In Theorem 1 and all theorems below, the condition that $\kappa_1 \equiv \kappa_2 \pmod 4$ is necessary in the sense that otherwise the product of the central values vanishes for root number reasons.

Our method works for fixed Maaß forms f_1, f_2, assuming that they satisfy the Ramanujan conjecture, which we use crucially in the course of the argument. In Section 14, we state the small modifications needed to prove the following result.

Theorem 2. *For $j = 1, 2$, let f_j be (fixed) cuspidal Maaß newforms of the same parity for the group $\mathrm{SL}_2(\mathbb{Z})$ with Hecke eigenvalues $\lambda_j(n)$. If f_1, f_2 satisfy the Ramanujan conjecture, i.e. if $\lambda_j(n) \ll n^\varepsilon$ for all $n \in \mathbb{N}$, then (1.9) holds.*

The first result in the direction of Theorems 1 and 2 in the case $f_1 = f_2$ is due to Stefanicki [St], who proved an asymptotic formula for the second moment with an error term that saves a small power of $\log q$, provided q has only few prime divisors. A formula with a $\log\log q$-saving was established by Gao-Khan-Ricotta [GKR] for almost all integers q. As either method saves less than a factor of $\log q$ in the error term, this type of argument cannot produce an asymptotic formula in the case $f_1 \neq f_2$, regardless of the factorization of q. An individual asymptotic formula with a power saving error term, and in case $f_1 \neq f_2$ an asymptotic formula with *any* saving in the error term, that would be valid for *any* infinite subset of moduli q has been a long-standing open problem until now. Theorems 1 and 2 cover, in a weak sense, almost all moduli.

Theorems 1 and 2 are concerned with the family of character twists to an individual modulus q. If an additional average over moduli q is introduced, the problem becomes easier, and indeed such versions of (B) are available due to Akbary [Ak] and, with a considerably shorter average, to Hoffstein and Lee [HL].

1.2. **Selected applications.** In addition to providing statistics in families of L-functions, asymptotic formulas with a power saving are an essential prerequisite to the analytic techniques of amplification, mollification, and resonators in questions of arithmetic importance, including upper bounds, nonvanishing, and extreme values. The allowable length of the Dirichlet polynomial (such as the amplifier), and thus the quality of arithmetic implications, is related to the strength of the power saving in the summation formula. Several such applications of Theorems 1 and 2 are featured here, beginning with the nonvanishing problem.

Combining Theorem 2 with a mollifier, one can improve the work of Stefanicki [St] to show that (for Maaß forms satisfying the Ramanujan conjecture) a *positive proportion* of L-functions with twists by primitive Dirichlet characters modulo q does not vanish at the central point, provided that q has a divisor in a suitable range. A non-vanishing result of positive proportion strength had been out of reach so far in this family.

We highlight a different application of Theorem 2 to *simultaneous* non-vanishing of twisted L-functions, as follows:

Theorem 3. *Let f_1, f_2 be two (fixed) cuspidal Maaß newforms of the same parity for $\mathrm{SL}_2(\mathbb{Z})$ that satisfy the Ramanujan conjecture, and let $\eta > 0$. Then, for every sufficiently large modulus $q \geqslant C = C(f_1, f_2, \eta)$ such that $q \not\equiv 2 \pmod 4$ and q has a divisor $q_1 \in [q^\eta, q^{1/2-\eta}]$ such that $(q, 6^\infty) \mid q_1$, there exist primitive Dirichlet characters χ modulo q such that*

$$L(1/2, f_1 \otimes \chi)\overline{L(1/2, f_2 \otimes \chi)} \neq 0,$$

and, in fact, the number of such characters is at least $q^{1/4-\varepsilon}$.

Nonvanishing results for central values of L-functions of character twists have a long history, in particular in connection with cusp forms associated to elliptic curves, but also for general automorphic forms (on fairly general reductive groups). We cannot quote here all the relevant literature, but we would like to emphasize that the focus in Theorem 3 is on the Maaß case, because in the holomorphic case one can establish extremely strong non-vanishing results by Galois-theoretic methods [Ro, Ch]. In the Maaß case, however, Theorem 3 is, at least under the assumption of the Ramanujan conjecture, the first instance of any simultaneous non-vanishing result for general twists of automorphic L-functions. The quantitative version comes from the best-known subconvexity results for twisted L-functions [BH2].

As another application of the asymptotic formula in Theorem 1 — and here the power saving is absolutely crucial — one obtains a lower bound of the correct order of magnitude for k^{th} moments of mixed products

$$\sum_{\chi \bmod q}^{*} \Big(L(1/2, f_1 \otimes \chi)\overline{L(1/2, f_2 \otimes \chi)}\Big)^k,$$

following the method of Rudnick and Soundararajan [RS, RS1]. As an illustration we provide complete details for the following result.

Theorem 4. *Let $p > 3$ be a fixed prime, and let $q = p^\kappa$ be large. Let f_1, f_2 be two fixed holomorphic cuspidal Hecke eigenforms of level 1 and respective weights κ_1, κ_2 with $\kappa_1 \equiv \kappa_2$ (mod 4). Then*

$$\sum_{\chi \bmod q}^{*} \Big(L(1/2, f_1 \otimes \chi)\overline{L(1/2, f_2 \otimes \chi)}\Big)^2 \gg q(\log q)^2.$$

We remark that with slightly more technical effort one can show by the same method the general lower bound

$$\sum_{\chi \bmod q}^{*} \Big(L(1/2, f_1 \otimes \chi)\overline{L(1/2, f_2 \otimes \chi)}\Big)^k \gg q(\log q)^{k^2/2}$$

for any even integer $k \geqslant 2$, as well as similar results (up to a factor of $(\log q)^{-\varepsilon}$) for more general q, as in Theorem 1. Note that $L(1/2, f_1 \otimes \chi)\overline{L(1/2, f_2 \otimes \chi)}$ is real, cf. (3.1) below. The proof of Theorem 4 will be given at the end of the paper.

1.3. **The methods.** In this section, we sketch the method of proof of Theorem 1 and highlight some auxiliary results of independent interest, in particular Lemma 2 and Theorems 5, 8, and 10.

A natural starting point is an approximate functional equation, and there are two options: one can either take an approximate functional equation for $L(s, f_1 \otimes \chi)\overline{L(s, f_2 \otimes \chi)}$ with root number independent of χ, or the product of two separate approximate functional equations for $L(s, f_1 \otimes \chi)$ and $L(s, f_2 \otimes \chi)$, each of which has a root number depending on χ. Summing over χ, one obtains either way an expression roughly of the shape

$$\sum_{\substack{nm \leqslant q^2 \\ n \equiv m \bmod q}} \lambda_1(m)\lambda_2(n),$$

where $\lambda_j(n)$ denotes the normalized n-th Hecke eigenvalue of f_j. We need to beat the trivial bound $\mathrm{O}(q^{1+\varepsilon})$ for the contribution of the off-diagonal terms $n \neq m$ by a small, but fixed power of q. There are two ways to interpret this double sum: either as a shifted convolution problem, or as a problem of summing Hecke eigenvalues in arithmetic progressions. The former point of view is useful if n and m are not too far apart, the latter if one variable is sufficiently small compared to the other variable. For clarity, let us restrict $n \asymp N$ and $m \asymp M$ to dyadic intervals and assume $N \geqslant M$ by symmetry and (for the sake of argument) $NM = q^2$, which is supposedly the hardest range. On the one hand, we can apply Voronoi summation to the inner sum in

$$\sum_{m \asymp M} \lambda_1(m) \sum_{\substack{n \asymp N \\ n \equiv m \bmod q}} \lambda_2(n),$$

getting roughly

$$\frac{N}{q^2} \sum_{m \asymp M} \sum_{n \asymp q^2/N} \lambda_1(m)\lambda_2(n) S(n, m, q). \tag{1.11}$$

The trivial bound at this point, using Deligne (or Rankin-Selberg) and Weil bounds, is $Mq^{1/2}$, which is admissible if $M \leqslant q^{1/2-\delta}$ (or equivalently $N \geqslant q^{3/2+\delta}$).

Alternatively, we can consider the average of shifted convolution problems

$$\sum_{r \asymp N/q} \sum_{\substack{n \asymp N, m \asymp M \\ n-m=rq}} \lambda_1(m)\lambda_2(n). \tag{1.12}$$

There is by now a well-developed toolbox of methods for handling shifted convolution problems. The first step is always to detect the linear condition $n - m = rq$ by additive characters. The corresponding (horocycle) integral can then be decomposed by a variant of the circle method. Voronoi summation in the n, m-variables leads to sums of Kloosterman sums which can be analyzed spectrally through the Kuznetsov formula. Alternatively (and quite similarly in spirit), one can apply Mellin inversion and the unfolding trick to express the horocycle integral directly as a triple product involving Poincaré series which can again be decomposed spectrally. This is the strategy followed by Good [Go] and Sarnak [Sa]. Finally, as a third option, one can use carefully chosen vectors in the representation space of the automorphic representations generated by f_1 and f_2 to spectrally decompose the horocycle integral directly [BH1]. In all approaches, the n, m-sum can be spectrally expanded, and the resulting expansion can then be summed over r.

In this paper, we follow [Bl1, BHM] and start with a very flexible variant of the circle method due to Jutila. To speed up the performance, we observe that, although $n \asymp N$, the n-sum is in reality relatively short, namely $n = rq + \mathrm{O}(M)$. One of the main devices in the argument is the well-known trick of attaching a redundant weight function that localizes n at $rq + \mathrm{O}(M)$ (which is of course automatic in (1.12), but gets "forgotten" in the course of the manifold transformations unless we remember it explicitly by an additional weight function). The price for this manoeuvre is a very subtle and delicate analysis with Bessel functions,

for which we prepare in Section 6. As a first order approximation, we end up with an expression roughly of the form

$$\frac{M^3}{C^3N^{3/2}}\sum_{t_j\leqslant(N/M)^{1/2}}\lambda_j(q)\sum_{r\asymp N/q}\lambda_j(r)\sum_{m\asymp C^2/M}\sum_{n\asymp C^2N/M^2}\lambda_1(m)\lambda_2(n)\lambda_j(n-m), \tag{1.13}$$

where $C=N^{1000}$ is a very large parameter and the outermost spectral sum runs over a basis of level 1 Maaß forms with spectral parameter $t_j\leqslant(N/M)^{1/2}$. We caution that (1.13) is a much oversimplified expression that reflects reality only in a very vague sense; in particular, some extra cost has to be paid to separate variables, there is also a continuous spectrum contribution, and the level is not always 1, but sometimes a bit larger. Note that the two innermost sums resemble the triple products that would arise in a direct spectral analysis.

One can now apply the Cauchy-Schwarz inequality and the spectral large sieve of Deshouillers-Iwaniec, thus obtaining the final bound $Nq^{\theta-1/2}$ plus some more terms that are smaller in typical ranges. (Here, as usual, θ denotes an admissible exponent towards the Ramanujan-Petersson conjecture.) We point out the interesting feature of Jutila's method that the auxiliary parameter C is only a catalyst that does not enter the final bound and conclude that this analysis is admissible if $N\leqslant q^{3/2-\theta-\delta}$.

Obviously, the ranges $N\geqslant q^{3/2+\delta}$ and $N\leqslant q^{3/2-\theta-\delta}$ do not overlap, not even assuming the Ramanujan conjecture ($\theta=0$). The overall strategy up to this point is similar to the analysis in [Y], and both here and there the main problem is to overcome the small gap in the two ranges. Young uses the fact that one can decompose the divisor function in order to get more variables with which one can apply Poisson summation. The corresponding saving is strong enough to close the gap. This route is not available in the present situation.

As the first step, we remove the dependence on the Ramanujan conjecture by applying Hölder's inequality to (1.13) with exponents 1/4, 1/4, 1/2, getting

$$\frac{M^3}{C^3N^{3/2}}\Big(\sum_{t_j\leqslant(\frac{N}{M})^{1/2}}|\lambda_j(q)|^4\Big)^{\frac14}\Big(\sum_{t_j\leqslant(\frac{N}{M})^{1/2}}\Big|\sum_{r\asymp\frac{N}{q}}\lambda_j(r)\Big|^4\Big)^{\frac14}\Big(\sum_{t_j\leqslant(\frac{N}{M})^{1/2}}\Big|\sum_{h\asymp\frac{C^2N}{M^2}}\lambda_j(h)\sum_{\substack{m\asymp\frac{C^2}{M}\\ n\asymp\frac{C^2N}{M^2}\\ n-m=h}}\lambda_1(m)\lambda_2(n)\Big|^2\Big)^{\frac12}.$$

After expanding one of the squares inside the fourth powers using multiplicativity, we apply the Kuznetsov formula for the first factor, and the large sieve (which is, of course, also based on the Kuznetsov formula) for the other two factors, getting a bound roughly of the strength $Nq^{-1/2}$ without dependence on the Ramanujan conjecture. A precise version can be found in Proposition 7 below. The crucial input is Theorem 8, which presents a flexible variant of the spectral large sieve that allows for additional divisibility conditions (that are, in turn, essential to the success of our method) without being wasteful. This requires, among other things, an orthonormalization of the collection of Maaß forms $\{f(dz):d\mid\ell\}$ for a newform f and some integer ℓ, where ℓ is not necessarily squarefree; see Lemma 2.

This procedure works in great generality. As observed by Fouvry, Kowalski and Michel, the methods we employ improve Young's result (1.4) on the fourth moment of Dirichlet L-functions:

$$\sideset{}{^*}\sum_{\chi\bmod q}|L(1/2,\chi)|^4=q\sum_{i=1}^4c_i(\log q)^i+\mathrm{O}\left(q^{1-\frac{1}{82}+\varepsilon}\right)$$

for primes p. At the current state of knowledge, this is better than (1.4) and – more importantly – independent of bounds towards the Ramanujan-Petersson conjecture. Inserting more algebraic geometry, the error term in (1.4) can in fact be improved to $\mathrm{O}(q^{1-1/32+\varepsilon})$ with no recourse on bound towards the Ramanujan-Petersson conjecture. This result is contained, among other things, in the companion paper [BFKMM], which imports the spectral analysis discussed in this subsection.

Returning to the situation of Theorem 1, it now remains to close the "small" gap where $N = q^{3/2+o(1)}$ and $M = q^{1/2+o(1)}$, for which an essentially new idea is necessary. We use the Cauchy-Schwarz inequality to bound (1.11) by

$$\frac{NM^{1/2}}{q^2}\Big(\sum_{n_1,n_2\asymp q^2/N}\Big|\sum_{m\asymp M}S(m,n_1,q)S(m,n_2,q)\Big|\Big)^{1/2}. \tag{1.14}$$

Weil's individual bound for Kloosterman sums yields an upper bound of $Mq^{1/2}$, and we win if we can prove some extra cancellation for generic pairs (n_1, n_2) in the short m-sum. Note that at this point all automorphic information is gone, and we are left with a problem of bounding exponential sums, namely short sums of products of two Kloosterman sums. Generically, the length of the m-sum is roughly the square-root of the modulus of the two Kloosterman sums, so this seems to be a hard problem in general.

1.4. **Short sums of products of Kloosterman sums.** The crucial new arithmetic input of this paper is a non-trivial estimation of the inner double sum in (1.14) if q is sufficiently factorable. In fact, we can estimate the individual m-sums with pleasing success generically and only use the sum over n_1, n_2 to control the frequency of "nearly diagonal" pairs (n_1, n_2). Our analysis is somewhat inspired by Heath-Brown's paper on hybrid bounds for Dirichlet L-functions [HB]. Our situation is more involved, since the function $b(m) = S(m, n_1, q)S(m, n_2, q)$ is not multiplicative in m (not even in some twisted sense), unlike a Dirichlet character $\chi(m)$ modulo q. Moreover, for higher prime powers $q = p^s$, the Kloosterman sum resembles the exponential of a p-adic square-root, and therefore much more genuine p-adic methods naturally enter the analysis of the corresponding multiple exponential sums.

Nevertheless, provided that we can factorize $q = r_1r_2$ with $(r_1, r_2) = 1$, a careful application of Weyl differencing with respect to r_2 (presented in Lemma 12), followed by an application of Poisson summation to effect the technique of "completion", yields a bound roughly of the form

$$\Big|\sum_{m\asymp M}S(m,n_1,q)S(m,n_2,q)\Big|^2 \ll M^2q^2\Big(\frac{r_2}{M}+\frac{r_2^2}{M^2}+\frac{\widehat{S}}{r_1^2M}+\frac{\widehat{S}}{r_1^3}\Big),$$

where $\widehat{S}$ is the average of *complete* sums of the type

$$\sum_{m \bmod r_1} S(m,n_1,r_1)S(m,n_2,r_1)S(m+h,n_1,r_1)S(m+h,n_2,r_1)e\left(\frac{km}{r_1}\right) \tag{1.15}$$

for various values of k and h. A general version of the underlying idea is presented in Theorem 10 in Section 9, which may be of use in other situations.

If r_1 is squarefree, one can use the independence of Kloosterman sheafs [Ka] to obtain square-root cancelation (in generic situations) in the multiple exponential sum (1.15). For the squareful parts, we obtain a bound of generically similar strength via an unexpectedly involved p-adic stationary phase argument that features, among other things, singular critical points; the latter are necessary to obtain results for the class of moduli of the stated generality and (as will be evident from our treatment) provably contribute to the correct order of magnitude. It is common belief that exponential sums to squareful moduli are easy to handle; while it is true that their treatment is *elementary* (in the sense that in most ranges no algebraic geometry is needed), the analysis is often extremely complicated, and the treatment of degenerate cases can turn out to be quite involved (see [DF] for an example of GL(3) Kloosterman sums). The upshot of the above discussion is the following result:

Theorem 5. *Let $r, q, n_1, n_2 \in \mathbb{N}$ with $r \mid q$, let $A \in \mathbb{R}$, $M > 1$. Then, for any $s \mid r$ satisfying $(r, 6^\infty) \mid s$ we have*

$$\sum_{\substack{A<m\leqslant A+M\\(m,q)=1}} S(m,n_1,r)S(m,n_2,r) \ll r^\varepsilon\Big(M^{1/2}rs^{1/2}+\frac{M^{1/2}r^{5/4}}{s^{1/4}}+Mr^{3/4}(r,n_1-n_2)^{1/4}s^{1/4}+rs+\sigma\Big),$$

where the term σ defined in (10.10) satisfies $\sigma = 0$ if $r/(r, s^\infty)$ is cube-free, and $\sigma \ll r^{11/8}s^{1/8}$ in all cases.

As in Theorem 1, with a bit more work the condition $(r, 6^\infty) \mid s$ could be removed in Theorem 5; it affects only moduli r divisible by extremely high powers of 2 or 3.

Comparing with the "trivial" bound $Mr^{1+o(1)}$ on the left-hand side, and assuming for simplicity that $(r, n_1n_2(n_1 - n_2)) = 1$, we obtain a power saving as long as

$$\frac{r}{M^2}(rM)^\eta \ll s \ll \min\left(M, \frac{M^8}{r^3}\right)(rM)^{-\eta},$$

where the term involving M^8/r^3 can simply be omitted (and the ranges of application in s extended) if $r/(r, s^\infty)$ is cube-free. In the important range $M \asymp r^{1/2}$, this gives a power saving as long as r has a divisor s in the (essentially full) range

$$r^\eta \ll s \ll r^{1/2-\eta}$$

(with the above constraint on high powers of 2 and 3). This holds for 99.9% of all r.

As an application, let us consider the most interesting range, the "square-root threshold" $M \asymp r^{1/2}$. If $(r, n_1n_2(n_1 - n_2)) = 1$ and r has a divisor in the range $s \asymp r^{1/3}$, we obtain the bound $r^{17/12+\varepsilon}$, an improvement of $r^{1/12}$ over the "trivial" bound $r^{3/2}$. For a sum such as that featured in Theorem 5, with $\asymp r^{1/2}$ terms of arithmetic nature to modulus r of size $\asymp r^{1+o(1)}$, it may be reasonable to speculate that the best possible bound (and the true order of magnitude) is $\asymp r^{5/4+o(1)}$. Our bound thus reaches $\frac{1}{3}$ of the way from the trivial to the best possible result and may be seen as the analogue of the "Weyl exponent" in this case.

In the case when $r = p^s$ is a sufficiently high prime power and $(r, n_1n_2(n_1 - n_2)) = 1$, Theorem 5 is concerned with a short sum of exponentials with a p-adically analytic phase that may be directly estimated by [Mi, Theorem 2]. In fact, in the crucial range $M \asymp r^{1/2}$ this yields a bound of sub-Weyl strength $r^{17/12-\delta}$ in the situation of Theorem 5 and consequently a stronger error term of the form $O(q^{65/66-\delta'})$ in Theorem 1 in the case of a prime power modulus q, with some small but fixed $\delta, \delta' > 0$. The corresponding route does not appear to be as readily available for more general r (not even at high prime power divisors of r), since, absent additional arithmetic conditions on the divisor s, degenerate critical points genuinely must be considered.

Using Theorem 5 in (1.14), we obtain Proposition 6 below, which enables us to complete the proof of Theorem 1. We finally remark that the pleasing generality of the moduli considered in this paper requires a lot of technical overhead (in both the automorphic and the algebro-arithmetic treatment) that contributes to the length of the paper.

Acknowledgements. We would like to take the opportunity to thank Étienne Fouvry, Emmanuel Kowalski, Philippe Michel, Lillian Pierce and Guillaume Ricotta for helpful remarks and discussions. This paper grew out of the conversations we had while the second author visited the Max Planck Institute for Mathematics in Bonn; it is a pleasure to acknowledge the support and excellent research infrastructure at MPIM.

2. Automorphic Preliminaries I

We follow the notation of [BHM]. We write the Fourier expansion of a holomorphic modular form f of level ℓ and weight k as

$$f(z) = \sum_{n\geqslant 1} \rho_f(n)(4\pi n)^{k/2}e(nz),$$

and similarly we write for a Maaß form f of level ℓ and spectral parameter $t = t_f \in \mathbb{R} \cup [-i\theta, i\theta]$ (where currently $\theta = 7/64$ is known)

$$f(z) = \sum_{n \neq 0} \rho_f(n) W_{0,it}(4\pi|n|y) e(nx) \tag{2.1}$$

where $W_{0,it}(y) = (y/\pi)^{1/2} K_{it}(y/2)$ is a Whittaker function. The inner product of two Maaß forms f and g of level ℓ is given by

$$\langle f, g \rangle := \int_{\Gamma_0(\ell)\backslash\mathbb{H}} f(z)\overline{g(z)} \frac{dx\, dy}{y^2}. \tag{2.2}$$

For each cusp $\mathfrak{a}$ of $\Gamma_0(\ell)$ there is an Eisenstein $E_{\mathfrak{a}}(z, s)$ series whose Fourier expansion at $s = 1/2 + it$ we write as

$$E_{\mathfrak{a}}(z, 1/2 + it) = \delta_{\mathfrak{a}=\infty} y^{1/2+it} + \varphi_{\mathfrak{a}}(1/2 + it) y^{1/2-it} + \sum_{n \neq 0} \rho_{\mathfrak{a}}(n, t) W_{0,it}(4\pi|n|y) e(nx).$$

If f is a cuspidal newform (and in particular an eigenform of all Hecke operators), we denote its normalized Hecke eigenvalues by $\lambda_f(n)$ and record the relation

$$\lambda_f(n)\rho_f(1) = \sqrt{n}\rho_f(n) \tag{2.3}$$

for $n \geqslant 1$, and $\rho_f(-n) = \pm\rho_f(n)$ in the Maaß case (since f is an eigenform of the involution $z \mapsto -\bar{z}$). For future reference we state the well-known bounds (e.g. [HM, (30)])

$$|\rho_f(1)|^2 = \frac{\cosh(\pi t_f)}{\ell} (\ell(1 + |t_f|))^{o(1)} \tag{2.4}$$

for a *newform* f of level ℓ which are essentially due to Hoffstein-Lockhart (upper bound) and Iwaniec (lower bound). We will frequently use the Hecke relation

$$\lambda_f(nm) = \sum_{d|(n,m)} \mu(d)\chi_0(d)\lambda_f\left(\frac{n}{d}\right)\lambda_f\left(\frac{m}{d}\right), \quad n, m \in \mathbb{N}, \tag{2.5}$$

where χ_0 is the trivial character modulo ℓ, and the Rankin-Selberg bound

$$\sum_{n \leqslant x} |\lambda_f(n)|^2 \ll_f x. \tag{2.6}$$

If f is in addition holomorphic, then we have Deligne's bound [De]

$$\lambda_f(n) \ll n^{\varepsilon}. \tag{2.7}$$

This is expected to hold for Maaß newforms (of arbitrary level) as well, but in general we only know

$$\lambda_f(n) \ll n^{\theta+\varepsilon}, \tag{2.8}$$

where θ is an admissible exponent for the Ramanujan-Petersson conjecture. Currently $\theta = 7/64$ is known [KS]. Wilton's bound gives

$$\sum_{n \leqslant x} \lambda_f(n) e(\alpha n) \ll_f x^{1/2+\varepsilon}, \tag{2.9}$$

uniformly in $\alpha \in \mathbb{R}$.

For a smooth, compactly supported function $V : (0, \infty) \to \mathbb{C}$ and fixed $\kappa \in \mathbb{N}$ define the Hankel-type transform

$$\mathring{V}(y) = 2\pi i^{\kappa} \int_0^{\infty} V(x) J_{\kappa-1}(4\pi\sqrt{xy}) dx. \tag{2.10}$$

It depends on κ, but this is not displayed in the notation. It is easy to see that $\mathring{V}$ is a Schwartz class function; indeed, by [BM, Section 2.6] we have

$$\int_0^\infty V(x)J_{\kappa-1}(4\pi\sqrt{xy})dx = \left(-\frac{1}{2\pi\sqrt{y}}\right)^j \int_0^\infty \frac{\partial^j}{\partial x^j}\left(V(x)x^{-\frac{\kappa-1}{2}}\right) x^{\frac{\kappa-1+j}{2}} J_{\kappa-1+j}(4\pi\sqrt{xy})dx \tag{2.11}$$

for any $j \in \mathbb{N}_0$, and now one can differentiate under the integral sign using [GR, 8.471.2].

The Mellin transform of a function f will always be denoted by $\widehat{f}$. More integral transforms will be introduced in the context of the Kuznetsov formula. The following formula is standard (e.g. [HM, Proposition 1]).

Lemma 1. *[Voronoi summation] Let $c \in \mathbb{N}$, $b \in \mathbb{Z}$, and assume $(b,c) = 1$. Let V be a smooth compactly supported function, and let $N > 0$. Let $\lambda(n)$ denote the normalized Hecke eigenvalues of a holomorphic cuspidal newform of weight κ for $\mathrm{SL}_2(\mathbb{Z})$. Then*

$$\sum_n \lambda(n)e\left(\frac{bn}{c}\right)V\left(\frac{n}{N}\right) = \frac{N}{c}\sum_n \lambda(n)e\left(-\frac{\bar{b}n}{c}\right)\mathring{V}\left(\frac{n}{c^2/N}\right).$$

3. The core argument

3.1. **The main term.** In this section, we present the backbone of the proof of the Theorem 1. By a standard approximate functional equation ([IK, Theorem 5.3]) we have for each primitive character χ modulo q that

$$L(1/2, f_1 \otimes \chi)\overline{L(1/2, f_2 \otimes \chi)} = \sum_{n,m} \frac{(\lambda_1(m)\lambda_2(n) + \lambda_2(m)\lambda_1(n))\chi(m)\bar{\chi}(n)}{(nm)^{1/2}} W\left(\frac{nm}{q^2}\right) \tag{3.1}$$

where

$$W(x) = \frac{1}{2\pi i}\int_{(2)} \frac{\Gamma(\kappa_1/2+s)\Gamma(\kappa_2/2+s)}{(2\pi)^{2s}\Gamma(\kappa_1/2)\Gamma(\kappa_2/2)} x^{-s}\frac{ds}{s} \tag{3.2}$$

satisfies $W^{(j)}(x) \ll_{A,j} (1+x)^{-A}$ for all $A, j \geqslant 0$. Note that by [IK, Proposition 14.20] the L-function $L(s, f_1 \otimes \chi)\overline{L(s, f_2 \otimes \chi)}$ has root number 1 if $\kappa_1 \equiv \kappa_2 \pmod 4$. Summing over all primitive characters χ and using the elementary identity

$$\sideset{}{^*}\sum_{\chi \bmod q} \chi(n) = \sum_{d|(n-1,q)} \phi(d)\mu(q/d),$$

for $(n,q) = 1$, we obtain

$$\sideset{}{^*}\sum_{\chi \bmod q)} L(1/2, f_1 \otimes \chi)\overline{L(1/2, f_2 \otimes \chi)} = 2\sum_{d|q}\phi(d)\mu\left(\frac{q}{d}\right) \sum_{\substack{n \equiv m \bmod d \\ (nm,q)=1}} \frac{\lambda_1(m)\lambda_2(n)}{(nm)^{1/2}} W\left(\frac{nm}{q^2}\right). \tag{3.3}$$

The diagonal term $n = m$ contributes

$$\Delta(q) = 2\psi(q)\sum_{(n,q)=1}\frac{\lambda_1(n)\lambda_2(n)}{n}W\left(\frac{n^2}{q^2}\right) = \frac{2\psi(q)}{2\pi i}\int_{(1)} \frac{L^{(q)}(1+2s, f_1 \times f_2)}{\zeta^{(q)}(2(1+2s))} q^{2s}\widehat{W}(s)ds$$

where the superscript (q) denotes omission of the Euler factors at primes dividing q. Define P, Q and c as in (1.7) – (1.10) so that in particular

$$\frac{L^{(q)}(1+2s, f_1 \times f_2)}{\zeta^{(q)}(2(1+2s))} = \frac{L(1+2s, f_1 \times f_2)}{\zeta(2(1+2s))}Q(s).$$

Shifting the contour to $\Re s = -1/4 + \varepsilon$, we obtain

$$\Delta(q) = 2\psi(q)\frac{P(1)L(1, \mathrm{sym}^2 f_1)}{\zeta(2)}\left(\log q + c + \frac{P'(1)}{P(1)} + O\left(q^{-\frac{1}{2}+\varepsilon}\right)\right), \quad f_1 = f_2,$$

and

$$\Delta(q) = 2\psi(q)\frac{Q(1)L(1, f_1 \times f_2)}{\zeta(2)}\left(1 + O\left(q^{-\frac{1}{2}+\varepsilon}\right)\right), \quad f_1 \neq f_2.$$

3.2. **The off-diagonal term.** We proceed to treat the off-diagonal contribution $n \neq m$ in (3.3). We attach a smooth partition of unity to the n- and m-sum, and localize the variables at $N \leqslant n \leqslant 2N$ and $M \leqslant m \leqslant 2M$ with weight functions v_1, v_2, where $N, M \geqslant 1$ and $NM \leqslant q^{2+\varepsilon}$ (at the cost of a negligible error). By Mellin inversion we are left with bounding

$$\sum_{d|q} d \Big| \int_{(\varepsilon)} \widehat{W}(s) \sum_{\substack{n \equiv m \bmod d \\ (nm,q)=1 \\ n \neq m}} \frac{\lambda_1(m)\lambda_2(n)}{(nm)^{1/2}} v_1\left(\frac{n}{N}\right) v_2\left(\frac{m}{M}\right)\left(\frac{nm}{q^2}\right)^{-s} \frac{ds}{2\pi i}\Big|.$$

By Stirling's formula, $\widehat{W}$ is exponentially decreasing on vertical lines, so that we can truncate the integral at $|\Im s| \leqslant (\log 5q)^2$ at a negligible cost. It therefore suffices to bound

$$S_{N,M,d,q} := \frac{d}{(NM)^{1/2}} \sum_{\substack{n \equiv m \bmod d \\ (nm,q)=1 \\ n \neq m}} \lambda_1(m)\lambda_2(n) V_1\left(\frac{m}{M}\right) V_2\left(\frac{n}{N}\right). \tag{3.4}$$

for $d \mid q$ and $N \geqslant M$ (by symmetry) for functions $V_{1,2}$ with compact support in $[1, 2]$ and derivatives bounded by

$$V_{1,2}^{(j)}(x) \ll (\log 5q)^{2j} \ll_j q^{\varepsilon}. \tag{3.5}$$

Using Deligne's bound[1] (2.7), we obtain immediately a trivial bound

$$S_{N,M,d,q} \ll \frac{d}{(NM)^{1/2-\varepsilon}} \sum_{M \leqslant m \leqslant 2M} \sum_{\substack{N \leqslant n \leqslant 2N \\ n \equiv m \bmod d \\ n \neq m}} 1 \ll (NM)^{1/2+\varepsilon}. \tag{3.6}$$

In the next section we will show

Proposition 6. *Let $q_1 \mid q$ be a divisior satisfying $(q, 6^\infty) \mid q_1$. Then*

$$S_{N,M,d,q} \ll q^{\varepsilon}\frac{q}{N^{1/2}}\left(M^{1/4}q^{1/2}q_1^{1/4} + \frac{M^{1/4}q^{5/8}}{q_1^{1/8}} + M^{1/2}q^{3/8}q_1^{1/8} + (qq_1)^{1/2} + q^{11/16}q_1^{1/16}\right)$$

for any $d \mid q$ whenever $N \geqslant 20M$.

To see when this result will be useful for us, we assume that $NM = q^2$. If $q_1 \leqslant q^{1/2}$, Proposition 6 covers the range $N \geqslant q^{3/2+\delta}$, for any fixed $\delta > 0$. (In fact, the trivial estimate on the upper bound on $S_{N,M,d,q}$ reached by (4.3) suffices in this range of parameters, and this requires no special divisibility properties of q.) However, if we can find q_1 such that $q^{\eta} \leqslant q_1 \leqslant q^{1/2-\eta}$, then we can extend the range for N slightly beyond $q^{3/2}$, so that it will overlap with the admissible range in Proposition 7 below.

Now let $\ell_1, \ell_2 \in \mathbb{N}$, $h \in \mathbb{N}$, and define

$$\mathcal{D}(\ell_1, \ell_2, h, N, M) = \sum_{\ell_1 n - \ell_2 m = h} \lambda_1(m)\lambda_2(n) V_1\left(\frac{\ell_2 m}{M}\right) V_2\left(\frac{\ell_1 n}{N}\right) \tag{3.7}$$

and

$$\mathcal{S}(\ell_1, \ell_2, d, N, M) = \sum_{r} \mathcal{D}(\ell_1, \ell_2, rd, N, M) \tag{3.8}$$

[1]This is the only point in the argument where Deligne's bound seems unavoidable.

where d is a positive integer. Note that the support of V_2 restricts $r \leqslant 2N/d$. From [Bl1, Theorem 3] we quote the individual uniform bound

$$\mathcal{D}(\ell_1, \ell_2, h, N, M) \ll (N+M)^{1/2+\theta}(NMq)^{\varepsilon}. \tag{3.9}$$

From (3.4) we obtain by Möbius inversion, (2.5) and (2.7) that

$$\begin{aligned} S_{N,M,d,q} &= \frac{d}{(NM)^{1/2}} \sum_{r=1}^{2N/d} \sum_{\substack{n-m=rd \\ (nm,q)=1}} \lambda_1(m)\lambda_2(n) V_1\left(\frac{m}{M}\right) V_2\left(\frac{n}{N}\right) \\ &= \frac{d}{(NM)^{1/2}} \sum_{g_1|f_1|q} \mu(g_1)\mu(f_1)\lambda_2\left(\frac{f_1}{g_1}\right) \sum_{g_2|f_2|q} \mu(g_2)\mu(f_2)\lambda_1\left(\frac{f_2}{g_2}\right) \sum_{r=1}^{2N/d} \mathcal{D}(f_1g_1, f_2g_2, rd, N, M) \\ &\ll \frac{d}{(NM)^{1/2}} \sum_{\substack{g_1|f_1|q \\ g_2|f_2|q}} (f_1f_2)^{\varepsilon} \Big|\mathcal{S}(f_1g_1, f_2g_2, d, N, M)\Big|. \end{aligned} \tag{3.10}$$

Sections 7 and 8 are devoted to the proof of

Proposition 7. *Let $\ell_1, \ell_2, d \in \mathbb{N}$, $N, M \geqslant 1$ and define $\mathcal{S}(\ell_1, \ell_2, d, N, M)$ as in (3.7) – (3.8). Assume that $N \geqslant 20M$. Then*

$$\mathcal{S}(\ell_1, \ell_2, d, N, M) \ll (dN)^{\varepsilon}\left(\frac{N}{d^{1/2}} + \frac{N^{5/4}M^{1/4}}{d} + \frac{N^{3/4}M^{1/4}}{d^{1/4}} + \frac{NM^{1/2}}{d^{3/4}}\right).$$

The implicit constant depends on ε alone.

This implies

$$S_{N,M,d,q} \ll \left(\frac{(Nq)^{1/2}}{M^{1/2}} + \frac{N^{3/4}}{M^{1/4}} + \frac{N^{1/4}q^{3/4}}{M^{1/4}} + N^{1/2}q^{1/4}\right)(qN)^{\varepsilon} \tag{3.11}$$

for $N \geqslant 20M$, while from (3.9) and (3.10) we conclude by trivial estimates

$$S_{N,M,d,q} \ll \frac{d}{(NM)^{1/2}} q^{\varepsilon} \frac{N}{d} N^{1/2+\theta} = \frac{q^{\varepsilon}N^{1+\theta}}{M^{1/2}} \tag{3.12}$$

in the slightly larger range $N \geqslant M$.

3.3. **An optimization problem.** We are now prepared to prove Theorem 1. First we observe that (3.12) in connection with our general assumption $NM \leqslant q^{2+\varepsilon}$ suffices to prove Theorem 1 whenever $N \asymp 20M$. Hence from now on we assume $N \geqslant M$ so that Proposition 6 and (3.11) are available. In preparation for later estimates, we observe that (3.11) implies

$$S_{N,M,d,q} \ll \frac{q^{3/4+\varepsilon}N^{1/4}}{M^{1/4}}, \quad \text{if} \quad NM \leqslant q^{2+\varepsilon}, N \leqslant Mq. \tag{3.13}$$

We distinguish two cases.

Case I: $q_1 \leqslant q^{1/3}$. In this case we need to show $S_{N,M,d,q} \ll q^{1+\varepsilon}q_1^{-1/22}$. The bound (3.6) is admissible unless

$$q^2 q_1^{-1/11} \leqslant NM \leqslant q^{2+\varepsilon}. \tag{3.14}$$

In this range, (3.13) is admissible unless

$$N/M \geqslant q q_1^{-2/11}. \tag{3.15}$$

If both (3.14) and (3.15) hold, then Proposition 6 implies that $S_{N,M,d,q}$ is, up to a factor q^{ε}, at most

$$\frac{q^{3/2}q_1^{1/4}}{(N/M)^{3/8}(NM)^{1/8}}+\frac{q^{13/8}q_1^{-1/8}}{(N/M)^{3/8}(NM)^{1/8}}+\frac{q^{11/8}q_1^{1/8}}{(N/M)^{1/2}}+\frac{q^{3/2}q_1^{1/2}}{(N/M)^{1/4}(NM)^{1/4}}+\frac{q^{27/16}q_1^{1/16}}{(N/M)^{1/4}(NM)^{1/4}}$$
$$\ll q^{7/8}q_1^{29/88}+qq_1^{-1/22}+q^{7/8}q_1^{19/88}+q^{3/4}q_1^{25/44}+q^{15/16}q_1^{23/176}\ll qq_1^{-1/22}$$

for $q_1\leqslant q^{1/3}$.

Case II: $q^{1/3}\leqslant q_1\leqslant q^{1/2}$. In this case we need to show $S_{N,M,d,q}\ll q^{21/22+\varepsilon}q_1^{1/11}$. The bound (3.6) is admissible unless

$$q^{21/11}q_1^{2/11}\leqslant NM\leqslant q^{2+\varepsilon}. \tag{3.16}$$

In this range, (3.13) is admissible unless

$$N/M\geqslant q^{9/11}q_1^{4/11}. \tag{3.17}$$

If both (3.16) and (3.17) hold, then Proposition 6 implies that $S_{N,M,d,q}$ is, up to a factor q^{ε}, at most

$$\frac{q^{3/2}q_1^{1/4}}{(N/M)^{3/8}(NM)^{1/8}}+\frac{q^{13/8}q_1^{-1/8}}{(N/M)^{3/8}(NM)^{1/8}}+\frac{q^{11/8}q_1^{1/8}}{(N/M)^{1/2}}+\frac{q^{3/2}q_1^{1/2}}{(N/M)^{1/4}(NM)^{1/4}}+\frac{q^{27/16}q_1^{1/16}}{(N/M)^{1/4}(NM)^{1/4}}$$
$$\ll q^{21/22}q_1^{1/11}+q^{95/88}q_1^{-25/88}+q^{85/88}q_1^{-5/88}+q^{9/11}q_1^{4/11}+q^{177/176}q_1^{-13/176}\ll q^{21/22+\varepsilon}q_1^{1/11}$$

for $q^{1/3}\leqslant q_1\leqslant q^{1/2}$.

4. Hecke eigenvalues in residue classes

In this section we prove Proposition 6, assuming the validity of Theorem 5 whose proof we postpone to the end of the paper. The method presented here is strong if N is much larger than M. Initially we only assume $N\geqslant 20M$, so that the condition $n\neq m$ is moot. We write

$$S_{N,M,d,q}=\frac{d}{(NM)^{1/2}}\sum_{(m,q)=1}\lambda_1(m)V_1\left(\frac{m}{M}\right)\sum_{\substack{n\equiv m \bmod d\\(n,q)=1}}\lambda_2(n)V_2\left(\frac{n}{N}\right).$$

Let us write $q=q_dq'$ where $q_d=(q,d^{\infty})$ and hence $(q',d)=1$. Since $(m,q)=1$ and $n\equiv m \pmod d$, the conditions $(n,q)=1$ and $(n,q')=1$ are equivalent. We remove the latter condition by Möbius inversion and (2.5), getting

$$\begin{aligned}S_{N,M,d,q}&=\frac{d}{(NM)^{1/2}}\sum_{f|q'}\mu(f)\sum_{(m,q)=1}\lambda_1(m)V_1\left(\frac{m}{M}\right)\sum_{n\equiv\bar{f}m \bmod d}\lambda_2(fn)V_2\left(\frac{fn}{N}\right)\\&=\frac{d}{(NM)^{1/2}}\sum_{g|f|q'}\mu(f)\mu(g)\lambda_2\left(\frac{f}{g}\right)\sum_{(m,q)=1}\lambda_1(m)V_1\left(\frac{m}{M}\right)\sum_{n\equiv\overline{fg}m \bmod d}\lambda_2(n)V_2\left(\frac{fgn}{N}\right).\end{aligned} \tag{4.1}$$

The innermost sum in (4.1) equals

$$\frac{1}{d}\sum_{r|d}\sideset{}{^*}\sum_{b \bmod r}e\left(\frac{\overline{fg}mb}{r}\right)\sum_n\lambda_2(n)e\left(-\frac{bn}{r}\right)V_2\left(\frac{fgn}{N}\right).$$

Applying the Voronoi summation formula (Lemma 1) to the n-sum, this is further equal to

$$\frac{1}{d}\sum_{r|d}\frac{N}{fgr}\sum_n S\big(\overline{fg}m,n,r\big)\lambda_2(n)\mathring{V}_2\left(\frac{nN}{fgr^2}\right).$$

Inserting this transformed sum back into (4.1), applying the Cauchy-Schwarz inequality to the m-sum, and using (2.6), we obtain

$$\begin{aligned} S_{N,M,d,q} &\ll \frac{1}{N^{1/2}} \sum_{g|f|q'} \mu^2(f) \Big|\lambda_2\Big(\frac{f}{g}\Big)\Big| \sum_{r|d} \frac{N}{fgr} \Big(\sum_{\substack{m \asymp M \\ (m,q)=1}} \Big| \sum_n S(\overline{fg}m, n, r) \lambda_2(n) \mathring{V}_2\Big(\frac{nN}{fgr^2}\Big)\Big|^2\Big)^{1/2} \\ &\ll \frac{1}{N^{1/2}} \sum_{g|f|q'} \mu^2(f) \Big|\lambda_2\Big(\frac{f}{g}\Big)\Big| \sum_{r|d} \frac{N}{fgr} \Big(\sum_{n_1,n_2 \ll fgr^2q^\varepsilon/N} |\lambda_2(n_1)\lambda_2(n_2) \mathcal{S}_M(\overline{fg}n_1, \overline{fg}n_2, r)|\Big)^{1/2} + q^{-10} \end{aligned}$$

by the rapid decay of $\mathring{V}_2$ (recall (3.5)), where

$$\mathcal{S}_M(n_1, n_2, r) = \sum_{\substack{m \asymp M \\ (m,q)=1}} S(m, n_1, r) S(m, n_2, r). \tag{4.2}$$

(This depends also on q, but this is not displayed in the notation.) Applying (2.7), we obtain our basic estimate

$$S_{N,M,d,q} \ll \frac{q^\varepsilon}{N^{1/2}} \sum_{g|f|q'} \sum_{r|d} \frac{N}{fgr} \Big(\sum_{n_1,n_2 \ll fgr^2q^\varepsilon/N} |\mathcal{S}_M(\overline{fg}n_1, \overline{fg}n_2, r)|\Big)^{1/2}. \tag{4.3}$$

Now let q_1 be a divisor of q with $(q, 6^\infty) \mid q_1$ and write $s = (r, q_1)$. Then in particular $(r, 6^\infty) \mid s$. Applying Theorem 5, we can bound $S_{N,M,d,q}$ by

$$\ll q^\varepsilon N^{\frac{1}{2}} \sum_{g|f|q'} \sum_{r|d} \frac{1}{fgr} \Big(\sum_{n_1,n_2 \ll fgr^2q^\varepsilon/N} M^{1/2} q q_1^{1/2} + \frac{M^{1/2} q^{5/4}}{(r,q_1)^{1/4}} + M q^{3/4} (q, n_1 - n_2)^{1/4} q_1^{1/4} + q q_1 + q^{11/8} q_1^{1/8}\Big)^{1/2},$$

and Proposition 6 follows.

5. Automorphic Preliminaries II

Unfortunately not all cusp forms are newforms. An L^2-basis $\mathcal{B}_k(\ell)$ for the finite-dimensional vector space $S_k(\ell)$, the space of holomorphic cusp forms of weight k and level ℓ, and an L^2-basis $\mathcal{B}(\ell, t)$ for $\mathcal{A}(\ell, t)$, the space of Maaß forms of level ℓ and spectral parameter t, will in general also include oldforms. We describe the procedure in detail for Maaß forms, the holomorphic case requires only small notational changes. For $\ell_1 \mid \ell$ let $\mathcal{B}^*(\ell_1, \ell, t) \subseteq \mathcal{B}(\ell, t)$ denote the set of all $L^2(\Gamma_0(\ell)\backslash\mathbb{H})$-normalized newforms of level ℓ_1 and spectral parameter t and write $f|_d(z) := f(dz)$. Then by newform theory we have

$$\mathcal{A}(\ell, t) = \bigoplus_{\ell_1|\ell} \bigoplus_{f \in \mathcal{B}^*(\ell_1,\ell,t)} \bigoplus_{d|\frac{\ell}{\ell_1}} f|_d \cdot \mathbb{C}. \tag{5.1}$$

The first two sums are orthogonal; the last one is, in general, not orthogonal and needs to be orthogonalized by Gram-Schmidt. In this way we get an orthogonal basis $\mathcal{B}(\ell, t)$ of $\mathcal{A}(\ell, t)$, and we collect all spectral parameters to obtain $\mathcal{B}(\ell) := \coprod_t \mathcal{B}(\ell, t)$, and correspondingly

$$\mathcal{B}^*(\ell_1, \ell) := \coprod_t \mathcal{B}^*(\ell_1, \ell, t).$$

The Fourier coefficients of the forms in the bases $\mathcal{B}_k(\ell)$ and $\mathcal{B}(\ell)$ are not exactly multiplicative, but almost so. More precisely [BHM, p. 74], if $m = qm' \in \mathbb{N}$ with $(m', q) = 1$, then

$$\sqrt{m}\rho_f(m) = \sum_{d|(\ell, q/(q,\ell))} \mu(d)\chi_0(d)\lambda_{f^*}\Big(\frac{q}{d(q,\ell)}\Big)\Big(\frac{(\ell,q)m'}{d}\Big)^{1/2} \rho_f\Big(\frac{(\ell,q)m'}{d}\Big) \tag{5.2}$$

where f^* is the underlying newform. In particular, if $(q, \ell) = 1$, then

$$\sqrt{m}\rho_f(m) = \lambda_{f^*}(q)\sqrt{m'}\rho_f(m'). \tag{5.3}$$

Moreover, if f^* satisfies the Ramanujan conjecture and a_m is any finite sequence of complex numbers supported on integers $m = qm'$ with $(m', q) = 1$, then

$$\Big|\sum_m a_m\sqrt{m}\rho_f(m)\Big|^2 \leqslant \tau(q)^2 \sum_{d|(q,\ell)} \Big|\sum_{m'} a_{qm'}\sqrt{dm'}\rho_f(dm')\Big|^2. \tag{5.4}$$

A somewhat involved explicit calculation shows a similar result [BHM, p. 80] for the coefficients $\rho_{\mathfrak{a}}(m,t)$ of Eisenstein series: if $q \in \mathbb{N}$ and a_m is any finite sequence of complex numbers supported on integers $m = qm'$ with $(m', q) = 1$, then

$$\sum_{\mathfrak{a}}\Big|\sum_m a_m\sqrt{m}\rho_{\mathfrak{a}}(m,t)\Big|^2 \leqslant 9\tau(\ell)^3\tau(q)^4 \sum_{d|(q,\ell)}\sum_{\mathfrak{a}}\Big|\sum_{m'} a_{qm'}\sqrt{dm'}\rho_{\mathfrak{a}}(dm',t)\Big|^2 \tag{5.5}$$

for all $t \in \mathbb{R}$.

The relation (5.2) is very useful, but not sufficient for all our purposes. We proceed to make the orthogonalization process in (5.1) explicit. For a newform $f \in \mathcal{B}^*(\ell_1, \ell)$ we define the following arithmetic functions:

$$r_f(c) := \sum_{b|c} \frac{\mu(b)\lambda_f(b)^2}{b\cdot\sigma_{-1}(b)^2}, \quad \alpha(c) := \sum_{b|c}\frac{\mu(b)}{b^2}, \quad \beta(c) = \sum_{b|c}\frac{\mu^2(b)}{b},$$

$$\mu_f(c) \text{ given by } L(f,s)^{-1} = \sum_c \frac{\mu_f(c)}{c^s}, \text{ so } \mu_f(p) = -\lambda_f(p),\ \mu_f(p^2) = \chi_0(p),\ \mu_f(p^\nu) = 0,\ \nu > 2,$$

where $\sigma_{-1}(b)$ is the sum of the reciprocal divisors of b and χ_0 is the trivial character modulo ℓ_1. For $d \mid g$ define

$$\xi'_g(d) := \frac{\mu(g/d)\lambda_f(g/d)}{r_f(g)^{1/2}(g/d)^{1/2}\beta(g/d)}, \quad \xi''_g(d) = \frac{\mu_f(g/d)}{(g/d)^{1/2}(r_f(g)\alpha(g))^{1/2}}.$$

Write uniquely $g = g_1g_2$ where g_1 is squarefree, g_2 is squarefull, and $(g_1, g_2) = 1$. Then for $d \mid g$ define

$$\xi_g(d) = \xi'_{g_1}((g_1,d))\xi''_{g_2}((g_2,d)) \ll g^\varepsilon(g/d)^{\theta-1/2}. \tag{5.6}$$

The following lemma is an extension of [ILS, Section 2] to non-squarefree levels. It is essentially contained [Ro, Proposition 5]. As this result is crucial for us, and the assumptions are a little different from [Ro], we provide a complete proof.

Lemma 2. *Let $\ell_1 \mid \ell$, and let $f^* \in \mathcal{B}^*(\ell_1, \ell) \subseteq \mathcal{B}(\ell)$ be an $L^2(\Gamma_0(\ell)\backslash\mathbb{H})$-normalized newform of level ℓ_1. Then the set of functions*

$$\Big\{ f^{(g)} := \sum_{d|g} \xi_g(d) f^*|_d \ :\ g \mid \frac{\ell}{\ell_1}\Big\}$$

is an orthonormal basis of the space $\bigoplus_{d|\frac{\ell}{\ell_1}} f^*|_d \cdot \mathbb{C}$.

If f is any member in this basis, then its Fourier coefficients satisfy the bound

$$\sqrt{n}\rho_f(n) \ll (n\ell)^\varepsilon n^\theta(\ell, n)^{1/2-\theta}|\rho_{f^*}(1)|. \tag{5.7}$$

Remark: We stress that f^* of level ℓ_1 is normalized as in (2.2), i.e. with respect to the group $\Gamma_0(\ell)$. The map $\mathcal{B}^*(\ell_1,\ell) \to \mathcal{B}^*(\ell_1,\ell_1) \subseteq \mathcal{B}(\ell_1)$ is not an isometry, but reduces the norm by a factor $[\Gamma_0(\ell_1) : \Gamma_0(\ell)]^{-1/2}$.

Although we do not need it in the present paper, we remark that with the definition $f|_d(z) := d^{k/2}f(dz)$ the same construction (and the same proof) works for holomorphic cusp forms of weight k, and in particular the bound (5.7) remains true with $\theta = 0$ for holomorphic cusp forms. Moreover, the trivial character χ_0 modulo ℓ_1 plays no special role, the same construction and the same proof work for any Dirichlet character χ modulo ℓ_1.

Proof. We write $\tilde{\ell} := \ell/\ell_1$. As a first step we need to compute the Gram matrix $(\langle f^*|_{d_1}, f^*|_{d_2}\rangle)_{d_1,d_2|\tilde{\ell}}$ where all inner products are as in (2.2). Write $d_1' = d_1/(d_1,d_2)$, $d_2' = d_2/(d_1,d_2)$. As in [ILS] we apply Rankin-Selberg theory. First we observe that $\langle f^*|_{d_1}, f^*|_{d_2}\rangle = \langle f^*|_{d_1'}, f^*|_{d_2'}\rangle$ since multiplication by a scalar (d_1,d_2) is an isometry. Let $E(z,s)$ be the standard non-holomorphic Eisenstein series of level ℓ. Then we unfold and use (2.1) and (2.3) to obtain

$$\langle E(\cdot,s)f^*|_{d_1'}, f^*|_{d_2'}\rangle = \int_0^\infty \int_0^1 y^s f^*(d_1'z)\bar{f}^*(d_2'z)\frac{dx\,dy}{y^2} = 2\sum_{n=1}^\infty \frac{\lambda_{f^*}(d_2'n)\lambda_{f^*}(d_1'n)}{(d_1'd_2')^{s-1/2}n^s}\int_0^\infty y^s|W_{0,it}(y)|^2\frac{dy}{y^2}.$$

We use (2.5) to evaluate the Dirichlet series

$$\sum_n \lambda_{f^*}(d_1'n)\lambda_{f^*}(d_2'n)n^{-s} = \sum_{(n,d_1'd_2')=1}\lambda_{f^*}(n)^2n^{-s}\prod_{p^{e_p}\|d_1'd_2'}\sum_{\nu=0}^\infty \lambda_{f^*}(p^{\nu+e_p})\lambda_{f^*}(p^\nu)p^{-\nu s}$$

and compare residues on both sides at $s=1$. In this way we obtain

$$\langle f^*|_{d_1}, f^*|_{d_2}\rangle = \langle f^*|_{d_1'}, f^*|_{d_2'}\rangle = A(d_1'd_2')\langle f^*, f^*\rangle = A\left(\frac{\mathrm{lcm}(d_1,d_2)}{\gcd(d_1,d_2)}\right)\langle f^*, f^*\rangle,$$

where A is the multiplicative function given by

$$A(p) = \frac{\lambda_{f^*}(p)}{\sqrt{p}(1+1/p)}, \quad A(p^{\nu+1}) = \frac{\lambda_{f^*}(p)}{\sqrt{p}}A(p^\nu) - \frac{\chi_0(p)}{p}A(p^{\nu-1}). \tag{5.8}$$

(Here again χ_0 is the trivial character modulo ℓ_1.) We need to verify that

$$\sum_{d_1|g_1}\sum_{d_2|g_2}\xi_{g_1}(d_1)\xi_{g_2}(d_2)A\left(\frac{\mathrm{lcm}(d_1,d_2)}{\gcd(d_1,d_2)}\right) = \delta_{g_1=g_2}.$$

By multiplicativity and symmetry it is enough to consider the case $g_1 = p^\alpha$, $g_2 = p^\beta$ for a prime p and $\beta \geqslant \alpha \geqslant 0$, so that it suffices to verify

$$I(\alpha,\beta) := \sum_{\delta_1\leqslant\alpha}\sum_{\delta_2\leqslant\beta}\xi_{p^\alpha}(p^{\delta_1})\xi_{p^\beta}(p^{\delta_2})A(p^{|\delta_1-\delta_2|}) = \delta_{\alpha=\beta}.$$

For prime powers, the arithmetic function $\xi_g(d)$ simplifies as follows:

$$\xi_1(1) = 1, \qquad \xi_p(p) = r_{f^*}(p)^{-1/2}, \quad \xi_p(1) = \frac{-\lambda_{f^*}(p)}{\sqrt{p}(1+1/p)}\xi_p(p),$$

$$\xi_{p^\nu}(p^\nu) = \big(r_{f^*}(p)(1-p^{-2})\big)^{-1/2}, \quad \xi_{p^\nu}(p^{\nu-1}) = \frac{-\lambda_{f^*}(p)}{\sqrt{p}}\xi_{p^\nu}(p^\nu), \quad \xi_{p^\nu}(p^{\nu-2}) = \frac{\chi_0(p)}{p}\xi_{p^\nu}(p^\nu), \quad \nu \geqslant 2,$$

and $\xi_{p^a}(p^b) = 0$ in all other cases. In particular, for $\nu \geqslant 2$ and $c \leqslant \nu$, the value $\xi_{p^\nu}(p^{\nu-c})$ depends only on p and c, but not on ν. Hence

$$I(\alpha,\beta) = I(\alpha+c,\beta+c) \tag{5.9}$$

for any $c \in \mathbb{N}$ and any $2 \leqslant \alpha \leqslant \beta$, and by the recurrence relation in (5.8) we also have

$$I(\alpha,\beta+1) = \frac{\lambda_{f^*}(p)}{\sqrt{p}}I(\alpha,\beta) - \frac{\chi_0(p)}{p}I(\alpha,\beta-1) \tag{5.10}$$

if $\beta \geqslant \alpha+3$ (this condition is needed to ensure that the summation indices δ_1,δ_2 satisfy $\delta_2-\delta_1 \geqslant 0$ in all arising sums). By (5.9), it suffices to assume $\alpha \leqslant 2$, and by (5.10) it suffices to assume $\beta-\alpha \leqslant 3$; the rest follows by induction. This leaves us with the 12 cases $0 \leqslant \alpha \leqslant \beta \leqslant \alpha+3 \leqslant 5$, which are straightforward to verify.

The bound (5.7) now follows from

$$\rho_{f^{(g)}}(n) = \sum_{d|g} \xi_g(d)\rho_{f^*}(n/d)$$

(with the convention $\rho(x) = 0$ for $x \notin \mathbb{Z}$), (2.3), (2.8), and (5.6). □

We define the following integral transforms for a smooth function $\phi : [0,\infty) \to \mathbb{C}$ satisfying $\phi(0) = \phi'(0) = 0$, $\phi^{(j)}(x) \ll (1+x)^{-3}$ for $0 \leqslant j \leqslant 3$:

$$\begin{aligned} \dot{\phi}(k) &= 4i^k \int_0^\infty \phi(x) J_{k-1}(x) \frac{dx}{x}, \\ \tilde{\phi}(t) &= 2\pi i \int_0^\infty \phi(x) \frac{J_{2it}(x) - J_{-2it}(x)}{\sinh(\pi t)} \frac{dx}{x}, \\ \check{\phi}(t) &= 8 \int_0^\infty \phi(x) \cosh(\pi t) K_{2it}(x) \frac{dx}{x}. \end{aligned} \tag{5.11}$$

With the already established notation, the following spectral sum formula holds (see e.g. [BHM, Theorem 2]).

Lemma 3. *[Kuznetsov formula] Let ϕ be as in the previous paragraph, and let $a, b, \ell > 0$ be integers. Then,*

$$\begin{aligned} \sum_{\ell|c} \frac{1}{c} S(a,b,c)\phi\left(\frac{4\pi\sqrt{ab}}{c}\right) &= \sum_{\substack{k \geqslant 2 \\ k \text{ even}}} \sum_{f \in \mathcal{B}_k(\ell)} \dot{\phi}(k)\Gamma(k)\sqrt{ab}\rho_f(a)\rho_f(b) \\ &\quad + \sum_{f \in \mathcal{B}(\ell)} \tilde{\phi}(t_f) \frac{\sqrt{ab}}{\cosh(\pi t_f)} \rho_f(a)\rho_f(b) \\ &\quad + \frac{1}{4\pi} \sum_{\mathfrak{a}} \int_{-\infty}^{\infty} \tilde{\phi}(t) \frac{\sqrt{ab}}{\cosh(\pi t)} \rho_{\mathfrak{a}}(a,t)\rho_{\mathfrak{a}}(b,t)dt \end{aligned}$$

and

$$\begin{aligned} \sum_{\ell|c} \frac{1}{c} S(a,-b,c)\phi\left(\frac{4\pi\sqrt{ab}}{c}\right) &= \sum_{f \in \mathcal{B}(\ell)} \check{\phi}(t_f) \frac{\sqrt{ab}}{\cosh(\pi t_f)} \rho_f(a)\rho_f(-b) \\ &\quad + \frac{1}{4\pi} \sum_{\mathfrak{a}} \int_{-\infty}^{\infty} \check{\phi}(t) \frac{\sqrt{ab}}{\cosh(\pi t)} \rho_{\mathfrak{a}}(a,t)\rho_{\mathfrak{a}}(-b,t)dt. \end{aligned}$$

Often the Kuznetsov formula is used hand in hand with the large sieve inequalities of Deshouillers-Iwaniec [DI].

Lemma 4. *[Spectral large sieve] Let $T, M \geqslant 1$, $\ell \in \mathbb{N}$, and let (a_m), $M \leqslant m \leqslant 2M$, be a sequence of complex numbers. Then all three quantities*

$$\sum_{\substack{2 \leqslant k \leqslant T \\ k \text{ even}}} \Gamma(k) \sum_{f \in \mathcal{B}_k(\ell)} \Big|\sum_m a_m \sqrt{m}\rho_f(m)\Big|^2, \quad \sum_{\substack{f \in \mathcal{B}(\ell) \\ |t_f| \leqslant T}} \frac{1}{\cosh(\pi t_f)} \Big|\sum_m a_m \sqrt{m}\rho_f(\pm m)\Big|^2,$$

$$\sum_{\mathfrak{a}} \int_{-T}^{T} \frac{1}{\cosh(\pi t)} \Big|\sum_m a_m \sqrt{m}\rho_{\mathfrak{a}}(\pm m, t)\Big|^2 dt$$

are bounded by

$$M^{\varepsilon}\left(T^2 + \frac{M}{\ell}\right) \sum_m |a_m|^2.$$

Another application of the Kuznetsov formula is the following bound.

Lemma 5. *Let $T \geqslant 1$, $m, \ell \in \mathbb{N}$. Then*

$$\sum_{\substack{|t_f|\leqslant T\\ f\in\mathcal{B}(\ell)}} \frac{1}{\cosh(\pi t_f)}|\sqrt{m}\rho_f(m)|^2 \ll \left(T^2 + \frac{(\ell,m)^{1/2}m^{1/2}}{\ell}\right)(Tm)^{\varepsilon}$$

with an implied constant depending only on ε.

Proof. This is [Mot, Lemma 2.4] for $\ell = 1$, and the proof in the more general case is verbatim the same, except that in [Mot, (2.3.7), (2.3.10)] an additional divisibility condition is added in the sum over Kloosterman sums that leads to an obvious modification of the last two displays in the proof. □

The following important result will be used to avoid the Ramanujan conjecture.

Theorem 8. *Let $\ell, s \in \mathbb{N}$, $R, T \geqslant 1$, and let $\alpha(r)$, $R \leqslant r \leqslant 2R$, be any sequence of complex numbers with $|\alpha(r)| \leqslant 1$. Then*

$$\sum_{\substack{|t_f|\leqslant T\\ f\in\mathcal{B}(\ell)}} \frac{1}{\cosh(\pi t_f)}\Big|\sum_{\substack{R\leqslant r\leqslant 2R\\ (r,s\ell)=1}} \alpha(r)\sqrt{rs}\rho_f(rs)\Big|^2 \ll (\ell s T R)^{\varepsilon}(\ell,s)\left(T + \frac{s^{1/2}}{\ell^{1/2}}\right)\left(T + \frac{R}{\ell^{1/2}}\right)R.$$

Proof. We call the left hand side Ξ. Fix an $f \in \mathcal{B}(\ell)$ and denote by $f^* \in \mathcal{B}^*(\ell_1, \ell)$ the underlying newform of level ℓ_1, say. An application of (5.2) and (5.3) shows for $(r, s\ell) = 1$ that

$$\begin{aligned}\sqrt{rs}\rho_f(rs) &= \sum_{\delta|(\ell, \frac{s}{(s,\ell)})} \mu(\delta)\chi_0(\delta)\lambda_{f^*}\left(\frac{s}{\delta(\ell,s)}\right)\left(\frac{(\ell,s)r}{\delta}\right)^{1/2}\rho_f\left(\frac{(\ell,s)r}{\delta}\right)\\ &= \sum_{\delta|(\ell, \frac{s}{(s,\ell)})} \mu(\delta)\chi_0(\delta)\lambda_{f^*}\left(\frac{s}{\delta(\ell,s)}\right)\left(\frac{(\ell,s)}{\delta}\right)^{1/2}\rho_f\left(\frac{(\ell,s)}{\delta}\right)\lambda_{f^*}(r).\end{aligned}$$

We apply the Cauchy-Schwarz inequality first to the sum over δ and then to the sum over $f \in \mathcal{B}(\ell_1\ell_2)$ to obtain

$$\Xi \leqslant \tau(s)^{1/2}\Theta_2^{1/2}\sum_{\delta|(\ell, \frac{s}{(s,\ell)})}\Theta_1^{1/2}$$

where

$$\Theta_1 = \sum_{\substack{f\in\mathcal{B}(\ell)\\ |t_f|\leqslant T}} \frac{1}{\cosh(\pi t_f)^2}\Big|\lambda_{f^*}\left(\frac{s}{\delta(\ell,s)}\right)\left(\frac{(\ell,s)}{\delta}\right)^{1/2}\rho_f\left(\frac{(\ell,s)}{\delta}\right)\Big|^4$$

and

$$\Theta_2 = \sum_{\substack{f\in\mathcal{B}(\ell)\\ |t_f|\leqslant T}}\Big|\sum_{\substack{R\leqslant r\leqslant 2R\\ (r,s\ell)=1}}\alpha(r)\lambda_{f^*}(r)\Big|^4.$$

The main idea is to transform the sums Θ_1 and Θ_2 into sums to which Lemma 5 and Lemma 4, respectively, may be applied. By a crude application of (5.7), the Möbius inverse of (2.5) and (2.3) we have

$$\begin{aligned}\Theta_1 &\ll (\ell,s)^2\ell^\varepsilon \sum_{\substack{f\in\mathcal{B}(\ell)\\ |t_f|\leqslant T}} \frac{1}{\cosh(\pi t_f)^2}\Big|\lambda_{f^*}\Big(\frac{s}{\delta(\ell,s)}\Big)\rho_{f^*}(1)\Big|^4\\ &\leqslant \tau(s)(\ell,s)^2\ell^\varepsilon \sum_{g|\frac{s}{\delta(\ell,s)}}\sum_{\substack{f\in\mathcal{B}(\ell)\\ |t_f|\leqslant T}} \frac{|\rho_{f^*}(1)|^4}{\cosh(\pi t_f)^2}\Big|\lambda_{f^*}\Big(\frac{s^2}{(g\delta(\ell,s))^2}\Big)\Big|^2\\ &= \tau(s)(\ell,s)^2\ell^\varepsilon \sum_{g|\frac{s}{\delta(\ell,s)}}\sum_{\substack{f\in\mathcal{B}(\ell)\\ |t_f|\leqslant T}} \frac{|\rho_{f^*}(1)|^2}{\cosh(\pi t_f)^2}\Big|\frac{s}{g\delta(\ell,s)}\rho_{f^*}\Big(\frac{s^2}{(g\delta(\ell,s))^2}\Big)\Big|^2.\end{aligned}$$

The newform $f^*\in\mathcal{B}^*(\ell_1,\ell)$ is counted $\tau(\ell/\ell_1)$ times in the sum over $\mathcal{B}(\ell)$, and we sum now over $L^2(\Gamma_0(\ell_1)\backslash\mathbb{H})$-normalized newforms $f\in\mathcal{B}^*(\ell_1,\ell_1)\subseteq\mathcal{B}(\ell_1)$ which by the remark following Lemma 2 leads to a renormalizing factor $(\ell/\ell_1)^{-2+o(1)}$. Hence by (2.4) we conclude

$$\begin{aligned}\Theta_1 &\ll \tau(s)(\ell,s)^2\ell^\varepsilon \sum_{g|\frac{s}{\delta(\ell,s)}}\sum_{\ell_1|\ell}\frac{\tau(\ell/\ell_1)}{(\ell/\ell_1)^2}\sum_{\substack{f\in\mathcal{B}^*(\ell_1,\ell_1)\\ |t_f|\leqslant T}} \frac{|\rho_f(1)|^2}{\cosh(\pi t_f)^2}\Big|\frac{s}{g\delta(\ell,s)}\rho_f\Big(\frac{s^2}{(g\delta(\ell,s))^2}\Big)\Big|^2\\ &\ll \frac{(\ell,s)^2}{\ell}(\ell sT)^\varepsilon \sum_{g|\frac{s}{\delta(\ell,s)}}\sum_{\ell_1|\ell}\frac{1}{\ell/\ell_1}\sum_{\substack{f\in\mathcal{B}^*(\ell_1,\ell_1)\\ |t_f|\leqslant T}} \frac{1}{\cosh(\pi t_f)}\Big|\frac{s}{g\delta(\ell,s)}\rho_f\Big(\frac{s^2}{(g\delta(\ell,s))^2}\Big)\Big|^2.\end{aligned}$$

By positivity we can extend the innermost sum to all of $\mathcal{B}(\ell_1)$. By Lemma 5 we finally obtain

$$\Theta_1 \ll (\ell sTR)^\varepsilon\frac{(\ell,s)^2}{\ell}\sum_{\ell_1|\ell}\frac{1}{\ell/\ell_1}\left(T^2+\frac{s\big(\ell_1,s^2/(s,\ell)^2\big)^{1/2}}{(\ell,s)\ell_1}\right)\leqslant(\ell sTR)^\varepsilon\frac{(\ell,s)^2}{\ell}\left(T^2+\frac{s}{\ell}\right).$$

Next we turn to the estimation of Θ_2. By a similar argument we have

$$\begin{aligned}\Theta_2 &= \sum_{\ell_1|\ell}\tau(\ell/\ell_1)\sum_{\substack{f\in\mathcal{B}^*(\ell_1,\ell_1)\\ |t_f|\leqslant T}}\Big|\sum_{\substack{R\leqslant r\leqslant 2R\\ (r,s\ell)=1}}\alpha(r)\lambda_f(r)\Big|^4\\ &= \sum_{\ell_1|\ell}\tau(\ell/\ell_1)\sum_{\substack{f\in\mathcal{B}^*(\ell_1,\ell_1)\\ |t_f|\leqslant T}}\Big|\sum_{\substack{R\leqslant r,r'\leqslant 2R\\ (rr',s\ell)=1}}\alpha(r)\alpha(r')\sum_{g|(r,r')}\lambda_f\Big(\frac{rr'}{g^2}\Big)\Big|^2\\ &\ll (\ell T)^\varepsilon\sum_{\ell_1|\ell}\ell_1\sum_{\substack{f\in\mathcal{B}^*(\ell_1,\ell_1)\\ |t_f|\leqslant T}}\frac{1}{\cosh(\pi t_f)}\Big|\sum_{\substack{R\leqslant r,r'\leqslant 2R\\ (rr',s\ell)=1}}\alpha(r)\alpha(r')\sum_{g|(r,r')}\Big(\frac{rr'}{g^2}\Big)^{1/2}\rho_{f^*}\Big(\frac{rr'}{g^2}\Big)\Big|^2\\ &= (\ell T)^\varepsilon\sum_{\ell_1|\ell}\ell_1\sum_{\substack{f\in\mathcal{B}^*(\ell_1,\ell_1)\\ |t_f|\leqslant T}}\frac{1}{\cosh(\pi t_f)}\Big|\sum_{r\ll R^2}\sqrt{r}\rho_f(r)\beta(r)\Big|^2\end{aligned}$$

where

$$\beta(r)=\sum_{\substack{R\leqslant r_1,r_2\leqslant 2R\\ (r_1r_2,s\ell)=1}}\alpha(r_1)\alpha(r_2)\sum_{\substack{g|(r_1,r_2)\\ r_1r_2=g^2r}}1\ll\sum_{g\ll R/\sqrt{r}}\tau(r)\ll\frac{R^{1+\varepsilon}}{\sqrt{r}}.$$

Again we complete the sum over f to all of $\mathcal{B}(\ell_1)$. The large sieve (Lemma 4) shows

$$\Theta_2 \ll (\ell TR)^{\varepsilon} \sum_{\ell_1 | \ell} \ell_1 \left(T^2 + \frac{R^2}{\ell_1} \right) R^2,$$

and the lemma follows. □

Remark. The important step in the proof in the application of the Cauchy-Schwarz inequality. A simpler strategy would apply (5.3) with $r = q$, $s = m'$ directly, estimate $\sqrt{s}\rho_f(s)$ by (5.7) and apply the large sieve to obtain

$$\sum_{\substack{|t_f| \leqslant T \\ f \in \mathcal{B}(\ell)}} \frac{1}{\cosh(\pi t_f)} \Big| \sum_{\substack{R \leqslant r \leqslant 2R \\ (r, s\ell)=1}} \alpha(r) \sqrt{rs} \rho_f(rs) \Big|^2 \ll (\ell s T R)^{\varepsilon} s^{2\theta} (\ell, s)^{1-2\theta} \left(T^2 + \frac{R}{\ell} \right) R. \tag{5.12}$$

6. Bessel functions

We collect here some useful formulas for future reference. In view of the integral transform appearing in the Kuznetsov formula we write

$$\begin{aligned} \mathcal{J}^+_{2it}(x) &:= \pi i \frac{J_{2it}(x) - J_{-2it}(x)}{\sinh(\pi t)}, \\ \mathcal{J}^-_{2it}(x) &:= 4 \cosh(\pi t) K_{2it}(x). \end{aligned} \tag{6.1}$$

We start with the power series expansion [GR, 8.402]

$$J_\nu(x) = \frac{x^\nu}{2^\nu} \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{2^{2k} k! \Gamma(\nu + k + 1)} \tag{6.2}$$

valid for $x > 0$ and $\nu \in \mathbb{C}$. Next, we record the uniform asymptotic expansion [EMOT, 7.13(17)]

$$\frac{J_{it}(x)}{\sinh(\pi t/2)} = \exp\left(i\sqrt{t^2 + x^2} - it \operatorname{arcsinh}(t/x) \right) \mathcal{J}_M(t, x) + \mathrm{O}\left((x+t)^{-M} \right) \tag{6.3}$$

for $t > 1$ and any fixed $M \in \mathbb{N}$, where $\mathcal{J}_M(t, x)$ satisfies

$$x^j \frac{\partial^j}{\partial x^j} \mathcal{J}_M(t, x) \ll_{M,j} (t + x)^{-1/2}$$

for any $j \in \mathbb{N}_0$. The original error term in [EMOT] is only $\mathrm{O}(x^{-M})$ in place of $\mathrm{O}((x+t)^{-M})$, but the stronger error term follows from the power series expansion (6.2) for $x < t^{1/3}$. A similar expansion holds for $J_{-it}(x) = \overline{J_{it}(x)}$. By [GR, 8.411.1] we have

$$J_{k-1}(x) = \frac{1}{\pi} \int_0^\pi \cos((k-1)\xi - x \sin \xi) d\xi. \tag{6.4}$$

for $k \in \mathbb{N}$, and by [GR, 6.561.16] we have

$$\widehat{\mathcal{J}}^-_{2it}(s) = \cosh(\pi t) 2^{s-2} \Gamma\left(\frac{s}{2} + it \right) \Gamma\left(\frac{s}{2} - it \right), \quad \Re s > 2 |\Im t|.$$

In particular, for $\Re s = 1$, we have the bound

$$\widehat{\mathcal{J}}^-_{2it}(1 + i\tau) \ll e^{-\pi \max(0, \frac{|\tau|}{2} - |t|)}. \tag{6.5}$$

Lemma 6. *Let $k \in \mathbb{N}$, $t \in \mathbb{R} \cup (-i/4, i/4)$, $x > 0$. Then*

$$\begin{aligned} \mathcal{J}^+_{2it}(x) &\ll x^{-1/2}, \\ J_{k-1}(x) &\ll x^{-1/2}, \quad x > 100k, \end{aligned} \tag{6.6}$$

with absolute implied constants. Moreover, for fixed $\nu \in \mathbb{C}$ and $j \in \mathbb{N}_0$, we have

$$\frac{d^j}{dx^j} J_\nu(x) \ll_{\nu,j} \begin{cases} x^{\Re\nu - j}, & x \leqslant 1, \\ x^{-1/2}, & x \geqslant 1. \end{cases} \tag{6.7}$$

Proof. The bound for $J_{k-1}(x)$ follows from [Ra, Lemma 4.2, 4.3] for $k \geqslant 16$, while for $k < 16$ the bound is a trivial consequence of the asymptotic formula [GR, 8.451.1]. The bound for $\mathcal{J}^+_{2it}(x)$ for $x \geqslant 1$ follows from (6.3) and for $x < 1$ from the power series expansion (6.2). This proves (6.6) The bound (6.7) follows similarly from (6.2) and [GR, 8.451.1]. □

Lemma 7. *Let $\nu \in \mathbb{C}$ with $\Re\nu \geqslant 0$ be fixed. There exist smooth functions $F^\pm_\nu(x)$ such that*

$$x^j (F^\pm_\nu)^{(j)}(x) \ll_{\nu,j} \min(x^{\Re\nu}, x^{-1/2}) \tag{6.8}$$

for all $j \in \mathbb{N}_0$ and

$$J_\nu(x) = F^+_\nu(x) e^{ix} + F^-_\nu(x) e^{-ix}. \tag{6.9}$$

Proof. The idea is to use the asymptotic formula for $x \geqslant 1$ and a trivial decomposition for $x < 1$ and then to glue these decompositions together. To make this precise, we define $H^{(1)}_\nu(x) = J_\nu(x) + iY_\nu(x)$ and $H^{(2)}_\nu(x) = J_\nu(x) - iY_\nu(x)$ as in [GR, (8.405)] and write

$$H^+_\nu(x) = H^{(1)}_\nu(x) e^{-ix}, \quad H^-_\nu(x) = H^{(2)}_\nu(x) e^{ix}.$$

By [GR, 8.476.10] we have $\overline{H^+_\nu(x)} = H^-_{\bar\nu}(x)$ for $x \in \mathbb{R}$. Then,

$$J_\nu(x) = \frac{1}{2}\left(H^+_\nu(x) e^{ix} + H^-_\nu(x) e^{-ix}\right)$$

by [GR, 8.481]. Finally, we choose a smooth function V with support in $[1,\infty)$ and $V(x) = 1$ on $[2,\infty)$ and define

$$F^+_\nu(x) := \frac{1}{2} H^+_\nu(x) V(x) + e^{-ix} J_\nu(x)(1 - V(x)), \quad F^-_\nu(x) := \frac{1}{2} H^-_\nu(x) V(x),$$

so that (6.9) holds.

We compute the derivatives of $H^\pm_\nu(x)$ for $x \geqslant 1$ using the integral representation ([GR, 8.421.9])

$$H^+_\nu(x) = \left(\frac{2}{\pi x}\right)^{1/2} \frac{e(-\frac{2\nu+1}{8})}{\Gamma(\nu + 1/2)} \int_0^\infty \left(1 + \frac{it}{2x}\right)^{\nu - 1/2} t^{\nu - 1/2} e^{-t} dt$$

and the derivatives of $e^{-ix} J_\nu(x)$ for $x \leqslant 2$ using (6.7). This implies (6.8). □

The next lemma shows when the integral transforms of the Kuznetsov formula are negligibly small.

Lemma 8. *Let $Z \geqslant 1$, $X, P, \alpha > 0$, and let $C \geqslant Z + X + P + \alpha$ be a large parameter. Let Ω be a smooth weight function of fixed compact support satisfying $\Omega^{(j)}(x) \ll P Z^j$ for all $j \in \mathbb{N}_0$. Then the following bounds hold for any fixed $A > 0$.*

$$\int_0^\infty \Omega\left(\frac{x}{X}\right) e^{\pm i\alpha x} \mathcal{J}^+_{2it}(x) \frac{dx}{x} \ll |t|^{-A}, \quad \text{if} \quad t \geqslant C^\varepsilon Z(X\sqrt{\alpha^2 - 1} + X^{1/2} + 1), \quad \alpha \geqslant 1; \tag{6.10}$$

$$\int_0^\infty \Omega\left(\frac{x}{X}\right) e^{\pm i\alpha x} \mathcal{J}^-_{2it}(x) \frac{dx}{x} \ll |t|^{-A}, \quad \text{if} \quad t \geqslant C^\varepsilon Z(X + \alpha X + 1); \tag{6.11}$$

$$\int_0^\infty \Omega\left(\frac{x}{X}\right) e^{\pm i\alpha x} J_{k-1}(x) \frac{dx}{x} \ll k^{-A}, \quad \text{if} \quad k \geqslant C^\varepsilon Z(X^{1/2} + 1), \quad \alpha \geqslant 1. \tag{6.12}$$

Proof. This is essentially [J3, Lemma 3, Remark 1 & 2]. We give a variant of the proof in [J3].

By [BHM, (2.14)], all three bounds (6.10), (6.11), (6.12) hold if

$$t,\, k \geqslant C^\varepsilon (X + Z + \alpha X). \tag{6.13}$$

In particular (6.11) is proved, and also (6.10) if $\alpha \geqslant 2$, so in order to complete the proof of (6.10) we may assume $\alpha = 1+\beta$ with $0 \leqslant \beta \leqslant 1$, and then we may also assume $X \geqslant Z^2$, for otherwise the size condition in (6.10) implies (6.13). Hence the range for t not yet covered by (6.13) and the condition in (6.10) is contained in $C^\varepsilon \leqslant t \ll C^\varepsilon X$. We insert the uniform asymptotic formula (6.3) getting (up to an admissible error)

$$\int_0^\infty \Omega\left(\frac{x}{X}\right) \mathcal{J}_M(t,x) e^{if(x)} \frac{dx}{x}.$$

where

$$f(x) = \pm\alpha x \pm \left(\sqrt{(2t)^2+x^2} - 2t \operatorname{arcsinh}(2t/x)\right),$$
$$f'(x) = \pm\alpha \pm \frac{\sqrt{(2t)^2+x^2}}{x}, \quad f^{(j)}(x) \asymp \frac{t^2}{x^j\sqrt{t+x}} \quad (j \geqslant 2).$$

Under the present size assumptions an integration by parts argument as in [BKY, Lemma 8.1] with $U = X/Z$, $Q = X$, $Y = t^2/\sqrt{t+X}$ now shows the bound in (6.10) provided

$$\frac{\sqrt{t^2+X^2}}{X} - 1 \geqslant \beta + \left(\frac{Z}{X} + \frac{t}{X^{3/2}}\right) C^\varepsilon$$

which is implied by the assumption (observe that the left hand side is of order t^2/X^2).

Finally we prove (6.12). Since $J_k(x) \ll e^{k/5}$ for $x \leqslant k/2$ (see [Ra, Lemma 4.2]), we may assume $k \ll X$. In combination with our current assumption this implies $X \gg (C^\varepsilon Z)^2$ and $X^{1/2} \leqslant k \ll X$. We insert (6.4) getting

$$\int_0^\infty \Omega(x/X) e^{\pm i\alpha x} \int_{-\pi}^{\pi} \cos((k-1)\xi - x\sin\xi) d\xi \, \frac{dx}{x}.$$

Repeated integrating by parts in the x-integral shows (6.12) if $\alpha \geqslant 1 + C^\varepsilon Z/X$ (in particular if $\alpha \geqslant 2$). More precisely, we may extract smoothly the range $\sin\xi = \pm 1 + \mathrm{O}(C^\varepsilon Z/X)$ from the ξ-integral at the cost of an admissible error. In the remaining ξ-integral we integrate by parts sufficiently to complete the proof of (6.12). □

Lemma 9. *Let W be a fixed smooth function with support in $[1/2,3]$ satisfying $W^{(j)}(x) \ll_j 1$ for all j. Let $\nu \in \mathbb{C}$ be a fixed number with $\Re\nu \geqslant 0$. For $z, w > 0$ define*

$$W^*(z,w) = \int_0^\infty W(y) J_\nu(4\pi\sqrt{yw+z}) dy.$$

Fix $C \geqslant 1$ and $A, \varepsilon > 0$. Then for $z \gg w$ we have

$$W^*(z,w) = W_+(z,w)e(2\sqrt{z}) + W_-(z,w)e(-2\sqrt{z}) + \mathrm{O}_A(C^{-A}) \tag{6.14}$$

for suitable functions $W_\pm$ (depending on ν) satisfying

$$z^i w^j \frac{\partial^i}{\partial z^i}\frac{\partial^j}{\partial w^j} W_\pm(z,w) \begin{cases} = 0, & \sqrt{z}/w \leqslant C^{-\varepsilon}, \\ \ll C^{\varepsilon(i+j)} \min(z^{-1/4}, 1), & \text{otherwise.} \end{cases} \tag{6.15}$$

for any $i, j \in \mathbb{N}_0$. The implied constants depend on i, j and ν.

Proof. Integration by parts in connection with [GR, 8.472.3] (cf. (2.11)) yields

$$\int_0^\infty W(y) J_\nu(4\pi\sqrt{yw+z}) dy = \int_0^\infty \left(\frac{-\nu}{4\pi\sqrt{yw+z}} W(y) + \frac{\sqrt{yw+z}}{2\pi w} W'(y)\right) J_{\nu+1}(4\pi\sqrt{yw+z}) dy,$$

for $z, w > 0$. Repeated application together with (6.7) shows

$$W^*(z,w) \ll_A \left(\frac{\sqrt{z}}{w}\right)^A$$

for $A \in \mathbb{N}_0$. For $\sqrt{z}/w \leqslant C^{-\varepsilon}$ we obtain an admissible decomposition satisfying (6.14) and (6.15) by putting $W_+(z,w) = W_-(z,w) = 0$. Let us now assume $\sqrt{z}/w \geqslant \frac{1}{2}C^{-\varepsilon}$. We insert the decomposition from Lemma 7 into the definition of $W^*(z,w)$. In this way we obtain a decomposition satisfying (6.14) by putting

$$W_\pm(z,w) := \int_0^\infty W(y) F_\nu^\pm\big(4\pi\sqrt{yw+z}\big)\exp\big(\pm 4\pi i(\sqrt{yw+z}-\sqrt{z})\big)dy.$$

Now the second line of (6.15) is easily verified. As in Lemma 7, we glue these decompositions together to complete the proof of the lemma. □

Corollary 10. *The double Mellin transform*

$$\widehat{W}_\pm(s,t) = \int_0^\infty\int_0^\infty W_\pm(z,w) z^s w^t \frac{dz\,dw}{zw}$$

is absolutely convergent in the tube domain defined by $\Re t > 0$, $0 < \Re s + \Re t/2 < 1/4$, and satisfies

$$\widehat{W}_\pm(s,t) \ll_{A,B,\varepsilon,\Re s,\Re t} C^\varepsilon |s|^{-A}|t|^{-B} \tag{6.16}$$

in this region. Moreover, the Mellin inversion formula

$$W_\pm(z,w) = \int_{(c_2)}\int_{(c_1)} \widehat{W}_\pm(s,t) z^{-s} w^{-t} \frac{ds}{2\pi i}\frac{dt}{2\pi i}$$

holds whenever $c_1, c_2 > 0$, $c_1 + c_2/2 < 1/4$.

Proof. Repeated integration by parts gives

$$\widehat{W}_\pm(s,t) \ll_{i,j} |s|^{-i}|t|^{-j}\int_0^\infty\int_0^\infty z^i y^j W_\pm^{(i,j)}(z,w) z^{s-1} w^{t-1} dz\,dw.$$

Inserting (6.15) proves (6.16) in the desired range, and the Mellin inversion formula follows easily (for instance by applying first the one-dimensional inversion formula in w and then in z). □

Remark: Lemma 9 and Corollary 10 play an important role in the analysis of shifted convolution sums for holomorphic cusp forms. In the Maaß case we need a small, but somewhat technical extension of these results. It is convenient to state it already at this point:

(1) Lemma 9 holds true for negative w as long as $4|w| \leqslant z$ (with $|w|$ in place of w in (6.15)). In this case the support condition of W implies $yw + z > 0$ and in fact $yw + z \asymp z$.

(2) In order to encode the condition $4|w| \leqslant z$ into Corollary 10, we proceed as follows: let $0 < z_0 < 1$ and let $W_0(z,w)$ be a smooth function on $[0,\infty)\times\mathbb{R}$ such that
- $W_0(z,w) = 1$ if $5|w| \leqslant z$ and $z \geqslant z_0$,
- $W_0(z,w) = 0$ if $4|w| \geqslant z$ or $z \leqslant \frac{1}{2}z_0$,
- $z^i|w|^j W_0^{(i,j)}(z,w) \ll_{i,j} 1$ for all $i,j \in \mathbb{N}_0$, uniformly in z_0.

Define $\mathtt{W}_\pm(z,w) := W_0(z,w)W_\pm(z,w)$ with $W_\pm$ as in Lemma 9, and define

$$\widehat{\mathtt{W}}_{\pm,\pm}(s,t) = \int_0^\infty\int_0^\infty \mathtt{W}_\pm(z,\pm w) z^s w^t \frac{dz\,dw}{zw}.$$

Then Corollary 10 holds with $\widehat{\mathtt{W}}_{\pm,\pm}(s,t)$ in place of $\widehat{W}_\pm(s,t)$, and (6.16) is uniform in z_0.

7. Spectral decomposition of shifted convolution sums

This section is devoted to the spectral decomposition of the shifted convolution sum $\mathcal{D}(\ell_1,\ell_2,h,N,M)$, defined in (3.7). We choose a large parameter

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

and make the general assumption

$$h \asymp N \geqslant 20M. \tag{7.2}$$

We can also assume without loss of generality that

$$\ell_1, \ell_2 \leqslant 2N,$$

for otherwise $\mathcal{D}(\ell_1, \ell_2, h, N, M)$ vanishes trivially. Slightly more generally than in (3.5) we only assume that

$$V_{1,2} \text{ are supported in } [1,2] \text{ and satisfy } V_{1,2}^{(j)} \ll C^{j\varepsilon}. \tag{7.3}$$

The weight function V_2 localizes $\ell_1 n$ in a dyadic interval of size N, but the summation condition $\ell_1 n - \ell_2 m = h$ suggests that $\ell_1 n$ can, for a given h, vary only in an interval of length M. Therefore we attach a redundant weight function $W(\frac{\ell_1 n - h}{M})$ to the sum where W is smooth with bounded derivatives, constantly 1 on $[1,2]$, and supported on $[1/2, 3]$. With this notation, we can re-write

$$\begin{aligned}\mathcal{D}(\ell_1, \ell_2, h, N, M) &= \sum_{\ell_1 n - \ell_2 m = h} \lambda_1(m)\lambda_2(n) V_1\left(\frac{\ell_2 m}{M}\right) V_2\left(\frac{\ell_2 m + h}{N}\right) W\left(\frac{\ell_1 n - h}{M}\right) \\ &= \int_{-\infty}^{\infty} V_2^\dagger(z) e\left(\frac{zh}{N}\right) \mathcal{D}_z(\ell_1, \ell_2, h, N, M) dz,\end{aligned}$$

where $V_2^\dagger$ is the Fourier transform of V_2 and

$$\mathcal{D}_z(\ell_1, \ell_2, h, N, M) = \sum_{\ell_1 n - \ell_2 m = h} \lambda_1(m)\lambda_2(n) V_z\left(\frac{\ell_2 m}{M}\right) W\left(\frac{\ell_1 n - h}{M}\right),$$

with $V_z(x) = V_1(x)e(zxM/N)$. We can truncate the z-integral at $|z| \leqslant C^\varepsilon$ at the cost of an error $\mathrm{O}(C^{-100})$.

7.1. **The circle method.** The following lemma is Jutila's variant of the circle method [J1, J2].

Lemma 11. *[Jutila's circle method] Let $Q \geqslant 1$ and $Q^{-2} \leqslant \delta \leqslant Q^{-1}$ be two parameters. Let w be a nonnegative function with support in $[Q, 2Q]$ satisfying $\|w\|_\infty \leqslant 1$ and $\sum_c w(c) > 0$. For $r \in \mathbb{Q}$ write $I_r(\alpha)$ for the characteristic function of the interval $[r - \delta, r + \delta]$ and define*

$$\Lambda := \sum_c w(c)\phi(c), \quad \tilde{I}(\alpha) = \frac{1}{2\delta\Lambda} \sum_c w(c) \sideset{}{^*}\sum_{d \bmod c} I_{d/c}(\alpha). \tag{7.4}$$

Then $\tilde{I}(\alpha)$ is a good approximation to the characteristic function on $[0,1]$ in the sense that

$$\int_0^1 (1 - \tilde{I}(\alpha))^2 d\alpha \ll_\varepsilon \frac{Q^{2+\varepsilon}}{\delta\Lambda^2}$$

for any $\varepsilon > 0$.

We apply this lemma with $Q = C$ and $\delta = C^{-1}$. Let w_0 be a fixed smooth function with support in $[1,2]$, and let

$$w(c) = \begin{cases} w_0(c/C), & \ell_1\ell_2 \mid c, \\ 0, & \text{else.} \end{cases} \tag{7.5}$$

With the notation as in Lemma 11, we have

$$\Lambda \asymp C^2(\ell_1\ell_2)^{-1} \tag{7.6}$$

and

$$\begin{aligned}\mathcal{D}_z(\ell_1, \ell_2, h, N, M) &= \int_0^1 \sum_{n,m} \lambda_1(m)\lambda_2(n) W\left(\frac{\ell_1 n - h}{M}\right) V_z\left(\frac{\ell_2 m}{M}\right) e(\alpha(\ell_1 n - \ell_2 m - h)) d\alpha \\ &= \frac{1}{2\delta} \int_{-\delta}^{\delta} \mathcal{D}_{z,\eta}(\ell_1, \ell_2, h, N, M) d\eta + E,\end{aligned}$$

where

$$\mathcal{D}_{z,\eta}(\ell_1,\ell_2,h,N,M)$$
$$= \frac{1}{\Lambda}\sum_{\ell_1\ell_2|c} w_0\Big(\frac{c}{C}\Big)\sideset{}{^*}\sum_{d \bmod c}\sum_{n,m}\lambda_1(m)\lambda_2(n)e\Big(\frac{d}{c}(\ell_1 n-\ell_2 m-h)\Big)W_{\eta M}\Big(\frac{\ell_1 n-h}{M}\Big)V_{z,\eta M}\Big(\frac{\ell_2 m}{M}\Big) \tag{7.7}$$

with $V_{z,\eta}(x)=V_z(x)e(-\eta x)=V_1(x)e(x(zM/N-\eta))$, $W_\eta(x)=W(x)e(\eta x)$, and

$$E=\int_0^1\sum_{n,m}\lambda_1(m)\lambda_2(n)W\Big(\frac{\ell_1 n-h}{M}\Big)V_z\Big(\frac{\ell_2 m}{M}\Big)e(\alpha(\ell_1 n-\ell_2 m-h))(1-\tilde{I}(\alpha))d\alpha$$
$$\ll \frac{C^{1+\varepsilon}}{\delta^{1/2}\Lambda}\Big(\sum_{m\ll M/\ell_2}|\lambda_1(m)|\Big)\Big(\sum_{n\ll N/\ell_1}|\lambda_2(n)|\Big)\ll \frac{C^{1+\varepsilon}}{\delta^{1/2}\Lambda}\frac{NM}{\ell_1\ell_2}\ll \frac{NM}{C^{1/2-\varepsilon}}\ll C^{-2/5}$$

by the Cauchy-Schwarz inequality and (2.6). Since $|\eta|\leqslant C^{-1}=N^{-1000}$ is very small (in particular $\eta\ll M^{-1}$), the functions $V_{z,\eta M}$ and $W_{\eta M}$ have again nice properties, in particular $W_{\eta M}^{(j)}\ll 1$ and $V_{z,\eta M}^{(j)}\ll C^{j\varepsilon}$, uniformly in $|z|\ll C^\varepsilon$, and $V_{z,\eta M}$, $W_{\eta M}$ have support in $[1,2]$ resp. $[1/2,3]$.

7.2. **Voronoi summation.** In the main term (7.7), we apply Lemma 1 to the n,m-sum, getting

$$\sum_m \lambda_1(m)e\Big(-\frac{dm}{c/\ell_2}\Big)V_{z,\eta M}\Big(\frac{\ell_2 m}{M}\Big)=\frac{M}{c}\sum_m\lambda_1(m)e\Big(\frac{\bar{d}\ell_2 m}{c}\Big)\mathring{V}_{z,\eta M}\Big(\frac{\ell_2 mM}{c^2}\Big) \tag{7.8}$$

and

$$\sum_n \lambda_2(n)e\Big(\frac{dn}{c/\ell_1}\Big)W_{\eta M}\Big(\frac{\ell_1 n-h}{M}\Big)$$
$$= \frac{\ell_1}{c}\sum_n\lambda_2(n)e\Big(-\frac{\bar{d}\ell_1 n}{c}\Big)2\pi i^{\kappa_2}\int_0^\infty W_{\eta M}\Big(\frac{\ell_1 x-h}{M}\Big)J_{\kappa_2-1}\Big(4\pi\frac{\sqrt{xn}}{c/\ell_1}\Big)dx \tag{7.9}$$
$$= \frac{M}{c}\sum_n\lambda_2(n)e\Big(-\frac{\bar{d}\ell_1 n}{c}\Big)W^*_{\eta M}\Big(\frac{h\ell_1 n}{c^2},\frac{M\ell_1 n}{c^2}\Big),$$

where

$$W^*_{\eta M}(z,w)=2\pi i^{\kappa_2}\int_0^\infty W_{\eta M}(y)J_{\kappa_2-1}(4\pi\sqrt{yw+z})dy \tag{7.10}$$

was analyzed in Lemma 9. Substituting (7.8) and (7.9) back into (7.7) and using (6.14), we obtain

$$\mathcal{D}_{z,\eta}(\ell_1,\ell_2,h,N,M)=\frac{M^2}{\Lambda C}\sum_{\ell_1\ell_2|c}w_1\Big(\frac{c}{C}\Big)\frac{1}{c}\sum_{n,m}\lambda_1(m)\lambda_2(n)S(\ell_1 n-\ell_2 m,h,c)$$
$$\times W_\pm\Big(\frac{h\ell_1 n}{c^2},\frac{M\ell_1 n}{c^2}\Big)e\Big(\pm 2\frac{\sqrt{h\ell_1 n}}{c}\Big)\mathring{V}_{z,\eta M}\Big(\frac{\ell_2 m}{c^2/M}\Big)+\mathrm{O}(C^{-A})$$

where

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

By (6.15) and the fact that $\mathring{V}_{z,\eta M}$ is a Schwartz class function (cf. (2.11)) we can restrict the n,m-sums to

$$\ell_1 n\leqslant \mathcal{N}_0:=\frac{C^{2+\varepsilon}N}{M^2},\quad \ell_2 m\leqslant\mathcal{M}_0:=\frac{C^{2+\varepsilon}}{M} \tag{7.11}$$

at the cost of a negligible error. It is convenient to restrict the n and m-variable to dyadic intervals. We use the notation $x\asymp X$ to mean $X\leqslant x\leqslant 2X$, and for $\mathcal{N}\leqslant\mathcal{N}_0$, $\mathcal{M}\leqslant\mathcal{M}_0$ we split $\mathcal{D}_\eta(\ell_1,\ell_2,h,N,M)$ into subsums $n\asymp\mathcal{N}$, $m\asymp\mathcal{M}$. It is also convenient to restrict to $|\ell_1 n-\ell_2 m|\asymp\mathcal{K}$. We split the arising subsums into three pieces $\sum_+$, $\sum_0$, and $\sum_-$, according to $\ell_1 n>\ell_2 m$, $\ell_1 n=\ell_2 m$, and $\ell_1 n<\ell_2 m$. Each

of $\sum_+$, $\sum_0$ and $\sum_-$ depends on $\ell_1, \ell_2, h, N, M, \mathcal{N}, \mathcal{M}$ and $\mathcal{K}$. We first treat the terms with $\ell_1 n = \ell_2 m$. A trivial estimate shows that their contribution is at most

$$\sum\nolimits_0 \ll \frac{M^2}{\Lambda C^{1-\varepsilon}} \sum_{C\leqslant c\leqslant 2C} \frac{(h,c)}{c} \sum_{\substack{\ell_1 n \asymp \mathcal{N}, \ell_2 m \asymp \mathcal{M} \\ \ell_1 n = \ell_2 m}} |\lambda_1(m)\lambda_2(n)|$$

$$\ll \frac{M^2\tau(h)}{\Lambda C^{1-\varepsilon}} \Big(\sum_{m\ll \mathcal{M}} |\lambda_1(m)|^2 \Big)^{1/2} \Big(\sum_{n\ll \mathcal{N}} |\lambda_2(n)|^2 \Big)^{1/2} \ll \frac{M^2\tau(h)(\mathcal{N}_0\mathcal{M}_0)^{1/2}}{\Lambda C^{1-\varepsilon}} \ll \frac{C^\varepsilon (NM)^{1/2}\ell_1\ell_2}{C^{1-\varepsilon}} \ll C^{-1/2}.$$

7.3. **Spectral analysis of $\sum_+$.** Next, we consider

$$\sum\nolimits_+ = \frac{M^2}{\Lambda C} \sum_{\substack{b>0 \\ |b|\asymp \mathcal{K}}} \sum_{\substack{\ell_1 n - \ell_2 m = b \\ \ell_1 n \asymp \mathcal{N}, \ell_2 m \asymp \mathcal{M}}} \lambda_1(m)\lambda_2(n) \sum_{\ell_1\ell_2 | c} \frac{S(b,h,c)}{c} \Phi\left(4\pi \frac{\sqrt{|b|h}}{c} \right), \tag{7.12}$$

where

$$\Phi(x) = w_1\left(\frac{4\pi\sqrt{|b|h}}{xC}\right) W_\pm\left(\frac{\ell_1 n x^2}{(4\pi)^2|b|}, \frac{M\ell_1 n x^2}{(4\pi)^2|b|h}\right) e\left(\pm\frac{x\sqrt{\ell_1 n}}{2\pi\sqrt{|b|}}\right) \mathring{V}_{z,\eta M}\left(\frac{\ell_2 m M x^2}{(4\pi)^2|b|h}\right)$$

and the inner sum over c in (7.12) is ready for an application of the Kuznetsov trace formula (Lemma 3). (We are writing here $|b|$ instead of b for notational consistency with next subsection.) The relevant Bessel transforms of Φ are given by

$$\tilde\Phi(t) = 2\int_0^\infty \Omega\Big(\frac{xC}{4\pi\sqrt{|b|h}}\Big) \exp\Big(\pm ix\sqrt{\ell_1 n/|b|}\Big) \mathcal{J}^+_{2it}(x) \frac{dx}{x},$$

$$\dot\Phi(k) = 4i^k \int_0^\infty \Omega\Big(\frac{xC}{4\pi\sqrt{|b|h}}\Big) \exp\Big(\pm ix\sqrt{\ell_1 n/|b|}\Big) J_{k-1}(x) \frac{dx}{x},$$

(cf. (5.11) and (6.1)), where

$$\Omega(x) := w_1\left(\frac{1}{x}\right) \mathring{V}_{z,\eta M}\left(\frac{x^2 M\ell_2 m}{C^2}\right) W_\pm\left(\frac{x^2 h\ell_1 n}{C^2}, \frac{x^2 M\ell_1 n}{C^2}\right).$$

Note that Ω has support on a fixed compact interval (inherited from w_1) and is almost non-oscillating, more precisely

$$\Omega^{(j)}(x) \ll_j C^{j\varepsilon} \min\left(\Big(\frac{\mathcal{N}N}{C^2}\Big)^{-1/4}, \Big(\frac{\mathcal{N}N}{C^2}\Big)^{-\varepsilon} \right)$$

by (6.15). By Lemma 8 with

$$X = \frac{4\pi\sqrt{|b|h}}{C}, \quad Z = C^\varepsilon, \quad \alpha = \left(\frac{\ell_1 n}{|b|}\right)^{1/2} \geqslant 1,$$

the transforms $\tilde\Phi(t)$ and $\dot\Phi(k)$ are negligible unless

$$\begin{aligned} |t| &\ll \mathcal{T}_+ := C^\varepsilon \left(1 + \Big(\frac{\mathcal{K}N}{C^2}\Big)^{1/4} + \Big(\frac{\mathcal{M}N}{C^2}\Big)^{1/2}\right), \\ k &\ll \mathcal{T}_h := C^\varepsilon \left(1 + \Big(\frac{\mathcal{K}N}{C^2}\Big)^{1/4}\right). \end{aligned} \tag{7.13}$$

By the Kuznetsov formula (Lemma 3), $\sum_+ = \mathcal{H}_+(h) + \mathcal{M}_+(h) + \mathcal{E}_+(h) + \mathrm{O}(C^{-A})$ can be decomposed as the sum of three main terms, corresponding to the holomorphic, Maaß and Eisenstein spectrum, where

(7.14)
$$\mathcal{H}_+(h) = \frac{M^2}{\Lambda C}\int_0^\infty \sum_{\substack{2\leqslant k\leqslant \mathcal{T}_h\\ k \text{ even}}} \sum_{f\in\mathcal{B}_k(\ell_1\ell_2)} 4i^k\Gamma(k)J_{k-1}(x)\sqrt{h}\rho_f(h)\sum_{\substack{b>0\\ |b|\asymp\mathcal{K}}} w_1\Big(\frac{4\pi\sqrt{|b|h}}{Cx}\Big)\sqrt{|b|}\rho_f(b)\gamma_+(b,h,x)\frac{dx}{x}$$

with

$$\gamma_+(b,h,x) = \sum_{\substack{\ell_1 n-\ell_2 m=b\\ \ell_1 n\asymp\mathcal{N},\ell_2 m\asymp\mathcal{M}}} \lambda_1(m)\lambda_2(n)\mathring{V}_{z,\eta M}\left(\frac{x^2\ell_2 mM}{(4\pi)^2|b|h}\right) W_\pm\left(\frac{x^2\ell_1 n}{(4\pi)^2|b|}, \frac{x^2\ell_1 nM}{(4\pi)^2|b|h}\right)\vartheta_x\left(\frac{\ell_2 m}{|b|}\right)$$

and

$$\vartheta_x(y) = \exp\left(\pm ix\sqrt{1+y}\right) v\left(\frac{y}{\mathcal{M}/\mathcal{K}}\right),$$

where v is an artificially added, redundant smooth weight function of compact support $[1/4, 3]$ that is constantly 1 on $[1/2, 2]$. We note that

$$y^j\frac{d^j}{dy^j}\vartheta_x(y) \ll \Big(1+\frac{x\mathcal{M}}{\sqrt{\mathcal{K}\mathcal{N}}}\Big)^j. \tag{7.15}$$

Analogous expressions hold for $\mathcal{M}_+(h)$ and $\mathcal{E}_+(h)$:

$$\begin{aligned}
\mathcal{M}_+(h) &= \frac{2M^2}{\Lambda C}\int_0^\infty \sum_{\substack{f\in\mathcal{B}(\ell_1\ell_2)\\ |t_f|\leqslant\mathcal{T}_+}} \frac{\mathcal{J}^+_{2it_f}(x)}{\cosh(\pi t_f)}\sqrt{h}\rho_f(h)\sum_{\substack{b>0\\ |b|\asymp\mathcal{K}}} w_1\Big(\frac{4\pi\sqrt{|b|h}}{Cx}\Big)\sqrt{|b|}\rho_f(b)\gamma_+(b,h,x)\frac{dx}{x},\\
\mathcal{E}_+(h) &= \frac{2M^2}{\Lambda C}\int_0^\infty \frac{1}{4\pi}\sum_{\mathfrak{a}}\int_{-\mathcal{T}_+}^{\mathcal{T}_+} \frac{\mathcal{J}^+_{2it}(x)}{\cosh(\pi t)}\sqrt{h}\rho_{\mathfrak{a}}(h,t)\sum_{\substack{b>0\\ |b|\asymp\mathcal{K}}} w_1\Big(\frac{4\pi\sqrt{|b|h}}{Cx}\Big)\sqrt{|b|}\rho_{\mathfrak{a}}(b,t)dt\,\gamma_+(b,h,x)\frac{dx}{x}.
\end{aligned} \tag{7.16}$$

7.4. **Spectral analysis of $\sum_-$.** The treatment of

$$\sum\nolimits_- = \frac{M^2}{\Lambda C}\sum_{\substack{b<0\\ |b|\asymp\mathcal{K}}}\sum_{\substack{\ell_1 n-\ell_2 m=b\\ \ell_1 n\asymp\mathcal{N},\ell_2 m\asymp\mathcal{M}}} \lambda_1(m)\lambda_2(n)\sum_{\ell_1\ell_2|c}\frac{S(b,h,c)}{c}\Phi\left(4\pi\frac{\sqrt{|b|h}}{c}\right)$$

is similar, but the details are slightly different. Note that $b<0$ implies

$$\mathcal{N}+\mathcal{K} \ll \mathcal{M} \leqslant \mathcal{M}_0. \tag{7.17}$$

By Lemma 8, the integral transform $\check{\Phi}(t)$ is negligible unless

$$|t| \ll \mathcal{T}_- := C^\varepsilon\left(1+\Big(\frac{\mathcal{M}N}{C^2}\Big)^{1/2}\right). \tag{7.18}$$

Applying the opposite sign Kuznetsov formula, we obtain $\sum_- = \mathcal{M}_-(h) + \mathcal{E}_-(h) + \mathrm{O}(C^{-A})$ where (after a change of variables $x\mapsto 4\pi\sqrt{|b|}x$)

$$\mathcal{M}_-(h) = \frac{2M^2}{\Lambda C}\int_0^\infty \sum_{\substack{f\in\mathcal{B}(\ell_1\ell_2)\\ |t_f|\leqslant\mathcal{T}_-}} \frac{\sqrt{h}\rho_f(h)}{\cosh(\pi t_f)}\sum_{\substack{b<0\\ |b|\asymp\mathcal{K}}}\mathcal{J}^-_{2it_f}\left(4\pi\sqrt{|b|}x\right) w_1\left(\frac{\sqrt{h}}{Cx}\right)\sqrt{|b|}\rho_f(b)\gamma_-(b,h,x)\frac{dx}{x}, \tag{7.19}$$

with

$$\gamma_-(b,h,x)=\sum_{\substack{\ell_1 n-\ell_2 m=b\\ \ell_1 n\asymp\mathcal{N},\ell_2 m\asymp\mathcal{M}}}\lambda_1(m)\lambda_2(n)\mathring{V}_{z,\eta M}\left(\frac{x^2\ell_2 mM}{h}\right)W_\pm\left(x^2\ell_1 n,\frac{x^2\ell_1 nM}{h}\right)e\left(\pm 2x\sqrt{\ell_1 n}\right).$$

By Mellin inversion and (6.5), we have up to a negligible error

$$\begin{aligned}\mathcal{M}_-(h)=\frac{2M^2}{\Lambda C}\int_0^\infty\int_{1-iC^\varepsilon\mathcal{T}_-}^{1+iC^\varepsilon\mathcal{T}_-}&\sum_{\substack{f\in\mathcal{B}(\ell_1\ell_2)\\ |t_f|\leqslant\mathcal{T}_-}}\widehat{\mathcal{J}}^-_{2it_f}(s)\frac{\sqrt{h}\rho_f(h)}{\cosh(\pi t_f)}w_1\left(\frac{\sqrt{h}}{Cx}\right)\\ &\times\sum_{\substack{b<0\\ |b|\asymp\mathcal{K}}}\left(4\pi\sqrt{|b|}x\right)^{-s}\sqrt{|b|}\rho_f(b)\gamma_-(b,h,x)\,\frac{ds}{2\pi i}\,\frac{dx}{x}.\end{aligned}\tag{7.20}$$

An analogous formula holds for $\mathcal{E}_-(h)$.

7.5. **Conclusion.** Before we sum over h in the next section, we pause for a moment and summarize our discussion by stating the following decomposition.

Proposition 9. *Let $\ell_1,\ell_2,h\in\mathbb{N}$, $M,N\geqslant 1$. Let $C=N^{1000}$, $\delta=1/C$, assume (7.2), and define $\mathcal{N}_0$, $\mathcal{M}_0$ by (7.11). Let w_0 be a fixed smooth function with support in $[1,2]$, define Λ as in (7.4) using (7.5), and let $w_1(x)=w_0(x)/x$. Define $\mathcal{T}_h$, $\mathcal{T}_+$ and $\mathcal{T}_-$ as in (7.13) and (7.18). Assume that $V_{1,2}$ satisfy (7.3), let W be as in the discussion after (7.3) and $V_2^\dagger$ be the Fourier transform of V_2, and define $\mathring{V}$ as in (2.10) and, for $z,\eta\in\mathbb{R}$, $V_{z,\eta M}(x)=V_1(x)e(-\eta Mx)e(zxM/N)$. Let $W^*(z,w)$ be defined by (7.10) and correspondingly $W_\pm$ by (6.14). Finally recall the special functions (6.1). With this notation define $\mathcal{H}_+(h)$, $\mathcal{M}_\pm(h)$, $\mathcal{E}_\pm(h)$ as in (7.14), (7.16), (7.19).*

Then the smooth shifted convolution sum $\mathcal{D}(\ell_1,\ell_2,h,N,M)$ defined in (3.7) equals

$$\begin{aligned}\mathcal{D}(\ell_1,\ell_2,h,N,M)=&\frac{1}{2\delta}\int_{-\delta}^{\delta}\int_{-C^\varepsilon}^{C^\varepsilon}V_2^\dagger(z)e\left(\frac{zh}{N}\right)\sum_{\mathcal{N}\leqslant\mathcal{N}_0}\sum_{\substack{\mathcal{M}\leqslant\mathcal{M}_0\\ \mathcal{M},\mathcal{K}\leqslant\mathcal{N}}}\sum_{\mathcal{K}\leqslant\mathcal{N}_0}\big(\mathcal{H}_+(h)+\mathcal{M}_+(h)+\mathcal{E}_+(h)\big)dz\,d\eta\\ &+\frac{1}{2\delta}\int_{-\delta}^{\delta}\int_{-C^\varepsilon}^{C^\varepsilon}V_2^\dagger(z)e\left(\frac{zh}{N}\right)\sum_{\mathcal{N}\leqslant\mathcal{M}_0}\sum_{\mathcal{M}\leqslant\mathcal{M}_0}\sum_{\mathcal{K}\leqslant\mathcal{M}_0}\big(\mathcal{M}_-(h)+\mathcal{E}_-(h)\big)\,dz\,d\eta+\mathrm{O}(C^{-1/3})\end{aligned}\tag{7.21}$$

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

8. Shifted convolution sums on average

In this section, we use Proposition 9 to study averages of shifted convolution sums $\mathcal{S}(\ell_1,\ell_2,d,N,M)=\sum_r\mathcal{D}(\ell_1,\ell_2,rd,N,M)$ over multiples of a positive integer d, which were defined in (3.8). In particular, we will prove Proposition 7. Write

$$\beta:=\mathrm{lcm}(\ell_1,\ell_2,d).$$

Our general assumption (7.2) is still in place, so that $\mathcal{D}(\ell_1,\ell_2,rd,N,M)$ vanishes unless $r\asymp N/d$. We keep the notation from the previous section and import in particular the inequalities (7.1), (7.6), (7.11), (7.13), (7.18). We start by considering

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

where $\mathcal{H}_+$ was defined in (7.14). We will sacrifice cancellation in the x-integral (in some typical ranges there is very little cancellation anyway) (7.14) and just note that the range of integration is

$$x \asymp X_+ := \frac{\sqrt{\mathcal{K}N}}{C}. \tag{8.1}$$

8.1. **Separation of variables.** We need to separate the variables $h = r_1r_2d, b, n, m$, scattered in the various smooth weight functions. We do this by brute force, expressing each weight function as an inverse Mellin transform. Since all of them are essentially non-oscillating (at least in typical ranges), this can be done with little loss. With this in mind we write

$$\begin{aligned}
&w_1\left(\frac{4\pi\sqrt{|b|r_1r_2d}}{Cx}\right)\mathring{V}_{z,\eta M}\left(\frac{x^2\ell_2mM}{(4\pi)^2|b|r_1r_2d}\right)W_\pm\left(\frac{x^2\ell_1n}{(4\pi)^2|b|},\frac{x^2\ell_1nM}{(4\pi)^2|b|r_1r_2d}\right)\vartheta_x\left(\frac{\ell_2m}{|b|}\right)\\
&=\frac{1}{(2\pi i)^5}\int_{(0)}\int_{(\varepsilon)}\int_{(1/4-\varepsilon)}\int_{(\varepsilon)}\int_{(0)}\widehat{w}_1(s_1)\widehat{\mathring{V}}_{z,\eta M}(s_2)\widehat{W}_\pm(s_3,s_4)\widehat{\vartheta_x}(s_5)\\
&\times\left(\frac{4\pi\sqrt{|b|r_1r_2d}}{Cx}\right)^{-s_1}\left(\frac{x^2\ell_2mM}{(4\pi)^2|b|r_1r_2d}\right)^{-s_2}\left(\frac{x^2\ell_1n}{(4\pi)^2|b|}\right)^{-s_3}\left(\frac{x^2\ell_1nM}{(4\pi)^2|b|r_1r_2d}\right)^{-s_4}\left(\frac{\ell_2m}{|b|}\right)^{-s_5}ds_5\,ds_4\,ds_3\,ds_2\,ds_1.
\end{aligned}$$

The multiple integral is absolutely convergent, and we recall in particular Corollary 10. The $s_1,\dots,s_4$-integrals are rapidly converging and can be truncated at $|\Im s_j|\leqslant C^\varepsilon$ at the cost of a negligible error. By (7.15) the s_5-integral can be truncated at

$$|\Im s_5|\leqslant S := C^\varepsilon\Big(1+\frac{X_+\mathcal{M}}{\sqrt{\mathcal{K}\mathcal{N}}}\Big). \tag{8.2}$$

It is convenient to re-write the last line of the penultimate display as

$$\begin{aligned}
&\left(\frac{x}{X_+}\right)^{s_1-2(s_2+s_3+s_4)}\left(\frac{|b|}{\mathcal{K}}\right)^{-\frac{s_1}{2}+s_2+s_3+s_4+s_5}\left(\frac{\ell_1n}{\mathcal{N}}\right)^{-s_3-s_4}\left(\frac{\ell_2m}{\mathcal{M}}\right)^{-s_2-s_5}(r_1r_2d)^{-\frac{s_1}{2}+s_2+s_4}\\
&\times\frac{C^{s_1}\mathcal{K}^{-\frac{s_1}{2}+s_2+s_3+s_4+s_5}}{M^{s_2+s_4}(X_+/(4\pi))^{-s_1+2(s_2+s_3+s_4)}\mathcal{N}^{s_3+s_4}\mathcal{M}^{s_2+s_5}}\ll C^\varepsilon\frac{\mathcal{K}^{1/4}}{X_+^{1/2}\mathcal{N}^{1/4}}.
\end{aligned}$$

We substitute this back into (7.14), estimate the x- and s_j-integrals trivially and finally apply the Cauchy-Schwarz inequality to get

$$\sum_{\substack{r_1\asymp N/(dr_2)\\(r_1,\beta)=1}}e\left(\frac{zr_1r_2d}{N}\right)\mathcal{H}_+(r_1r_2d)\ll\frac{C^\varepsilon M^2}{\Lambda C}\frac{\mathcal{K}^{1/4}}{X_+^{1/2}\mathcal{N}^{1/4}}S\,\big(\Xi^{\mathcal{H}}_{1,+}\,\Xi^{\mathcal{H}}_{2,+}\big)^{1/2} \tag{8.3}$$

where

$$\Xi^{\mathcal{H}}_{1,+}=\max_{|u_4|\leqslant C^\varepsilon}\sum_{\substack{2\leqslant k\leqslant\mathcal{T}_h\\k\text{ even}}}\Gamma(k)\sum_{f\in\mathcal{B}_k(\ell_1\ell_2)}\Big|\sum_{\substack{r_1\asymp N/(dr_2)\\(r_1,\beta)=1}}e\left(\frac{zr_1r_2d}{N}\right)r_1^{2\varepsilon+iu_4}\sqrt{r_1r_2d}\rho_f(r_1r_2d)\Big|^2,$$

$$\Xi^{\mathcal{H}}_{2,+}=\max_{\substack{|u_2|\leqslant C^\varepsilon\\|u_1|,|u_3|\leqslant S\\x\asymp X_+}}\sum_{\substack{2\leqslant k\leqslant\mathcal{T}_h\\k\text{ even}}}|J_{k-1}(x)|^2\,\Gamma(k)\sum_{f\in\mathcal{B}_k(\ell_1\ell_2)}\Big|\sum_{|b|\asymp\mathcal{K}}\sqrt{|b|}\rho_f(b)\gamma^*(b)\Big|^2,$$

with

$$\gamma^*(b)=\left(\frac{|b|}{\mathcal{K}}\right)^{1/4+\varepsilon+iu_3}\sum_{\substack{\ell_1n-\ell_2m=b\\\ell_1n\asymp\mathcal{N},\ell_2m\asymp\mathcal{M}}}\left(\frac{\ell_1n}{\mathcal{N}}\right)^{-1/4+iu_2}\left(\frac{\ell_2m}{\mathcal{M}}\right)^{-\varepsilon+iu_1}\lambda_1(m)\lambda_2(n).$$

The same analysis works *mutatis mutandis* for the Eisenstein and Maaß spectrum, giving similar expressions $\Xi^{\mathcal{E}}_{1/2,+}$ and $\Xi^{\mathcal{M}}_{1/2,+}$.

8.2. **The spectral large sieve.** We proceed to estimate the various $\Xi^{\star}_{j,+}$ for $j \in \{1,2\}$, $\star \in \{\mathcal{H}, \mathcal{E}, \mathcal{M}\}$. We have

$$\sum_b |\gamma^*(b)|^2 \ll \int_0^1 \Big| \sum_{\ell_2 m \asymp \mathcal{M}} \lambda_1(m) \Big(\frac{\ell_2 m}{\mathcal{M}}\Big)^{-\varepsilon+iu_1} e(\ell_2 m\alpha)\Big|^2 \Big| \sum_{\ell_1 n \asymp \mathcal{N}} \lambda_2(n) \Big(\frac{\ell_1 n}{\mathcal{N}}\Big)^{-1/4+iu_2} e(-\ell_1 n\alpha)\Big|^2 d\alpha.$$

Since u_2 is small, we can successfully apply Wilton's bound (2.9) and partial summation. This does not work efficiently for the m-sum, but having estimated the n-sum by its sup-norm, we can open the square and use (2.6) to conclude that

$$\sum_b |\gamma^*(b)|^2 \ll C^{\varepsilon} \frac{\mathcal{N}}{\ell_1} \sum_{m \asymp \mathcal{M}/\ell_2} |\lambda_1(m)|^2 \ll C^{\varepsilon} \frac{\mathcal{N}\mathcal{M}}{\ell_1\ell_2}, \tag{8.4}$$

uniformly in u_1, u_2, u_3.

In order to bound the Bessel function $J_{k-1}(x)$ in $\Xi^{\mathcal{H}}_{2,+}$, we recall the size of x in (8.1) and k in (7.13). If $X_+ \geqslant 10^3 C^{\varepsilon}(1+X_+^{1/2})$, then $J_{k-1}(x) \ll x^{-1/2}$ by (6.6). The opposite assumption $X_+ < 10^3 C^{\varepsilon}(1+X_+^{1/2})$ implies $X_+ \ll C^{2\varepsilon}$ and hence trivially $J_{k-1}(x) \ll 1 \ll C^{\varepsilon} x^{-1/2}$. By the large sieve inequality (Lemma 4) and (8.4) we obtain

$$\Xi^{\mathcal{H}}_{2,+} \ll \frac{C^{\varepsilon}}{X_+} \Big(\mathcal{T}_h^2 + \frac{\mathcal{K}}{\ell_1\ell_2}\Big) \frac{\mathcal{N}\mathcal{M}}{\ell_1\ell_2}. \tag{8.5}$$

Similarly one shows

$$|\Xi^{\mathcal{E}}_{2,+}| + |\Xi^{\mathcal{M}}_{2,+}| \ll \frac{C^{\varepsilon}}{X_+} \Big(\mathcal{T}_+^2 + \frac{\mathcal{K}}{\ell_1\ell_2}\Big) \frac{\mathcal{N}\mathcal{M}}{\ell_1\ell_2}. \tag{8.6}$$

By (2.7) and (5.4) we obtain

$$\Xi^{\mathcal{H}}_{1,+} \ll \max_{|u_4| \leqslant C^{\varepsilon}} C^{\varepsilon} \sum_{\delta | \ell_1\ell_2} \sum_{\substack{2 \leqslant k \leqslant \mathcal{T}_h \\ k \text{ even}}} \Gamma(k) \sum_{f \in \mathcal{B}_k(\ell_1\ell_2)} \Big| \sum_{\substack{r_1 \asymp N/(dr_2) \\ (r_1,\beta)=1}} \alpha(r_1) \sqrt{r_1\delta}\, \rho_f(r_1\delta)\Big|^2.$$

where

$$\alpha(r_1) = \alpha_{r_2 d, u_4}(r_1) = e\Big(\frac{z r_1 r_2 d}{N}\Big) r_1^{2\varepsilon + iu_4}. \tag{8.7}$$

The large sieve (Lemma 4) yields

$$\Xi^{\mathcal{H}}_{1,+} \ll C^{\varepsilon} \sum_{\delta | \ell_1\ell_2} \Big(\mathcal{T}_h^2 + \frac{N\delta}{dr_2\ell_1\ell_2}\Big) \frac{N}{dr_2} \ll C^{\varepsilon} \Big(\mathcal{T}_h^2 + \frac{N}{dr_2}\Big) \frac{N}{dr_2}. \tag{8.8}$$

(We could be more careful here with powers of $\ell_1\ell_2$, but this is not necessary.) Using (5.5) we obtain analogously

$$\Xi^{\mathcal{E}}_{1,+} \ll C^{\varepsilon} \Big(\mathcal{T}_+^2 + \frac{N}{dr_2}\Big) \frac{N}{dr_2}. \tag{8.9}$$

Note that the upper bounds in (8.6) and (8.9) majorize those in (8.5) and (8.8). Finally we apply Theorem 8 to obtain

$$\begin{aligned} \Xi^{\mathcal{M}}_{1,+} &= \max_{|u_4| \leqslant C^{\varepsilon}} \sum_{\substack{|t_f| \leqslant \mathcal{T}_+ \\ f \in \mathcal{B}(\ell_1\ell_2)}} \frac{1}{\cosh(\pi t_f)} \Big| \sum_{\substack{r_1 \asymp N/(dr_2) \\ (r_1,\beta)=1}} \alpha(r_1) \sqrt{r_1 r_2 d}\, \rho_f(r_1 r_2 d)\Big|^2 \\ &\ll C^{\varepsilon}(\ell_1\ell_2, r_2 d) \Big(\mathcal{T}_+ + \frac{(r_2 d)^{1/2}}{(\ell_1\ell_2)^{1/2}}\Big) \Big(\mathcal{T}_+ + \frac{N}{dr_2(\ell_1\ell_2)^{1/2}}\Big) \frac{N}{dr_2}. \end{aligned} \tag{8.10}$$

Combining (8.5) – (8.10), we conclude the final bound

$$\begin{aligned}&\big(|\Xi^{\mathcal{H}}_{1,+}|+|\Xi^{\mathcal{E}}_{1,+}|+|\Xi^{\mathcal{M}}_{1,+}|\big)\big(|\Xi^{\mathcal{H}}_{2,+}|+|\Xi^{\mathcal{E}}_{2,+}|+|\Xi^{\mathcal{M}}_{2,+}|\big)\\ &\ll \frac{C^\varepsilon}{X_+}\left(\Big(\mathcal{T}_+ + \frac{(r_2d)^{1/2}}{(\ell_1\ell_2)^{1/2}}\Big)\Big(\mathcal{T}_+ + \frac{N}{dr_2(\ell_1\ell_2)^{1/2}}\Big) + \frac{N}{dr_2}\right)\frac{N}{dr_2}\left(\mathcal{T}_+^2 + \frac{\mathcal{K}}{\ell_1\ell_2}\right)\frac{(\ell_1\ell_2, r_2d)\mathcal{N}\mathcal{M}}{\ell_1\ell_2}.\end{aligned} \tag{8.11}$$

8.3. **Conclusion of the plus-case.** It is now a matter of book-keeping. Combining (7.6), (7.13), (8.1), (8.2), (8.3) and (8.11), we obtain

$$\begin{aligned}&\sum_{\substack{r_1\asymp N/(dr_2)\\(r_1,q)=1}} e\left(\frac{zr_1r_2d}{N}\right)\big(|\mathcal{H}_+(r_1r_2d)|+|\mathcal{M}_+(r_1r_2d)|+|\mathcal{E}_+(r_1r_2d)|\big)\\ &\ll C^\varepsilon\frac{M^2\mathcal{N}^{1/4}\mathcal{M}^{1/2}(\ell_1\ell_2(\ell_1\ell_2,r_2d))^{1/2}}{C^2(dr_2)^{1/2}\mathcal{K}^{1/4}}\left(1+\frac{N^{1/2}\mathcal{M}}{C\mathcal{N}^{1/2}}\right)\left(1+\Big(\frac{\mathcal{K}N}{C^2}\Big)^{1/4}+\Big(\frac{\mathcal{M}N}{C^2}\Big)^{1/2}+\Big(\frac{\mathcal{K}}{\ell_1\ell_2}\Big)^{1/2}\right)\\ &\times\left(\left(1+\Big(\frac{\mathcal{K}N}{C^2}\Big)^{1/8}+\Big(\frac{\mathcal{M}N}{C^2}\Big)^{1/4}+\frac{(r_2d)^{1/4}}{(\ell_1\ell_2)^{1/4}}\right)\left(1+\Big(\frac{\mathcal{K}N}{C^2}\Big)^{1/8}+\Big(\frac{\mathcal{M}N}{C^2}\Big)^{1/4}+\frac{N^{1/2}}{(dr_2)^{1/2}(\ell_1\ell_2)^{1/4}}\right)+\frac{N^{1/2}}{(dr_2)^{1/2}}\right).\end{aligned}$$

We multiply out the 136 terms, and and write each term as

$$\mathcal{M}^\alpha\mathcal{K}^\beta\mathcal{N}^\gamma C^\delta\times\text{expression in } N, M, \ell_1, \ell_2, d, r_2.$$

At this point it is important to recall (7.1), (7.11) and the size conditions $\mathcal{M},\mathcal{K}\leqslant\mathcal{N}$ in the summation condition of the first line of (7.21). We conclude that all terms with

$$2\alpha+2\gamma+2\max(\beta,0)<-\delta-1/3$$

are less than $C^{-1/4}$ and therefore negligible. This applies to all terms except those involving the last term $(\mathcal{K}/\ell_1\ell_2)^{1/2}$ in the second parenthesis on the right hand side. Hence we obtain the bound

$$\begin{aligned}&C^\varepsilon\frac{M^2\mathcal{M}^{1/2}(\ell_1\ell_2,r_2d)^{1/2}}{C^2(dr_2)^{1/2}}\left((\mathcal{N}\mathcal{K})^{1/4}+\frac{N^{1/2}\mathcal{M}\mathcal{K}^{1/4}}{C\mathcal{N}^{1/4}}\right)\Bigg[\left(1+\Big(\frac{\mathcal{K}N}{C^2}\Big)^{1/8}+\Big(\frac{\mathcal{M}N}{C^2}\Big)^{1/4}+\frac{(r_2d)^{1/4}}{(\ell_1\ell_2)^{1/4}}\right)\\ &\times\left(1+\Big(\frac{\mathcal{K}N}{C^2}\Big)^{1/8}+\Big(\frac{\mathcal{M}N}{C^2}\Big)^{1/4}+\frac{N^{1/2}}{(dr_2)^{1/2}(\ell_1\ell_2)^{1/4}}\right)+\frac{N^{1/2}}{(dr_2)^{1/2}}\Bigg]+C^{-1/4}.\end{aligned}$$

In the first parenthesis we cancel $(\mathcal{K}/\mathcal{N})^{1/4}\leqslant 1$. Having done this, all terms are increasing in $\mathcal{K},\mathcal{M},\mathcal{N}$, and we insert (7.11). This gives the final bound

$$\begin{aligned}&\sum_{\substack{r_1\asymp N/(dr_2)\\(r_1,q)=1}} e\left(\frac{zr_1r_2d}{N}\right)\big(|\mathcal{H}_+(r_1r_2d)|+|\mathcal{M}_+(r_1r_2d)|+|\mathcal{E}_+(r_1r_2d)|\big)\\ &\ll C^\varepsilon\frac{M^{3/2}(\ell_1\ell_2,r_2d)^{1/2}}{C(dr_2)^{1/2}}\frac{CN^{1/2}}{M}\left(\left(\frac{N^{1/4}}{M^{1/4}}+\frac{(r_2d)^{1/4}}{(\ell_1\ell_2)^{1/4}}\right)\left(\frac{N^{1/4}}{M^{1/4}}+\frac{N^{1/2}}{(dr_2)^{1/2}(\ell_1\ell_2)^{1/4}}\right)+\frac{NM^{1/2}}{dr_2}\right)\\ &\ll C^\varepsilon(\ell_1\ell_2,r_2d)^{1/2}\left(\frac{N}{(dr_2)^{1/2}}+\frac{N^{5/4}M^{1/4}}{dr_2(\ell_1\ell_2)^{1/4}}+\frac{N^{3/4}M^{1/4}}{(dr_2\ell_1\ell_2)^{1/4}}+\frac{NM^{1/2}}{(dr_2)^{3/4}(\ell_1\ell_2)^{1/2}}+\frac{NM^{1/2}}{dr_2}\right)\\ &\ll C^\varepsilon(\ell_1\ell_2,d)^{1/2}\left(\frac{N}{d^{1/2}}+\frac{N^{5/4}M^{1/4}}{d(\ell_1\ell_2)^{1/4}}+\frac{N^{3/4}M^{1/4}}{d^{1/4}}+\frac{NM^{1/2}}{d^{3/4}(\ell_1\ell_2)^{1/2}}+\frac{NM^{1/2}}{d}\right).\end{aligned} \tag{8.12}$$

(Here, of course, the term $C^{-1/4}$ can be absorbed.)

8.4. **The minus-case.** The treatment of $\mathcal{M}_-$ and $\mathcal{E}_-$ is similar in spirit, but the details are slightly different and considerably less involved. In particular, we can afford to be somewhat lossy in our estimations. We recall from (7.19) that the range of integration is

$$x \asymp X_- := \frac{\sqrt{N}}{C} \tag{8.13}$$

which is quite different from the previous case. We separate variables in

$$\begin{aligned}&\mathring{V}_{\eta M}\left(\frac{x^2\ell_2 mM}{r_1r_2d}\right) W_\pm\left(x^2\ell_1 n, \frac{x^2\ell_1 nM}{r_1r_2d}\right)\\ &= \frac{1}{(2\pi i)^3}\int_{(\varepsilon)}\int_{(1/4-\varepsilon)}\int_{(\varepsilon)} \widehat{\mathring{V}}_{z,\eta M}(s_1)\widehat{W}_\pm(s_2,s_3)\left(\frac{x^2\ell_2 mM}{r_1r_2d}\right)^{-s_1}\left(x^2\ell_1 n\right)^{-s_2}\left(\frac{x^2\ell_1 nM}{r_1r_2d}\right)^{-s_3} ds_3\, ds_2\, ds_1\end{aligned}$$

by Mellin inversion. All integrals are rapidly converging and can be truncated at $|\Im s_j| \leqslant C^\varepsilon$ at the cost of a negligible error. We substitute this back into (7.20), estimate the x-, s- and s_j-integrals trivially (using (6.5)) and apply the Cauchy-Schwarz inequality getting

$$\sum_{\substack{r_1 \asymp N/(dr_2)\\(r_1,q)=1}} e\left(\frac{zr_1r_2d}{N}\right)\mathcal{M}_-(r_1r_2d) \ll C^\varepsilon \frac{M^2\ell_1\ell_2}{C^3}\frac{1}{X_-^{1/2}\mathcal{N}^{1/4}}\frac{\mathcal{T}_-}{\mathcal{K}^{1/2}X_-}\Xi_{1,-}^{1/2}\Xi_{2,-}^{1/2} \tag{8.14}$$

where

$$\Xi_{1,-} = \max_{\substack{|u_3|\leqslant C^\varepsilon\\ x\asymp X_-}} \sum_{\substack{f\in\mathcal{B}(\ell_1\ell_2)\\|t_f|\leqslant\mathcal{T}_-}} \frac{1}{\cosh(\pi t_f)}\Big|\sum_{\substack{r_1\asymp N/(dr_2)\\(r_1,\beta)=1}} \tilde{\alpha}(r_1) w_1\left(\frac{\sqrt{r_1r_2d}}{Cx}\right)\sqrt{r_1r_2d}\rho_f(r_1r_2d)\Big|^2,$$

$$\Xi_{2,-} = \max_{\substack{|u_1|,|u_2|\leqslant C^\varepsilon\\ x\asymp X_-}} \sum_{\substack{f\in\mathcal{B}(\ell_1\ell_2)\\|t_f|\leqslant\mathcal{T}_-}} \frac{1}{\cosh(\pi t_f)}\Big|\sum_{|b|\asymp\mathcal{K}} \sqrt{|b|}\rho_f(b)\gamma^*(b)\Big|^2$$

with $\mathcal{T}_-$ as in (7.18),

$$\tilde{\alpha}(r_1) = r_1^{2\varepsilon+iu_3} e\left(\frac{zr_1r_2d}{N}\right)$$

and

$$\gamma^*(b) = \sum_{\substack{\ell_1 n-\ell_2 m=b\\ \ell_1 n\asymp\mathcal{N},\ell_2 m\asymp\mathcal{M}}} \left(\frac{\ell_1 n}{\mathcal{N}}\right)^{-1/4+iu_1}\left(\frac{\ell_1 m}{\mathcal{M}}\right)^{-\varepsilon+iu_2}\lambda_1(m)\lambda_2(n)e(\pm 2x\sqrt{\ell_1 n}).$$

As in (8.4) we find

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

uniformly in x, u_1, u_2, and hence by the large sieve

$$\Xi_{2,-} \ll \left(\mathcal{T}_-^2 + \frac{\mathcal{K}}{\ell_1\ell_2}\right) C^\varepsilon \frac{\mathcal{N}\mathcal{M}}{\ell_1\ell_2}.$$

The estimation of $\Xi_{1,-}$ is similar to the preceding analysis, but simpler. Here we apply (5.12) (in a weak version without the denominator ℓ) to obtain

$$\Xi_{1,-} \ll C^\varepsilon\left(\mathcal{T}_-^2 + \frac{N}{dr_2}\right)\frac{N}{dr_2}(dr_2)^{2\theta}(\ell_1\ell_2, dr_2)^{1-2\theta}$$

For the treatment of Eisenstein case we can directly apply (5.5) and the large sieve as in (8.8) – (8.9) getting a slightly stronger bound. Substituting back into (8.14) and recalling (7.18) and (8.13), we obtain

$$\sum_{\substack{r_1 \asymp N/(dr_2)\\(r_1,\beta)=1}} e\left(\frac{zr_1r_2d}{N}\right)\left(|\mathcal{M}_-(r_1r_2d)| + |\mathcal{E}_-(r_1r_2d)|\right) \ll C^{\varepsilon}\frac{M^2N^{1/4}(\ell_1\ell_2)^{1/2}\mathcal{M}^{1/2}\mathcal{N}^{1/4}(\ell_1\ell_2, dr_2)^{1/2-\theta}}{C^{5/2}(dr_2)^{1/2-\theta}}$$
$$\times \frac{1+(\mathcal{M}N/C^2)^{1/2}}{(\mathcal{K}N/C^2)^{1/2}}\left(1+\left(\frac{\mathcal{M}N}{C^2}\right)^{1/2}+\left(\frac{\mathcal{K}}{\ell_1\ell_2}\right)^{1/2}\right)\left(1+\left(\frac{\mathcal{M}N}{C^2}\right)^{1/2}+\left(\frac{N}{dr_2}\right)^{1/2}\right).$$

As before, we use (7.1) to argue that in the penultimate parenthesis only the third term contributes non-negligibly. The resulting expression is increasing in $\mathcal{M}, \mathcal{N}, \mathcal{K}$ each of which are bounded by $\mathcal{M}_0$, see (7.17). Now a straightforward calculation similar to the above shows the bound

(8.15)
$$\sum_{\substack{r_1 \asymp N/(dr_2)\\(r_1,\beta)=1}} e\left(\frac{zr_1r_2d}{N}\right)\left(|\mathcal{M}_-(r_1r_2d)| + |\mathcal{E}_-(r_1r_2d)|\right)$$
$$\ll C^{\varepsilon}(dr_2)^{\theta}(\ell_1\ell_2, dr_2)^{1/2-\theta}\left(\frac{M^{1/4}N^{3/4}}{(dr_2)^{1/2}}+\frac{M^{3/4}N^{3/4}}{dr_2}\right) \ll C^{\varepsilon}d^{\theta}(\ell_1\ell_2,d)^{1/2}\left(\frac{M^{1/4}N^{3/4}}{d^{1/2}}+\frac{M^{3/4}N^{3/4}}{d}\right)$$

8.5. **Conclusion.** We sum (8.12) and (8.15) over $r_2 \mid \beta^{\infty}$; by Rankin's trick it is easy to see that

$$\sum_{\substack{r\leqslant X\\ r\mid\beta^{\infty}}} 1 \ll (X\beta)^{\varepsilon}.$$

Using $\theta \leqslant 1/4$ and $N \geqslant M$, we conclude

$$\mathcal{S}(\ell_1,\ell_2,d,N,M) \ll N^{\varepsilon}(\ell_1\ell_2,d)^{1/2}\left(\frac{N}{d^{1/2}}+\frac{N^{5/4}M^{1/4}}{d}+\frac{N^{3/4}M^{1/4}}{d^{1/4}}+\frac{NM^{1/2}}{d^{3/4}}\right). \tag{8.16}$$

We remove the factor $(\ell_1\ell_2,d)^{1/2}$ as follows. We decompose

$$\ell_1 = \ell_1'\tilde{\ell}\delta_1\delta, \quad \ell_2 = \ell_2'\tilde{\ell}\delta_2\delta, \quad d = d'\delta_1\delta_2\delta$$

where $\delta = (d,\ell_1,\ell_2)$, $\delta_1 = (d,\ell_1)/\delta$, $\delta_2 = (d,\ell_2)/\delta$, $\tilde{\ell} = (\ell_1,\ell_2)/\delta$. Using (2.5), we find

$$\begin{aligned}\mathcal{S}(\ell_1,\ell_2,d,N,M) &= \sum_r \sum_{\ell_1 n-\ell_2 m=rd} \lambda_1(m)\lambda_2(n)V\left(\frac{\ell_1 n}{N}\right)V\left(\frac{\ell_2 m}{M}\right)\\ &= \sum_r \sum_{\ell_1' n-\ell_2 m=d'r} \lambda_1(\delta_1 m)\lambda_2(\delta_2 n)V\left(\frac{\delta_2\ell_1 n}{N}\right)V\left(\frac{\delta_1\ell_2 m}{M}\right)\\ &= \sum_{g\mid\delta_2}\sum_{h\mid\delta_1}\mu(g)\mu(h)\lambda_2\left(\frac{\delta_2}{g}\right)\lambda_1\left(\frac{\delta_1}{h}\right)\mathcal{S}\left(\ell_1' g, \ell_2' h, d', \frac{N}{\delta\delta_1\delta_2\tilde{\ell}}, \frac{M}{\delta\delta_1\delta_2\tilde{\ell}}\right).\end{aligned}$$

Using only a trivial bound for the Hecke eigenvalues ($\lambda(n) \ll n^{1/2}$) and noting that $(\ell_1'\delta_2, \ell_2'\delta_1, d') = 1$, an application of (8.16) now completes the proof of Proposition 7.

9. Weyl Differencing

The rest of the paper is devoted to the proof of Theorem 5. We begin with the following differencing lemma.

Lemma 12. *Let the functions $b, b_{1i}, b_{2i} : \mathbb{Z} \to \mathbb{C}$ $(1 \leqslant i \leqslant I)$, $r_2 \in \mathbb{N}$, and $R_2 \in \mathbb{R}$ be such that*

$$b(m) = \sum_{i=1}^{I} b_{1i}(m) b_{2i}(m) \quad (m \in \mathbb{Z})$$

as well as

$$b_{2i}(m + r_2) = b_{2i}(m), \quad |b_{2i}(m)| \leqslant R_2 \quad (m \in \mathbb{Z},\ 1 \leqslant i \leqslant I).$$

Further, assume that the support of each b_{1i} is contained in $(A, A + M]$, and let $H \in \mathbb{N}$. Then

$$\Big|\sum_m b(m)\Big|^2 \ll \Big(Hr_2R_2^2 + \frac{R_2^2H^2r_2^2}{M}\Big) I \sum_{i=1}^{I} \sum_{A<m\leqslant A+M} |b_{1i}(m)|^2 + Hr_2R_2^2 I \sum_{0<|h|\leqslant \frac{M}{Hr_2}} \Big|\sum_{i=1}^{I} \sum_m b_{1i}(m + hHr_2)\overline{b_{1i}(m)}\Big|.$$

Proof. Let initially $b : \mathbb{Z} \to \mathbb{C}$ be arbitrary. We have

$$\begin{aligned}
&\sum_{A<m\leqslant A+M} \sum_{\substack{h\in\mathbb{Z}\\ A<m+hHr_2\leqslant A+M}} b(m + hHr_2) \\
&= \sum_{A<m\leqslant A+M} b(m) \cdot \#\big\{(m_1, h) : A < m_1 \leqslant A + M,\ m = m_1 + hHr_2\big\} \\
&= \sum_{A<m\leqslant A+M} b(m) \cdot \#\big\{h \in \mathbb{Z} : A < m - hHr_2 \leqslant A + M\big\} \\
&= \sum_{A<m\leqslant A+M} b(m) \Big(\frac{M}{Hr_2} + \mathrm{O}(1)\Big) = \frac{M}{Hr_2} \sum_{A<m\leqslant A+M} b(m) + \mathrm{O}\Big(\sum_{A<m\leqslant A+M} |b(m)|\Big).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{M^2}{H^2r_2^2}\Big|\sum_{A<m\leqslant A+M} b(m)\Big|^2 &\ll \Big|\sum_{A<m\leqslant A+M} \sum_{\substack{h\in\mathbb{Z}\\ A<m+hHr_2\leqslant A+M}} b(m + hHr_2)\Big|^2 + \Big(\sum_{A<m\leqslant A+M} |b(m)|\Big)^2 \\
&\ll M \sum_{A<m\leqslant A+M} \Big|\sum_{\substack{h\in\mathbb{Z}\\ A<m+hHr_2\leqslant A+M}} b(m + hHr_2)\Big|^2 + M \sum_{A<m\leqslant A+M} |b(m)|^2.
\end{aligned} \tag{9.1}$$

Let $b(m)$ be as in the statement of Lemma 12. Using the Cauchy-Schwarz inequality and applying (9.1) with $b_{[i]}(m) = b_{1i}(m)b_{2i}(m)$, we have that

$$\frac{M^2}{H^2r_2^2 I}\Big|\sum_{A<m\leqslant A+M} b(m)\Big|^2 \ll M \sum_{i=1}^{I} \sum_{A<m\leqslant A+M} \Big|\sum_{\substack{h\in\mathbb{Z}\\ A<m+hHr_2\leqslant A+M}} b_{[i]}(m+hHr_2)\Big|^2 + M \sum_{i=1}^{I} \sum_{A<m\leqslant A+M} |b_{[i]}(m)|^2.$$

Since each b_{2i} is r_2-periodic and bounded by R_2, we have for every individual i, m that

$$\begin{aligned}
\Big|\sum_{\substack{h\in\mathbb{Z}\\ A<m+hHr_2\leqslant A+M}} b_{[i]}(m + hHr_2)\Big|^2 &= \Big|b_{2i}(m) \sum_{\substack{h\in\mathbb{Z}\\ A<m+hHr_2\leqslant A+M}} b_{1i}(m + hHr_2)\Big|^2 \\
&\leqslant R_2^2\Big|\sum_{\substack{h\in\mathbb{Z}\\ A<m+hHr_2\leqslant A+M}} b_{1i}(m + hHr_2)\Big|^2.
\end{aligned}$$

Substituting this estimate above, we obtain

$$\begin{aligned}
&\frac{M^2}{H^2r_2^2I}\left|\sum_{A<m\leqslant A+M} b(m)\right|^2\\
&\quad\ll MR_2^2\sum_{i=1}^{I}\sum_{A<m\leqslant A+M}\left|\sum_h b_{1i}(m+hHr_2)\right|^2+M\sum_{i=1}^{I}\sum_m|b_{1i}(m)b_{2i}(m)|^2\\
&\quad\ll MR_2^2\sum_{i=1}^{I}\sum_{A<m\leqslant A+M}\sum_h|b_{1i}(m+hHr_2)|^2\\
&\qquad+MR_2^2\sum_{i=1}^{I}\sum_{A<m\leqslant A+M}\sum_{h_1\neq h_2}\sum b_{1i}(m+h_1Hr_2)\overline{b_{1i}(m+h_2Hr_2)}+MR_2^2\sum_{i=1}^{I}\sum_m|b_{1i}(m)|^2\\
&\ll MR_2^2\sum_{i=1}^{I}\sum_{A<m\leqslant A+M}|b_{1i}(m)|^2\left(\frac{M}{Hr_2}+\mathrm{O}(1)\right)\\
&\qquad+MR_2^2\sum_{0<|g|\leqslant\frac{M}{Hr_2}}\left|\sum_{i=1}^{I}\sum_{A<m\leqslant A+M}\sum_h b_{1i}\big(m+(h+g)Hr_2\big)\overline{b_{1i}(m+hHr_2)}\right|\\
&\ll MR_2^2\cdot\left(\frac{M}{Hr_2}+1\right)\sum_{i=1}^{I}\sum_m|b_{1i}(m)|^2+MR_2^2\sum_{0<|g|\leqslant\frac{M}{Hr_2}}\left|\sum_{i=1}^{I}\sum_m b_{1i}(m+gHr_2)\overline{b_{1i}(m)}\left(\frac{M}{Hr_2}+\mathrm{O}(1)\right)\right|\\
&\ll\left(\frac{M^2R_2^2}{Hr_2}+MR_2^2\right)\sum_{i=1}^{I}\sum_m|b_{1i}(m)|^2+\frac{M^2R_2^2}{Hr_2}\sum_{0<|g|\leqslant\frac{M}{Hr_2}}\left|\sum_{i=1}^{I}\sum_m b_{1i}(m+gHr_2)\overline{b_{1i}(m)}\right|,
\end{aligned}$$

using again that $\#\big\{(m_1,h):A<m_1\leqslant A+M,\,m=m_1+hHr_2\big\}=M/Hr_2+\mathrm{O}(1)$ as well as the Cauchy-Schwarz inequality to estimate the error terms in the off-diagonal summands. Rearranging, we conclude the lemma. □

The procedure used in the proof of Lemma 12, the "q-analogue of Weyl differencing", goes back at least to Postnikov [Po] and Heath-Brown [HB]. Similar ideas are also prominent in [PM]. The important point here is the generality in which the procedure applies: no particular structure (such as being a character, or an exponential of a rational function) is assumed for terms b_{1i} and b_{2i} beyond periodicity and a uniform bound for b_{2i}.

There are no conditions whatsoever on the coefficients $b_{1i}(m)$. In the applications we have in mind, however, the term $b_{1i}(m+gHr_2)\overline{b_{1i}(m)}$ will have a period that is a proper divisor of r. (This can happen for two reasons: either because b_{1i} are already periodic modulo a proper divisor of r, or because we take H to be a suitable divisor of r that causes a shortening of the period for the particular sequence b_{1i}.) On the other hand, the length of the m-summation in the off-diagonal terms in the upper bound of Lemma 12 is unchanged at M. In a typical situation, M may be too short compared to the original modulus r to expect any nontrivial bound (such as $M\asymp r^{1/2}$ or less with chaotically behaving summands $b(m)$), but its size may well be more favorable compared to the newly smaller modulus.

Finally, it will be important for our purposes that $b(m)$ is allowed to be a sum of finitely many terms $b_{1i}(m)b_{2i}(m)$ $(1\leqslant i\leqslant I)$ to which differencing is applied separately although the i-sum in the off-diagonal contribution to the upper bound is kept inside the absolute values. The case $I=1$, on the other hand, already contains the full idea of differencing.

Incomplete exponential sums whose length exceeds the square-root of the modulus, can often be efficiently estimated by the process sometimes referred to as completion. This procedure, which for clarity we record separately as the following simple technical result, applies in great generality, see [IK, Lemma 12.1]. For an r_1-periodic function $c : \mathbb{Z} \to \mathbb{C}$, let

$$\hat{c}(k) := \sum_{n=1}^{r_1} c(n) e\left(-\frac{nk}{r_1}\right)$$

be its discrete Fourier transform. The important point is that $\hat{c}(k)$ are complete exponential sums. (The notation for discrete Fourier transform in this section and the Mellin transform in earlier sections will not lead to confusion.)

Lemma 13. *Let $A \in \mathbb{Z}$, $r_1, M \in \mathbb{N}$, and let $c : \mathbb{Z} \to \mathbb{C}$ be such that $c(m + r_1) = c(m)$ for $m \in \mathbb{Z}$). Then*

$$\sum_{A < m \leqslant A+M} c(m) \ll \sum_{|k| \leqslant r_1/2} |\hat{c}(k)| \min\left(\frac{M}{r_1}, \frac{1}{|k|}\right).$$

Combining Lemmas 12 and 13, we have the following general result:

Theorem 10. *Let $r, r_1, r_2 \in \mathbb{N}$ be such that $r = r_1 r_2$. Let the functions $b, b_{1i}, b_{2i} : \mathbb{Z} \to \mathbb{C}$ ($1 \leqslant i \leqslant I$), $R_1, R_2 \in \mathbb{R}$ be such that*

$$b(m) = \sum_{i=1}^{I} b_{1i}(m) b_{2i}(m) \quad (m \in \mathbb{Z})$$

as well as

$$\begin{aligned} b_{1i}(m + r_1) &= b_{1i}(m), \quad |b_{1i}(m)| \leqslant R_1, \\ b_{2i}(m + r_2) &= b_{2i}(m), \quad |b_{2i}(m)| \leqslant R_2. \end{aligned} \qquad (m \in \mathbb{Z},\ 1 \leqslant i \leqslant I).$$

Let $H \in \mathbb{N}$, and let, for every $h, k \in \mathbb{Z}$ and $1 \leqslant i \leqslant I$,

$$\hat{B}_{1i,hH}(k) = \sum_{m \bmod r_1} b_{1i}(m + hHr_2)\overline{b_{1i}(m)} e\left(-\frac{km}{r_1}\right). \tag{9.2}$$

Then, for every $A \in \mathbb{Z}$, $M \in \mathbb{N}$,

$$\Big|\sum_{A < m \leqslant A+M} b(m)\Big|^2 \ll (M + Hr_2)Hr_2(R_1R_2)^2 I^2 + Hr_2 R_2^2 I \sum_{0 < |h| \leqslant \frac{M}{Hr_2}} \sum_{|k| \leqslant r_1/2} \Big|\sum_{i=1}^{I} \hat{B}_{1i,hH}(k)\Big| \min\left(\frac{M}{r_1}, \frac{1}{|k|}\right).$$

Proof. The proof is immediate from Lemmas 12 and 13. Specifically, we apply Lemma 12 to

$$b(m)\chi_{(A,A+M]}(m) = \sum_{i=1}^{I} \big(b_{1i}(m)\chi_{(A,A+M]}(m)\big) b_{2i}(m).$$

We estimate the resulting first, diagonal term trivially, while for off-diagonal terms we use Lemma 13 with the r_1-periodic function

$$c(m) = \sum_{i=1}^{I} b_{1i}(m + hHr_2)\overline{b_{1i}(m)}.$$

□

The role of the parameter H in Theorem 10 will become clear later. Importantly in the applications such as the central application for our problem, the sum defining $\hat{B}_{1i,hH}(k)$ is a complete exponential sum modulo r_1. Note that the trivial bound is

$$|\hat{B}_{1i,hH}(k)| \ll r_1 R_1^2,$$

so the trivial bound on the right-hand side is $\ll (R_1R_2)^2 I^2 (MHr_2 + (Hr_2)^2 + Mr_1 + M^2 \log r_1)$. This is, for general b_{1i}, a step backwards from the trivial bound $\ll M^2 (R_1R_2)^2 I^2$ on the left-hand side.

For arithmetically defined functions b_{1i}, however, the complete sum defining $\hat{B}_{1i,hH}(k)$ inherits this arithmetic structure. It will often be the case that the sum $\hat{B}_{1i,hH}(k)$ can be multiplicatively split in a certain sense. For r_1 a prime, the remaining complete sum can be estimated using techniques of algebraic geometry. For r_1 a higher prime power, the sum can be treated by the method of p-adic stationary phase. We remark that completion followed by the method of p-adic stationary phase acts as the proper p-adic analogue of the B-process in the classical van der Corput's theory of exponential sums [Mi]; see also [BM] for an example involving Kloosterman sums. In either case, for b_{1i} of algebro-geometric origin, we can often recover square-root cancellation in $\hat{B}_{1i,hH}(k)$.

10. Proof of Theorem 5

We now prepare for the proof of Theorem 5. We first make a small reduction to the case $q = r$ in the situation of Theorem 5. Indeed, suppose that Theorem 5 is proved in this special case, and write $q = rr'q'$ where $r' \mid r^\infty$ and $(q', r) = 1$. Then by Möbius inversion we have

$$\sum_{\substack{A<m\leqslant A+M\\(m,q)=1}} S(m,n_1,r)S(m,n_2,r) = \sum_{f\mid q'}\mu(f)\sum_{\substack{A/f<m\leqslant (A+M)/f\\(m,r)=1}} S(m,fn_1,r)S(m,fn_2,r)$$

so that the general case follows from the special case. Thus we are interested in the sequence $b(m)$ given by

$$b(m) = \begin{cases} S(m,n_1,r)S(m,n_2,r), & (m,r)=1,\\ 0, & (m,r)>1,\end{cases}$$

for integers n_1, n_2 (not necessarily coprime to r). From now on, we implicitly assume that $(m, r) = 1$. Moreover, the letter q is now free, and we will use it (in a different meaning than in the rest of paper) with or without indices as prime powers occurring in the prime factorization of r.

Before we apply Theorem 10 to this particular function $b(m)$, we explain briefly some technical difficulties. Kloosterman sums enjoy twisted multiplicativity, but of course only for coprime moduli. In order to apply Theorem 10, we need to decompose $r = r_1r_2$ with $(r_1, r_2) = 1$ and r_1, r_2 in certain ranges. However, if r is highly squareful (for example, if r is a pure prime power), such a decomposition may not be possible. In this case, however, one can choose the parameter H in Theorem 10 to be a suitable divisor of r_1, which produces partly degenerate Kloosterman sums and reduces the period of the sequence $b_{1i}(m+hHr_2)\overline{b_{1i}(m)}$, so that correspondingly $\widehat{B}_{1i,hH}(k)$ vanishes often (see Lemmas 19 and 20). In other words, the parameters H and r_2, each in its own way, act to make the range of summation in the off-diagonal terms in the upper bound of Lemma 12 more favorable compared to the period of the summands, but they apply separately, depending on the factorization of the modulus r. The previous discussion motivates a different treatment of the squarefree and the squareful part of r that we proceed to make precise now. We start with some notation.

Let $p > 2$ be a prime. For $\kappa \in \mathbb{N}$, we denote by $\mathbf{M}_{p^\kappa}$ an arbitrary element of $p^\kappa\mathbb{Z}_p$, which may be different from line to line. This notation serves as a p-adic analogue of Landau's O-notation in Taylor expansions. For $s \geqslant 1$ and $(A, p) = 1$, let

$$\tau(A,p^s) = \begin{cases} 1, & 2\mid s,\ p \text{ odd},\\ \left(\frac{A}{p}\right), & 2\nmid s,\ p\equiv 1 \pmod 4,\\ \left(\frac{A}{p}\right)i, & 2\nmid s,\ p\equiv 3 \pmod 4,\end{cases} \tag{10.1}$$

be the sign of the Gauß sum $\sum_{x \bmod p^s} e(Ax^2/p^s) = p^{s/2}\tau(A, p^s)$.

Next, we collect facts and notations pertaining to square roots to prime power moduli, which arise in connection with the explicit evaluation of Kloosterman sums as in Lemma 14. While these square roots

naturally arise in p-adic towers as in [BM], we keep our exposition elementary and only discuss square roots to a prime power modulus p^κ. This discussion applies separately at every odd prime p. For every $x \in (\mathbb{Z}/p^\kappa\mathbb{Z})^{\times 2}$, there are exactly two solutions $u \in (\mathbb{Z}/p^\kappa\mathbb{Z})^\times$ of the congruence $u^2 \equiv x \pmod{p^\kappa}$. Fix once and for all a choice function $s : (\mathbb{Z}/p\mathbb{Z})^{\times 2} \to (\mathbb{Z}/p\mathbb{Z})^\times$ such that, for every $r \in (\mathbb{Z}/p\mathbb{Z})^\times$, the class $s(r) \in (\mathbb{Z}/p\mathbb{Z})^\times$ satisfies $s(r)^2 \equiv r \pmod p$. Then, for every $x \in (\mathbb{Z}/p^\kappa\mathbb{Z})^\times$, we denote by $u_{1/2}^{[\kappa]}(x)$ the unique class $u \in (\mathbb{Z}/p^\kappa\mathbb{Z})^\times$ such that $u^2 \equiv x \pmod{p^\kappa}$ and $u \in s(x + p\mathbb{Z})$. This gives way to a unique function $u_{1/2}^{[\kappa]} : (\mathbb{Z}/p^\kappa\mathbb{Z})^{\times 2} \to (\mathbb{Z}/p^\kappa\mathbb{Z})^\times$, which we may think of as a branch of the square-root. (Each choice of s gives rise to a different branch of the square-root, but we will never need to consider other possible choices.) The values of $u_{1/2}^{[\kappa]}$ are compatible across different values of κ, in the sense that $u_{1/2}^{[\kappa_1]}(x) \equiv u_{1/2}^{[\kappa_2]}(x) \pmod{p^{\min(\kappa_1,\kappa_2)}}$, and hence we simply write $x_{1/2}$ for $u_{1/2}^{[\kappa]}(x)$ with a sufficiently high value of κ (for example, the highest power of p occurring as a modulus in the exponential sum of interest).

The following (essentially well-known) lemma appears for instance in [BM, Lemma 6].

Lemma 14. *Let $p > 2$ be a prime, let $s \geqslant 2$, and let $S(m, n; p^s)$ be the usual Kloosterman sum. Let $(m, p) = 1$ and $p^\nu \parallel n$. Then $S(m, n; p^s) = 0$ unless*

$$\nu = 0, \ mn \in (\mathbb{Z}/p\mathbb{Z})^{\times 2},$$

in which case it equals

$$S(m, n; p^s) = p^{s/2} \sum_{\pm} \tau\big(\pm (mn)_{1/2}, p^s\big) e\left(\pm \frac{2(mn)_{1/2}}{p^s}\right).$$

Suppose that $r = r_1 r_2$ with

$$(r_1, 6r_2) = 1,$$

and let

$$r_1 = \prod_{j=1}^{J} q_j, \quad q_j = p_j^{s_j}, \quad p_j > 3, \tag{10.2}$$

be the canonical factorization of r_1 into prime powers. We write

$$Q_j = r_1/q_j, \qquad Q_j \bar{Q}_j \equiv 1 \pmod{q_j}.$$

We denote all moduli $q_j = p_j^{s_j}$ with $s_j \geqslant 2$ as $q_1, \dots, q_\rho$, and for later purposes, we fix a divisor $r_1^\sharp$ of r_1 which will the product of some of the moduli q_j, $1 \leqslant j \leqslant \rho$. By rearranging, we may write

$$r_1^\sharp = \prod_{j=1}^{\varrho} q_j. \tag{10.3}$$

for some $\varrho \leqslant \rho$. By definition, $r_1^\sharp$ is squareful. For the moment, we do not impose any further condition on $r_1^\sharp$. (The final choice will satisfy $r_1^\sharp = (r_1, H^\infty)$, but the need for this choice will only become apparent later.)

Using the twisted multiplicativity of Kloosterman sums (which follows from the Chinese remainder theorem), we have that $b(m) = b_1(m)b_2(m)$ with

$$b_1(m) = S(\bar{r}_2 m, \bar{r}_2 n_1, r_1) S(\bar{r}_2 m, \bar{r}_2 n_2, r_1) = \prod_{j=1}^{J} S\big(m, \bar{Q}_j^2 \bar{r}_2^2 n_1, q_j\big) S\big(m, \bar{Q}_j^2 \bar{r}_2^2 n_2, q_j\big),$$

$$b_2(m) = S(\bar{r}_1 m, \bar{r}_1 n_1, r_2) S(\bar{r}_1 m, \bar{r}_1 n_2, r_2).$$

Keeping in mind that $(\bar{r}_1 m, r_2) = 1$, we have according to Weil's bound

$$|b_2(m)| \leqslant R_2 := d(r_2)^2 r_2. \tag{10.4}$$

Since $(2m, r_1) = 1$, we see from Lemma 14 that $b_1(m)$ vanishes unless

$$mn_1, mn_2, n_1n_2 \in (\mathbb{Z}/p_j\mathbb{Z})^{\times 2} \quad (1 \leqslant j \leqslant \varrho),$$

in which case $b_1(m)$ splits as a sum of $4^\varrho \ll r^\varepsilon$ terms, which we naturally index by $\boldsymbol{\epsilon} \in \{\pm 1\}^{2\times\varrho} = (\epsilon_{ij})_{i=1}^{2}{}_{j=1}^{\varrho}$ as follows:

$$b_1(m) = \sum_{\boldsymbol{\epsilon}\in\{\pm1\}^{2\times\varrho}} b_1^{\boldsymbol{\epsilon}}(m),$$

$$b_1^{\boldsymbol{\epsilon}}(m) = \prod_{j=1}^{\varrho} S^{\epsilon_{1j}}\big(m, \bar{Q}_j^2\bar{r}_2^2 n_1; q_j\big) S^{\epsilon_{2j}}\big(m, \bar{Q}_j^2\bar{r}_2^2 n_2; q_j\big) \prod_{j=\varrho+1}^{J} S\big(m, \bar{Q}_j^2\bar{r}_2^2 n_1; q_j\big) S\big(m, \bar{Q}_j^2\bar{r}_2^2 n_2; q_j\big),$$

$$S^{\epsilon}(m, n; p^s) = p^{s/2}\tau\big(\epsilon(mn)_{1/2}, p^s\big) e\left(\frac{2\epsilon(mn)_{1/2}}{p^s}\right) \quad (s \geqslant 2,\ mn \in (\mathbb{Z}/p\mathbb{Z})^{\times 2}).$$

Note that the Kloosterman sums $S(m, n, q)$ are real-valued, but the terms $S^{\epsilon}(m, n, p^s)$, in general, are not.

We are now ready to apply Theorem 10, with

$$r = r_1r_2, \quad b(m) = \sum_{\boldsymbol{\epsilon}\in\{\pm1\}^{2\times\varrho}} b_1^{\boldsymbol{\epsilon}}(m) b_2(m),$$

R_2 as in (10.4), and

$$R_1 = \max_{\boldsymbol{\epsilon}\in\{\pm1\}^{2\times\varrho}} \big|b_1^{\boldsymbol{\epsilon}}(m)\big| \ll d(r_1)^2 r_1.$$

We can conclude that

$$\begin{aligned} \Big|\sum_{\substack{A<m\leqslant A+M\\(m,r)=1}} S(m, n_1, r) S(m, n_2, r)\Big|^2 &\ll r^\varepsilon (M + Hr_2) Hr_2 r^2 \\ &+ r^\varepsilon Hr_2^3 \sum_{0<|h|\leqslant \frac{M}{Hr_2}} \sum_{|k|\leqslant \frac{r_1}{2}} \Big|\sum_{\boldsymbol{\epsilon}\in\{\pm1\}^{2\times\varrho}} \hat{B}_{1,hH}^{\boldsymbol{\epsilon}}(r_1, r_2, k)\Big| \min\left(\frac{M}{r_1}, \frac{1}{|k|}\right), \end{aligned} \tag{10.5}$$

where, as in (9.2), the terms $\hat{B}_{1,hH}^{\boldsymbol{\epsilon}}(r_1, r_2, k)$ are given by complete sums

$$\hat{B}_{1,hH}^{\boldsymbol{\epsilon}}(r_1, r_2, k) = \sideset{}{^*}\sum_{m \bmod r_1} b_1^{\boldsymbol{\epsilon}}\big(m + hHr_2\big)\overline{b_1^{\boldsymbol{\epsilon}}(m)} e\left(-\frac{km}{r_1}\right).$$

The sum of these terms $\hat{B}_{1,hH}^{\boldsymbol{\epsilon}}(r_1, r_2, k)$ is the central object of our estimation. We introduce some additional notation that allows us to state our results succinctly.

For $q = p^s$, $s \geqslant 2$, $n_1n_2 \in (\mathbb{Z}/q\mathbb{Z})^{\times 2}$, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) \in \{\pm1\}^4$, denote

$$\begin{aligned} \Sigma^{\boldsymbol{\epsilon}}(n_1, n_2, a, k; p^s) = \sideset{}{^*}\sum_{\substack{m \bmod p^s\\ m, m+a \in n_1(\mathbb{Z}/p\mathbb{Z})^{\times 2}}} & S^{\epsilon_1}(m + a, n_1, p^s)\overline{S^{\epsilon_2}(m, n_1, p^s)} \\ & S^{\epsilon_3}(m + a, n_2, p^s)\overline{S^{\epsilon_4}(m, n_2, p^s)} e\left(-\frac{km}{p^s}\right). \end{aligned} \tag{10.6}$$

For a general (prime or a) prime power q, we let

$$\Sigma(n_1, n_2, a, k; q) = \sum_{\substack{m \bmod q\\(m(m+a),q)=1}} S(m + a, n_1, q) S(m + a, n_2, q) S(m, n_1, q) S(m, n_2, q) e\left(-\frac{km}{q}\right). \tag{10.7}$$

Denote

$$A_0 = \{\pm1\}^4, \quad A^\sharp = \{\boldsymbol{\epsilon} \in A_0 : \epsilon_1 = \epsilon_2,\ \epsilon_3 = \epsilon_4\},$$

and, for an odd prime $q = p^s$ with $s \geqslant 2$,

$$\Sigma^\sharp(n_1, n_2, a, k; q) = \sum_{\boldsymbol{\epsilon} \in A^\sharp} \Sigma^{\boldsymbol{\epsilon}}(n_1, n_2, a, k; q), \quad \Sigma(n_1, n_2, a, k; q) = \sum_{\boldsymbol{\epsilon} \in A_0} \Sigma^{\boldsymbol{\epsilon}}(n_1, n_2, a, k; q).$$

We may rewrite the innermost sum in (10.5) as

$$\begin{aligned} \hat{B}_{1,hH}[r_1, r_2, k] &:= \sum_{\boldsymbol{\epsilon} \in \{\pm 1\}^{2\times \varrho}} \hat{B}^{\boldsymbol{\epsilon}}_{1,hH}(r_1, r_2, k) \\ &= \prod_{j=1}^{\varrho} \Sigma^\sharp(\bar{Q}_j^2 \bar{r}_2^2 n_1, \bar{Q}_j^2 \bar{r}_2^2 n_2, hHr_2, \bar{Q}_j k; q_j) \prod_{j=\varrho+1}^{J} \Sigma(\bar{Q}_j^2 \bar{r}_2^2 n_1, \bar{Q}_j^2 \bar{r}_2^2 n_2, hHr_2, \bar{Q}_j k; q_j). \end{aligned} \tag{10.8}$$

We see that it suffices to obtain upper bounds for the complete sums $\Sigma(n_1, n_2, a, k; q)$ and $\Sigma^\sharp(n_1, n_2, a, k; p^s)$ as above. These bounds are provided in the following result whose proof we postpone to the next section.

We need just a bit more notation. For an integer n let $\mathrm{rad}(n)$ denote its squarefree kernel and $\omega(n)$ the number of its prime factors. For a finite set T and $q \in \mathbb{N}$, we denote

$$(T, q) = \mathrm{lcm}\{(t, q) : t \in T\}, \tag{10.9}$$

and $n + T = \{n + t : t \in T\}$ as usual. Finally for a positive integer n we denote by $n_\square$ the largest integer whose square divides n. (In particular, for a prime power p^s we have $(p^s)_\square = p^{[s/2]}$.)

Then, collecting the results of Lemma 18, the decomposition (11.3), the reduction formula (11.6), and Lemmata 19 and 20, we obtain the following result.

Lemma 15. *Let $q = p^s$, where $p > 3$ is a prime and $s \geqslant 1$, and let $n_1, n_2, a, k \in \mathbb{Z}$.*

(1) *If $p \mid a$ and $s \geqslant 2$, then*

$$\Sigma^\sharp(n_1, n_2, a, k; q) \ll q^{5/2} \sum_{\substack{\delta \in \{1, (q, n_1 - n_2)\} \\ \delta \mid k,\ (\delta a, q/p) \mid k}} (q, \delta a, k)^{1/2}.$$

(2) *If $p \mid a$, or if $s = 1$, then*

$$\Sigma(n_1, n_2, a, k; q) \ll q^{5/2} \sum_{\substack{\delta \in \{1, (q, n_1 - n_2)\} \\ \delta \mid k}} \sum_{\substack{\delta' \in \{1, (q, a)\} \\ (\delta\delta', q/p) \mid k}} (q, \delta\delta', k)^{1/2}.$$

(3) *There exists a finite set $T \subset \mathbb{Z} \setminus p\mathbb{Z}$, of absolutely bounded size, depending on q, n_1, and n_2 only, such that, for every $k \in \mathbb{Z}$ and every $p \nmid a$,*

$$\Sigma(n_1, n_2, a, k; q) \ll q^{5/2} \sum_{\substack{\delta \in \{1, (q, n_1 - n_2)\} \\ \delta \mid k}} \delta^{1/2} \left(\left(\frac{k}{\delta}\right)^2 a - T, \left(\frac{q}{\delta}\right)_\square \right)^{1/2},$$

and the second factor in the sum may be omitted whenever q/δ is cube-free.

Proof. We show how Lemma 15 follows from the results of Section 11.

If $s = 1$, then Lemma 18 shows that $\Sigma(n_1, n_2, a, k; q) \ll q^{5/2}$, except if $q \mid (a(n_1 - n_2))$ and $q \mid k$, in which case the upper bound obtained is $\Sigma(n_1, n_2, a, k; q) \ll q^3$. This estimate is absorbed by the upper bound in (2), specifically by the term corresponding to $\delta = \delta' = 1$ in the former and by the term corresponding to $\delta = (q, n_1 - n_2)$, $\delta' = (q, a)$ in the latter case. Moreover, if $s = 1$ and $q \nmid a$, the estimate of Lemma 18 is also allowable in (3) with $\delta = (q, n_1 - n_2, k)$.

Consider now the case $s \geqslant 2$. According to the decomposition (11.3), the sum $\Sigma^A(n_1, n_2, a, k; q)$ (with $A \in \{A_0, A^\sharp\}$ and $A = A^\sharp$ only if $p \mid a$) can be written as a finite linear combination

$$\Sigma = q^2 \sum_{\boldsymbol{\epsilon} \in A} \tau^{[\boldsymbol{\epsilon}]} \Sigma[A^{[\boldsymbol{\epsilon}]}(n_1, n_2), B^{[\boldsymbol{\epsilon}]}(n_1, n_2), a, k; q],$$

with the parameters $A = A^{[\epsilon]}(n_1, n_2)$ and $B = B^{[\epsilon]}(n_1, n_2)$ given explicitly as in (11.2) and the sum $\Sigma[A, B, a, k; q]$ defined in (11.4). The contribution of terms with $p \nmid A$ or $p \nmid B$ can be estimated by Lemmas 19 and 20 and absorbed in the terms corresponding to $\delta = 1$ in (1)–(3) above as follows:

- If $p \nmid a$, we apply Lemma 20 (1) to obtain (3) (expanding T to account for all choices of A and B).
- If $p \mid a$, we estimate the terms with $A \equiv B \pmod p$ (and then, as will be seen from (11.2), $A = B$) and $A \not\equiv B \pmod p$ separately. For the terms in which $A = B$, which are the only ones that arise in the estimation of $\Sigma^\sharp(n_1, n_2, a, k; q)$, we apply Lemma 19 and obtain (1) and the terms in (2) with $\delta' = (q, a)$. For the terms in which $A \not\equiv B \pmod p$, we apply Lemma 20 (2) and obtain the terms in (2) with $\delta' = 1$.

Terms with $p \mid A$ and $p \mid B$ will be seen to appear if and only if $p \mid (q, n_1 - n_2)$, in which case, denoting $\delta = (q, n_1 - n_2)$, we have $\delta \parallel A, B$, and $\Sigma[A, B, a, k; q] = 0$ unless $\delta \mid k$. If $\delta = q$, then all of (1)–(3) hold for the trivial reason that all upper bounds are at least q^3. Otherwise, by applying the reduction formula (11.6), we have that

$$\Sigma[A, B, a, k; q] = \delta \cdot \Sigma[A/\delta, B/\delta, a, k/\delta; q/\delta],$$

where $p \nmid (A/\delta)$ and $p \nmid (B/\delta)$. The remaining sum is treated as above and is seen to be bounded by the terms corresponding to $\delta = (q, n_1 - n_2)$ in (1)–(3). □

Applying Lemma 15 to the individual factors in (10.8), and with a quick application of the Chinese Remainder Theorem, we obtain the following crucial estimate.

Proposition 11. *Let $r = r_1 r_2$ with $(r_1, 6r_2) = 1$, and let $r_1^\sharp$ be a squareful divisor of r_1, with factorizations of r_1 and $r_1^\sharp$ as in (10.2) and (10.3). Let h and H be non-zero integers with $r_1^\sharp \mid (hH)^\infty$, and let $k \in \mathbb{Z}$. Write*

$$\tilde{r}_1 := \prod_{\substack{q_j \parallel r_1,\ \mu(q_j)=0,\\ (q_j, hH)=1}} q_j, \quad r_1 = r_1^\flat \tilde{r}_1;$$

in particular, $r_1^\sharp \mid r_1^\flat$. Then, there exists for every $\tilde{\delta} \mid \tilde{r}_1$ a set $T_{\tilde{\delta}}$, of cardinality $\mathrm{O}\big(C^{\omega(\tilde{r}_1)}\big)$ for some absolute constant C, with elements depending on $r_1, \tilde{r}_1, \tilde{\delta}, n_1, n_2$ only, and with all elements coprime to $\tilde{r}_1$, such that the sum $\hat{B}_{1,hH}[r_1, r_2, k]$ defined in (10.8) satisfies

$$\hat{B}^{\epsilon}_{1,hH}[r_1, r_2, k] \ll r_1^{5/2} \sum_{\delta^\flat \mid (r_1^\flat, n_1 - n_2, k)} \sum_{\substack{(r_1^\sharp, hH) \mid \delta' \mid (r_1^\flat, hH)\\ (\delta^\flat \delta', r_1^\flat / \operatorname{rad} r_1^\flat) \mid k}} \sum_{\tilde{\delta} \mid (\tilde{r}_1, n_1 - n_2, k)} (r_1, \delta^\flat \delta' \tilde{\delta}, k)^{1/2} \left(\left(\frac{k}{\tilde{\delta}}\right)^2 hHr_2 - T_{\tilde{\delta}}, \left(\frac{\tilde{r}_1}{\tilde{\delta}}\right)_\square \right)^{1/2},$$

where the second factor may be omitted whenever $\tilde{r}_1/\tilde{\delta}$ is cube-free.

With Proposition 11, we are ready for the proof of Theorem 5. Denote the sum to be estimated as

$$S = \sum_{\substack{A < m \leqslant A + M\\ (m, r) = 1}} S(m, n_1, r) S(m, n_2, r).$$

Fix a decomposition $r = r_1 r_2$ with $(r_1, 6r_2) = 1$ and a divisor

$$H \mid \frac{r_1}{\operatorname{rad} r_1},$$

both to be suitably specified later. We set

$$r_1^\sharp = (r_1, H^\infty).$$

It is then clear that $r_1^\sharp$ is a squareful divisor of r_1 of the type considered in (10.3), and that $r_1^\sharp \mid H^\infty$.

Using the basic estimate on S in (10.5) and Proposition 11, we have that

$$|S|^2 \ll r^\varepsilon (M+Hr_2)Hr_2r^2 + r^\varepsilon Hr_1^{5/2}r_2^3 \sum_{\substack{r_1=r_1^\flat \tilde{r}_1\\ H\mid r_1^\flat,\,(r_1^\flat,\tilde{r}_1)=1}} \sum_{d^\flat \mid (r_1^\flat, n_1-n_2)} \sum_{\tilde{d}\mid(\tilde{r}_1,n_1-n_2)} \sum_{\substack{0<|h|\leqslant \frac{M}{Hr_2}\\ (h,\tilde{r}_1)=1}} \sum_{(r_1^\sharp,hH)\mid d'\mid (r_1^\flat,hH)}$$

$$\sum_{\substack{|k|\leqslant r_1/2,\ d^\flat\tilde{d}\mid k,\\ (d^\flat d', r_1^\flat/\operatorname{rad}(r_1^\flat))\mid k}} (r_1, d^\flat d'\tilde{d}, k)^{1/2} \min\left(\frac{M}{r_1}, \frac{1}{|k|}\right) \left(\left(\frac{k}{\tilde{d}}\right)^2 hHr_2 - T_{\tilde{d}}, \left(\frac{\tilde{r}_1}{\tilde{d}}\right)_\square\right)^{1/2}$$

$$= r^\varepsilon (M+Hr_2)Hr_2r^2 + r^\varepsilon Hr_1^{5/2}r_2^3 \sum_{\substack{r_1=r_1^\flat \tilde{r}_1\\ H\mid r_1^\flat,\,(r_1^\flat,\tilde{r}_1)=1}} \sum_{d^\flat \mid (r_1^\flat, n_1-n_2)} \sum_{\tilde{d}\mid(\tilde{r}_1,n_1-n_2)} \sum_{\substack{0<|h|\leqslant \frac{M}{Hr_2}\\ (h,\tilde{r}_1)=1}} \sum_{(r_1^\sharp,hH)\mid d'\mid (r_1^\flat,hH)}$$

$$\sum_{\substack{d',\,(d^\flat d', r_1^\flat/\operatorname{rad}(r_1^\flat))\mid d^\sharp\\ d^\sharp\mid (d^\flat d', r_1^\flat)}} (d^\sharp\tilde{d})^{1/2} \sum_{|\ell|\leqslant \frac{r_1}{2d^\sharp\tilde{d}}} \min\left(\frac{M}{r_1}, \frac{1}{d^\sharp\tilde{d}|\ell|}\right) \left(d^{\sharp 2}\ell^2 hHr_2 - T_{\tilde{d}}, \left(\frac{\tilde{r}_1}{\tilde{d}}\right)_\square\right)^{1/2}.$$

We collect various contributions to the right-hand side. The contribution of the terms with $\ell = 0$ is

$$\begin{aligned}
&\ll r^\varepsilon Hr_1^{5/2}r_2^3 (r_1, H(n_1-n_2))^{1/2} \cdot \frac{M}{Hr_2}\cdot\frac{M}{r_1}\\
&\ll r^\varepsilon M^2 H^{1/2} r_2^2 r_1^{3/2} (r_1, n_1-n_2)^{1/2}\\
&\ll r^\varepsilon M^2 r^{3/2} (Hr_2)^{1/2} (r, n_1-n_2)^{1/2}.
\end{aligned}$$

As for the contributions of the terms with $h, \ell \neq 0$, we majorize the contribution of the four innermost (h, d', $d^\sharp$, and ℓ) sums above by

$$\begin{aligned}
&\ll \sum_{H\mid d'\mid r_1^\flat} \sum_{d'\mid d^\sharp \mid r_1^\flat} \sum_{0<|\ell|\leqslant \frac{r_1}{2d^\sharp\tilde{d}}} \frac{1}{(d^\sharp\tilde{d})^{1/2}|\ell|} \sum_{0<|h|\leqslant \frac{M}{Hr_2}} \left(h - (\overline{d^{\sharp 2}\ell^2 Hr_2}\cdot T_{\tilde{d}}), \left(\frac{\tilde{r}_1}{\tilde{d}}\right)_\square\right)^{1/2}\\
&\ll \frac{r^\varepsilon}{H^{1/2}} \sum_{\delta\mid(\tilde{r}_1/\tilde{d})_\square} \delta^{1/2}\left(1+\frac{M}{Hr_2\delta}\right) \ll \frac{r^\varepsilon}{H^{1/2}}\left((\tilde{r}_1)_\square^{1/2} + \frac{M}{Hr_2}\right).
\end{aligned}$$

We remark that, if r_1 (and hence $\tilde{r}_1$) is cube-free, then the term involving $(\tilde{r}_1)_\square^{1/2}$ may be omitted.

Executing the outside three ($r_1 = r_1^\flat\tilde{r}_1$, $d^\flat$, and $\tilde{d}$) summations and collecting all terms, we have that

$$|S|^2 \ll r^\varepsilon (M+Hr_2)Hr_2r^2 + r^\varepsilon M^2 r^{3/2}(Hr_2)^{1/2}(r, n_1-n_2)^{1/2} + r^\varepsilon r^{5/2}(Hr_2)^{1/2}(r_1/r_1^\sharp)_\square^{1/2} + r^\varepsilon \frac{Mr^{5/2}}{(Hr_2)^{1/2}}.$$

This estimate holds for every decomposition $r = r_1r_2$ with $(r_1, 6r_2) = 1$ and every divisor $H \mid (r_1/\operatorname{rad} r_1)$. Note that the upper bound depends only on the product Hr_2 rather than on the individual factors of H and r_2. Conceptually, this comes as no surprise, since the product Hr_2 was used as the single differencing step in Lemma 12. Also, note that $r_1/r_1^\sharp = r/(r, (Hr_2)^\infty)$.

This brings us to the statement of Theorem 5. For a given divisor $s \mid r$ satisfying $(r, 6^\infty) \mid s$, define

$$H = \left(s, \left(s, \frac{r}{s}\right)^\infty\right), \quad r_2 = \frac{s}{H}, \quad r_1 = \frac{r}{r_2}.$$

This choice of H and the decomposition $r = r_1r_2$ satisfy all our conditions, and we have proved

$$|S|^2 \ll r^\varepsilon Mr^2 s + r^\varepsilon \frac{Mr^{5/2}}{s^{1/2}} + r^\varepsilon r^2 s^2 + r^\varepsilon M^2 r^{3/2} s^{1/2} (r, n_1-n_2)^{1/2} + \sigma^2,$$

where

$$\sigma^2 = r^{\varepsilon} r^{5/2} s^{1/2} \left(\frac{r}{(r, s^{\infty})} \right)_{\square}^{1/2} \tag{10.10}$$

satisfies all the stated properties. This completes the proof of Theorem 5. □

11. Estimation of complete sums

11.1. **Preliminaries.** We start with two important lemmas that we will use at several stages of the fairly long and technical proof of Lemma 15. The following lemma is a special case of [Bo, Theorem 5] which is already implicit in Weil's work.

Lemma 16. *Let p be a prime, and let $f_1, f_2 \in (\mathbb{Z}/p\mathbb{Z})[x]$ be two coprime polynomials, not both of which are constant. Then*

$$\Big| \sum_{\substack{x \bmod p \\ f_2(x) \not\equiv 0 \bmod p}} e\left(\frac{f_1(x)\bar{f}_2(x)}{p} \right) \Big| \leqslant (\deg f_1 + 2\deg f_2 - 1)\sqrt{p} + 1.$$

The next lemma is of Hensel type.

Lemma 17. *Let $1 \leqslant \kappa \leqslant \lambda$, $A \subseteq \mathbb{Z}/p^{\lambda}\mathbb{Z}$, $A + p^{\kappa}\mathbb{Z} \subseteq A$, $f : A \to \mathbb{Z}/p^{\lambda}\mathbb{Z}$, $f_1 : A \to (\mathbb{Z}/p^{\lambda}\mathbb{Z})^{\times}$ be such that*

$$f(m + p^{\mu}t) - f(m) - p^{\mu} f_1(m) t \in p^{\mu+1}\mathbb{Z}/p^{\lambda}\mathbb{Z}$$

for all $m \in A$, $t \in \mathbb{Z}$, and $\kappa \leqslant \mu < \lambda$. Then, for all $\kappa \leqslant \mu \leqslant \lambda$, the number $K(p^{\mu})$ of solutions of the congruence

$$f(m) \equiv \omega \pmod{p^{\mu}}$$

in $m \in A$ modulo $p^{\mu}\mathbb{Z}$ satisfies

$$K(p^{\mu}) = K(p^{\kappa}).$$

Before heading to the proof, we remark that, in applications of Lemma 17, the condition that $f_1(m) \subseteq (\mathbb{Z}/p\mathbb{Z})^{\times}$ only needs to be checked for $m \in A$ satisfying $f(m) \equiv \omega \pmod{p^{\kappa}}$. This is immediate from the proof but also follows from the statement by applying it with the restricted domain $A \cap f^{-1}(\omega)$.

Proof. Let $\kappa \leqslant \mu < \lambda$. We prove that $K(p^{\mu}) = K(p^{\mu+1})$. Indeed, let $m \in A$ be such that $f(m) \equiv \omega \pmod{p^{\mu}}$. Every solution $m_1 \in A$ modulo $p^{\mu+1}$ such that $m_1 \equiv m \pmod{p^{\mu}}$ is of the form $m + p^{\mu}t$ for some $t \in \mathbb{Z}/p\mathbb{Z}$. According to the condition of the problem, we have that

$$f(m + p^{\mu}t) - f(m) - p^{\mu} f_1(m) t \in p^{\mu+1}\mathbb{Z}/p^{\lambda}\mathbb{Z}.$$

We are given that $f(m) \equiv \omega \pmod{p^{\mu}}$, so we can write $f(m) \equiv \omega + p^{\mu}F_m \pmod{p^{\lambda}}$ for some $F_m \in \mathbb{Z}/p^{\lambda-\mu}\mathbb{Z}$. In light of the above display, the congruence $f(m_1) \equiv \omega \pmod{p^{\mu+1}}$ is equivalent to

$$\begin{aligned} \omega + p^{\mu}F_m + p^{\mu} f_1(m) t &\equiv \omega \pmod{p^{\mu+1}}, \\ f_1(m) t &\equiv -F_m \pmod{p}. \end{aligned}$$

Since $f_1(m) \in (\mathbb{Z}/p^{\lambda}\mathbb{Z})^{\times}$, we above congruence is equivalent to $t \equiv -\overline{f_1(m)}F_m \pmod{p}$, and hence

$$m_1 \equiv m + p^{\mu}t \equiv m - p^{\mu}\overline{f_1(m)}F_m \pmod{p^{\mu+1}}.$$

Denoting by $A(p^{\mu})$ the set of solutions of $f(m) \equiv \omega \pmod{p^{\mu}}$ in $m \in A$ modulo $p^{\mu}\mathbb{Z}_p$, this shows in one move that the canonical reduction map $A(p^{\mu+1}) \to A(p^{\mu})$ is both surjective and injective; hence $K(p^{\mu}) = K(p^{\mu+1})$. The equality $K(p^{\mu}) = K(p^{\kappa})$ for every $\kappa \leqslant \mu \leqslant \lambda$ follows immediately. □

11.2. **Prime case.** We now turn to the estimation of $\Sigma(n_1, n_2, a, k; q)$ for q prime. The following result settles the second half of Lemma 15(2). A more general version is contained in the forthcoming preprint [FKM].

Lemma 18. *Let q be a prime, and let $n_1, n_2, a, k \in \mathbb{Z}$. Then, the sum $\Sigma(n_1, n_2, a, k; q)$ defined in (10.7) satisfies the bound*

$$\Sigma(n_1, n_2, a, k; q) \ll q^{5/2}(q, a(n_1 - n_2), k)^{1/2}$$

with an absolute implied constant.

Proof. Let us first assume that $q \mid n_1$, but $q \nmid n_2$. Then by Weil's bound for Kloosterman sums and standard bounds for Ramanujan sums we have

$$|\Sigma(n_1, n_2, a, k; q)| \leqslant 4q \sum_{\substack{m \bmod q \\ (m(m+a), q)=1)}} |S(m+a, 0, q)S(m, 0, q)| \leqslant 4q^2.$$

The same bound holds by symmetry if $q \nmid n_1$, but $q \mid n_2$. Similarly, if $q \mid n_1$ and $q \mid n_2$, then

$$|\Sigma(n_1, n_2, a, k; q)| \leqslant \sum_{\substack{m \bmod q \\ (m(m+a), q)=1}} |S(m+a, 0, q)S(m, 0, q)|^2 \leqslant q.$$

This leaves us with the generic case $q \nmid n_1 n_2$. Here, $\Sigma(n_1, n_2, a, k; q) = q^2\Sigma^\circ$ where

$$\Sigma^\circ = \sum_{\substack{m \bmod q \\ (m(m+a), q)=1}} \mathrm{Kl}_2(n_1(m+a), q)\mathrm{Kl}_2(n_2(m+a), q)\mathrm{Kl}_2(n_1 m, q)\mathrm{Kl}_2(n_2 m, q) e\left(-\frac{km}{q}\right)$$

where $\mathrm{Kl}_2(m, q) = q^{-1/2}S(1, m, q)$. If $q \mid a(n_1 - n_2)$ and $q \mid k$, then we estimate trivially with Weil's bound, getting the bound $|\Sigma^\circ| \leqslant 16q$.

On the hand, if $q \nmid a(n_1 - n_2)$ or $q \nmid k$, then we use independence of Kloosterman sheafs (as developed by Katz). We use this in the form of the explicit result on uniform distribution of angles of Kloosterman sums due to Fouvry–Michel–Rivat–Sárkőzy [FMRS] (see also [FGKM, Proposition 3.2]). Among the four linear forms $\ell_1(m) = n_1(m+a)$, $\ell_2(m) = n_2(m+a)$, $\ell_3(m) = n_1 m$, and $\ell_4(m) = n_2 m$, there may be four, two, or one distinct form(s) modulo q, depending on whether neither, one, or both of $q \mid a$ and $q \mid (n_1 - n_2)$ hold. We group terms corresponding to the same forms together and find a finite set $\mathcal{L}$ of linear forms over $\mathbb{F}_q$ and integers $\lambda_\ell \in \{1, 2, 4\}$ such that

$$\Sigma^\circ = \sum_{\substack{m \in \mathbb{F}_q \\ \ell(m) \neq 0\ (\forall \ell \in \mathcal{L})}} \prod_{\ell \in \mathcal{L}} \mathrm{Kl}_2(\ell(m), q)^{\lambda_\ell} e\left(-\frac{km}{q}\right).$$

Writing $\mathrm{Kl}_2(\ell(m), q) = 2\cos\theta(\ell(m))$ and using elementary trigonometry, the term

$$\mathrm{Kl}_2(\ell(m), q)^{\lambda_\ell} = [2\cos\theta(\ell(m))]^{\lambda_\ell}$$

can be rewritten as a finite linear combination of $\mathrm{sym}_k \theta(\ell(m))$ for some $|k| \leqslant \lambda_\ell$, $k \equiv \lambda_\ell \pmod 2$, where $\mathrm{sym}_k \theta = \sin((k+1)\theta)/\sin\theta$. Corresponding to this, the sum Σ° can be written as a finite linear combination (with coefficients of absolutely bounded size) of sums of the form

$$\Sigma_j^\circ = \sum_{\substack{m \in \mathbb{F}_q \\ \ell(m) \neq 0\ (\forall \ell \in \mathcal{L}}} \prod_{\ell \in \mathcal{L}} \mathrm{sym}_{k_{\ell,j}}(\theta(\ell(m))) e\left(-\frac{km}{q}\right)$$

for some $|k_{\ell,j}| \leqslant \lambda_\ell$, $k_{\ell,j} \equiv \lambda_\ell \pmod 2$.

According to [FMRS, Lemma 2.1], we have the estimate

$$\Sigma_j^\circ \ll q^{1/2}$$

as long as it is not the case that all $k_{\ell,j} = 0$ for all $\ell \in \mathcal{L}$ *and* $k = 0$ in $\mathbb{F}_q$. This is ensured by our non-degeneracy condition that $q \nmid a(n_1 - n_2)$ or $q \nmid k$; in the former case, $|\mathcal{L}| = 4$ and $|k_{\ell,j}| = \lambda_\ell = 1$ for all $\ell \in \mathcal{L}$, while, if $q \nmid k$, then $k \neq 0$ in $\mathbb{F}_p$. Putting everything together, we have that

$$\Sigma\big(n_1, n_2, a, k; q\big) \ll q^{5/2}$$

if $q \nmid n_1 n_2$ and if in addition $q \nmid a(n_1 - n_2)$ or $q \nmid k$, with an absolute implied constant. □

11.3. **Setup of the prime power case.** In the case of squareful moduli, the estimation of the multiple exponential sum $\Sigma(n_1, n_2, a, k; q)$ requires the deep tools of algebraic geometry only in some degenerate cases, but nevertheless (or because of this) the argument turns out to be very involved. In this subsection, we prepare ground for this estimation by reducing and decomposing the problem to one of the two distinctly different cases.

We are considering a sum of the form

$$\begin{aligned}\Sigma := \Sigma^A(n_1, n_2, a, k; p^s) = \sideset{}{^*}\sum_{\substack{m \bmod p^s \\ m, m+a \in n_1(\mathbb{Z}/p\mathbb{Z})^{\times 2}}} \sum_{\boldsymbol{\epsilon} \in A} S^{\epsilon_1}(m+a, n_1; p^s)\overline{S^{\epsilon_2}(m, n_1; p^s)} \\ S^{\epsilon_3}(m+a, n_2; p^s)\overline{S^{\epsilon_4}(m, n_2; p^s)} e\left(-\frac{km}{p^s}\right),\end{aligned} \tag{11.1}$$

where $A \in \{A_0, A^\sharp\}$, $A = A^\sharp$ only if $p \mid a$,

$$\begin{aligned}S^\epsilon(m, n; p^s) &= p^{s/2}\tau(\epsilon \cdot (mn)_{1/2}, p^s) e\left(\frac{2\epsilon \cdot (mn)_{1/2}}{p^s}\right) \\ &= p^{s/2}\tau(\epsilon \cdot \epsilon(mu, n\bar{u})(mu)_{1/2}(n\bar{u})_{1/2}, p^s) e\left(\frac{2\epsilon \cdot \epsilon((mu)_{1/2}, (n\bar{u})_{1/2})(mu)_{1/2}(n\bar{u})_{1/2}}{p^s}\right),\end{aligned}$$

and $u \in (\mathbb{Z}/p\mathbb{Z})^\times$ is a fixed representative of the class $n_1(\mathbb{Z}/p\mathbb{Z})^{\times 2} = n_2(\mathbb{Z}/p\mathbb{Z})^{\times 2}$.

Considering the product of the τ-factors in (11.1), we note that

$$\begin{aligned}T(m, n_1, n_2, a; p^s) = \tau\big(\epsilon_1((m+a)u)_{1/2}(n_1\bar{u})_{1/2}, p^s\big)\overline{\tau\big(\epsilon_2(mu)_{1/2}(n_1\bar{u})_{1/2}, p^s\big)} \\ \tau\big(\epsilon_3((m+a)u)_{1/2}(n_2\bar{u})_{1/2}, p^s\big)\overline{\tau\big(\epsilon_4(mu)_{1/2}(n_2\bar{u})_{1/2}, p^s\big)} = \tau^{[\boldsymbol{\epsilon}]}\end{aligned}$$

depends only on the product $\epsilon_1\epsilon_2\epsilon_3\epsilon_4$ and the parity of s (using the explicit formula (10.1) for the sign of the Gauß sum). By relabeling $\boldsymbol{\epsilon}$ as necessary, we can write

$$\Sigma = \sum_{\boldsymbol{\epsilon} \in A}^* \tau^{[\boldsymbol{\epsilon}]} p^{2s} \sideset{}{^*}\sum_{\substack{m \bmod p^s \\ m, m+a \in n_1(\mathbb{Z}/p\mathbb{Z})^{\times 2}}} e\left(\frac{f^{[\boldsymbol{\epsilon}]}(m, n_1, n_2, a, k)}{p^s}\right).$$

Here, we have denoted

$$\begin{aligned}f^{[\boldsymbol{\epsilon}]}(m, n_1, n_2, a, k) &= 2\epsilon_1((m+a)u)_{1/2}(n_1\bar{u})_{1/2} - 2\epsilon_2(mu)_{1/2}(n_1\bar{u})_{1/2} \\ &\quad + 2\epsilon_3((m+a)u)_{1/2}(n_2\bar{u})_{1/2} - 2\epsilon_4(mu)_{1/2}(n_2\bar{u})_{1/2} - km \\ &= 2A\big((m+a)u\big)_{1/2} - 2B(mu)_{1/2} - km,\end{aligned}$$

where

$$A = A^{[\boldsymbol{\epsilon}]}(n_1, n_2) = \epsilon_1(n_1\bar{u})_{1/2} + \epsilon_3(n_2\bar{u})_{1/2}, \quad B = B^{[\boldsymbol{\epsilon}]}(n_1, n_2) = \epsilon_2(n_1\bar{u})_{1/2} + \epsilon_4(n_2\bar{u})_{1/2}. \tag{11.2}$$

Corresponding to the above, we may further write

$$\Sigma = p^{2s} \sum_{\boldsymbol{\epsilon} \in A} \tau^{[\boldsymbol{\epsilon}]} \Sigma[A^{[\boldsymbol{\epsilon}]}(n_1, n_2), B^{[\boldsymbol{\epsilon}]}(n_1, n_2), a, k; p^s], \tag{11.3}$$

where we write more generally

$$\Sigma[A,B,a,k;p^s]=\sum_{\substack{m \bmod p^s\\ m,m+a\in u(\mathbb{Z}/p\mathbb{Z})^{\times 2}}}^{*} e\left(\frac{f[m,A,B,a,k]}{p^s}\right) \tag{11.4}$$

and

$$f[m,A,B,a,k]=2A\big((m+a)u\big)_{1/2}-2B(mu)_{1/2}-km. \tag{11.5}$$

Note that, in any case,

$$A^2-B^2\in\big\{0,\pm 4(n_1\bar u)_{1/2}(n_2\bar u)_{1/2}\big\}.$$

We also make the important remark that

$$\big((n_1\bar u)_{1/2}+(n_2\bar u)_{1/2}\big)\big((n_1\bar u)_{1/2}-(n_2\bar u)_{1/2}\big)=\bar u(n_1-n_2).$$

This shows that $A\equiv 0 \pmod p$ or $B\equiv 0 \pmod p$ is possible only if $n_1\equiv n_2 \pmod p$. Moreover, if $p^\nu \parallel (n_2-n_1)$, then $p^\nu \parallel A$ if $\epsilon_3=-\epsilon_1$ and $p\nmid A$ otherwise, and analogously for B; this also formally holds for $\nu=\infty$.

Suppose that $\nu>0$ and $p\mid A,B$; then, $p^\nu\parallel A,B$. It is immediate that $\Sigma[A,B,a,k;p^s]=0$ unless $p^{\nu'}\mid k$, where $\nu'=\min(\nu,s)$. From now on, assume that $p^{\nu'}\mid k$. It is also obvious that, if $\nu\geqslant s$, then $\Sigma[A,B,a,k;p^s]=p^s$. If, on the other hand, $\nu<s$ and $p^\nu\parallel A$, $p^\nu\parallel B$, then

$$\Sigma[A,B,a,k;p^s]=p^\nu\cdot\Sigma\left[\frac{A}{p^\nu},\frac{B}{p^\nu},a,\frac{k}{p^\nu};p^{s-\nu}\right]. \tag{11.6}$$

Therefore, it suffices to prove an estimate for the sum $\Sigma[A,B,a,k;p^s]$ defined in (11.4) (or a finite $\boldsymbol{\epsilon}$-average thereof) for $p\nmid A$ or $p\nmid B$, and for $s\geqslant 1$. We consider the following two situations separately, keeping as a standing condition that $p\nmid A$ or $p\nmid B$.

The case when $p\mid a$ and $A\equiv B \pmod p$ is addressed in Section 12. Note that, in this case, actually $A=B$. Referring back to (11.5), we see that this case is distinguished in that the branches of the square-root in $\big((m+a)u\big)_{1/2}$ and $(mu)_{1/2}$ are aligned so that the leading terms cancel out and, as will be seen, an additional factor of size $|a|_p$ emerges.

The remaining cases, when $p\mid a$ and $A\not\equiv B \pmod p$ as well as when $p\nmid a$, are treated in Section 13. In this case, no particular alignment of square-roots occurs, but Hensel liftings become much more delicate, and, if $p\nmid a$, singular critical points are encountered in the stationary phase analysis.

The final results of Sections 12 and 13 are the following Lemmas 19 and 20, respectively.

Lemma 19. *Let $q=p^s$, where $p>3$ is a prime and $s\geqslant 1$, and let $A,a,k\in\mathbb{Z}$ with $p\mid a$. Then, the sum $\Sigma[A,A,a,k;q]$ defined in (11.4) satisfies*

$$\sum_{\epsilon\in\{\pm 1\}}\Sigma[\epsilon A,\epsilon A,a,k;q]\ll q^{1/2}(q,Aa,k)^{1/2}$$

with an absolute implied constant. Moreover, the left-hand side vanishes unless $(Aa,q/p)\mid k$.

Lemma 20. *Let $q=p^s$, where $p>3$ is a prime and $s\geqslant 1$, and let $A,B\in\mathbb{Z}$ be such that $p\nmid A$ or $p\nmid B$.*

(1) *There exists a finite set $T\subset\mathbb{Z}\setminus p\mathbb{Z}$, of absolutely bounded size, depending on q, A, and B only, such that, for every $k\in\mathbb{Z}$ and every $p\nmid a$,*

$$\sum_{\boldsymbol{\epsilon}\in\{\pm 1\}^2}\Sigma[\epsilon_1 A,\epsilon_2 B,a,k;q]\ll q^{1/2}\big(k^2a-T,q_\square\big)^{1/2},$$

where $q=q_\square^2 q_1$ with $q_1\in\{1,p\}$, the sum on the left-hand side may be omitted for $s\geqslant 2$, and the second factor may be omitted if $s=2$ or (more generally) if $(k^2a-T,q)\mid p^2$.

(2) *If $A \not\equiv B \pmod p$, then, for every $p \mid a$ and every $k \in \mathbb{Z}$,*

$$\sum_{\epsilon\in\{\pm1\}} \Sigma[\epsilon A, \epsilon B, a, k; q] \ll q^{1/2}.$$

12. Proof of Lemma 19

In this section, we estimate $\Sigma[A, B, a, k; q]$ for $q = p^s$ with $s \geqslant 2$, $p \mid a$, and $A = B$, and prove Lemma 19. We start by noting that, in the case $s = 1$,

$$\Sigma[A, A, a, k; p^s] = \sideset{}{^*}\sum_{\substack{m \bmod p\\ m\in u(\mathbb{Z}/p\mathbb{Z})^{\times 2}}} e\left(-\frac{km}{p}\right),$$

which can be estimated (and anyway formally falls under the same condition $\nu + \alpha \geqslant s$) as in (12.12) below. Therefore, in what follows we may and do assume that $s \geqslant 2$.

By the assumption $p \mid a$, we can write

$$a = p^\alpha a_0, \quad \alpha \geqslant 1, \quad p \nmid a_0.$$

In this case, the summation in (11.4) is over $m \in u(\mathbb{Z}/p\mathbb{Z})^{\times 2}$, so that we may write $m = \bar{u}x^2$ for some $x \in (\mathbb{Z}/p\mathbb{Z})^\times$. The phase $f[m, A, A, a, k]$ defined in (11.5) can be rewritten as

$$f[\bar{u}x^2, A, A, a, k] = 2A(x^2 + au)_{1/2} - 2A(x^2)_{1/2} - k\bar{u}x^2 = 2A\epsilon_x x\big((1 + p^\alpha a_0 u\bar{x}^2)^{1/2} - 1\big) - k\bar{u}x^2,$$

where, for $p \nmid x$, $\epsilon_x := (x^2)_{1/2}\bar{x}$ depends on $x \bmod p$ only.

As $x \in (\mathbb{Z}/p\mathbb{Z})^{\times 2}$, we see that $m = \bar{u}x^2$ runs over all admissible values of m twice. Thus,

$$\begin{aligned}\sum_{\epsilon\in\{\pm1\}} \Sigma[\epsilon A, \epsilon A, a, k; p^s] &= \frac{1}{2}\sum_{\epsilon\in\{\pm1\}} \sideset{}{^*}\sum_{x \bmod p^s} e\left(\frac{2A\epsilon\epsilon_x x\big((1 + p^\alpha a_0 u\bar{x}^2)^{1/2} - 1\big) - k\bar{u}x^2}{p^s}\right)\\ &= \sum_{\epsilon\in\{\pm1\}} \tilde{\Sigma}[\epsilon A, a, k; p^s],\end{aligned}$$

where

$$\tilde{\Sigma}[A, a, k; p^s] = \frac{1}{2}\sideset{}{^*}\sum_{x \bmod p^s} e\left(\frac{\tilde{f}(A, a, k; x)}{p^s}\right), \tag{12.1}$$

$$\tilde{f}(A, a, k; x) = 2Ax\big((1 + p^\alpha a_0 u\bar{x}^2)^{1/2} - 1\big) - k\bar{u}x^2. \tag{12.2}$$

We proceed to estimate the sum $\tilde{\Sigma}[A, a, k; p^s]$ defined as in (12.1) for an arbitrary $A \in \mathbb{Z}$, and we define $\nu = \min(\operatorname{ord}_p A, s)$. For every $\kappa \geqslant 1$, we find that

$$\begin{aligned}&\overline{x + p^\kappa t} = \bar{x} - \bar{x}^2 \cdot p^\kappa t + \bar{x}^3 \cdot p^{2\kappa}t^2 + \mathbf{M}_{p^{3\kappa}},\\ &\overline{x + p^\kappa t}^2 = \bar{x}^2 - 2\bar{x}^3 \cdot p^\kappa t + 3\bar{x}^4 \cdot p^{2\kappa}t^2 + \mathbf{M}_{p^{3\kappa}},\\ &\big(1 + p^\alpha a_0 u \cdot \overline{x + p^\kappa t}^2\big)^{1/2} = \Big(\big(1 + p^\alpha a_0 u\bar{x}^2\big) - 2a_0 u\bar{x}^3 \cdot p^{\kappa+\alpha}t + 3a_0 u\bar{x}^4 \cdot p^{2\kappa+\alpha}t^2 + \mathbf{M}_{p^{3\kappa+\alpha}}\Big)^{1/2}\\ &\qquad = \big(1 + p^\alpha a_0 u\bar{x}^2\big)^{1/2} - \overline{1 + p^\alpha a_0 u\bar{x}^2}^{1/2} \cdot a_0 u\bar{x}^3 \cdot p^{\kappa+\alpha}t + 3 \cdot \bar{2} \cdot a_0 u\bar{x}^4 \cdot p^{2\kappa+\alpha}t^2 + \mathbf{M}_{p^{2\kappa+\alpha+1}},\end{aligned}$$

and so, finally,

$$\begin{aligned}
&\big(x+p^\kappa t\big)\Big(\big(1+p^\alpha a_0 u\cdot\overline{x+p^\kappa t}^2\big)^{1/2}-1\Big)\\
&\quad= x\Big(\big(1+p^\alpha a_0u\bar{x}^2\big)^{1/2}-1\Big)-\overline{1+p^\alpha a_0u\bar{x}^2}^{1/2}\cdot a_0u\bar{x}^2\cdot p^{\kappa+\alpha}t+3\cdot\bar{2}\cdot a_0u\bar{x}^3\cdot p^{2\kappa+\alpha}t^2\\
&\qquad+\Big(\big(1+p^\alpha a_0u\bar{x}^2\big)^{1/2}-1\Big)\cdot p^\kappa t-a_0u\bar{x}^3\cdot p^{2\kappa+\alpha}t^2+\mathbf{M}_{p^{2\kappa+\alpha+1}}\\
&\quad= x\Big(\big(1+p^\alpha a_0u\bar{x}^2\big)^{1/2}-1\Big)+\Big(\overline{1+p^\alpha a_0u\bar{x}^2}^{1/2}-1\Big)\cdot p^\kappa t+\bar{2}\cdot a_0u\bar{x}^3\cdot p^{2\kappa+\alpha}t^2+\mathbf{M}_{p^{2\kappa+\alpha+1}}.
\end{aligned}\tag{12.3}$$

Using (12.2) and (12.3), we have that, for every $\kappa\geqslant 1$,

$$\begin{aligned}
\tilde{f}\big(A,a,k;x+p^\kappa t\big)=\tilde{f}(A,a,k;x)+2\Big[A\Big(\overline{1+p^\alpha a_0u\bar{x}^2}^{1/2}-1\Big)-k\bar{u}x\Big]\cdot p^\kappa t\\
+\big(Aa_0u\bar{x}^3p^\alpha-k\bar{u}\big)\cdot p^{2\kappa}t^2+\mathbf{M}_{p^{2\kappa+\nu+\alpha+1}}.
\end{aligned}$$

At this point, note that

$$\operatorname{ord}_p\Big[A\Big(\overline{1+p^\alpha a_0u\bar{x}^2}^{1/2}-1\Big)\Big]=\operatorname{ord}_p\big(Aa_0u\bar{x}^3p^\alpha\big)=\nu+\alpha.$$

We first consider the principal case when

$$\nu+\alpha\leqslant s-1.\tag{12.4}$$

Let $\omega=\operatorname{ord}_p k$, and define $\kappa_\star$ and j by

$$s=\min\big(\nu+\alpha,\omega\big)+2\kappa_\star+j,\quad \kappa_\star\geqslant 0,\ j\in\{0,1\}.\tag{12.5}$$

Then, for $\upsilon\in\{0,1\}$ (and $\upsilon=1$ if $\kappa_\star=0$), we have that

$$\begin{aligned}
\tilde{\Sigma}[A,a,k;p^s]&=\frac{1}{2}\frac{1}{p^{s-\kappa_\star-\upsilon}}\sideset{}{^*}\sum_{x\bmod p^s}\sum_{t\bmod p^{s-\kappa_\star-\upsilon}}e\left(\frac{\tilde{f}\big(A,a,k;x+p^{\kappa_\star+\upsilon}t\big)}{p^s}\right)\\
&=\frac{1}{2}p^{-s+\kappa_\star+\upsilon}\sideset{}{^*}\sum_{x\bmod p^s}e\left(\frac{\tilde{f}(A,a,k;x)}{p^s}\right)\times\\
&\qquad\times\sum_{t\bmod p^{s-\kappa_\star-\upsilon}}e\left(\frac{A\big(\overline{1+p^\alpha a_0u\bar{x}^2}^{1/2}-1\big)-k\bar{u}x}{p^{s-\kappa_\star-\upsilon}}t+\frac{Aa_0u\bar{x}^3p^\alpha-k\bar{u}}{p^{s-2\kappa_\star-2\upsilon}}t^2\right).
\end{aligned}\tag{12.6}$$

We first use this formula with $\upsilon=j$. With this choice, there is no quadratic term in the inner sum, and in fact it vanishes unless

$$\operatorname{ord}_p\Big[A\Big(\overline{1+p^\alpha a_0u\bar{x}^2}^{1/2}-1\Big)-k\bar{u}x\Big]\geqslant s-\kappa_\star-j=\min\big(\nu+\alpha,\omega\big)+\kappa_\star,\tag{12.7}$$

when it equals $p^{s-\kappa_\star-j}$. We see that we cannot have $\omega<\nu+\alpha$, for then (12.4) and (12.5) would imply $\kappa_\star\geqslant 1$, contradicting (12.7). Hence from now on we assume

$$\omega\geqslant\nu+\alpha.\tag{12.8}$$

We now distinguish two subcases, namely $\nu+\alpha\leqslant s-2$ and $\nu+\alpha=s-1$. In the former case, we have that $\omega=\nu+\alpha$, for if $\omega>\nu+\alpha$, then (12.7) implies $\kappa=0$, contradicting (12.5). Now, (12.7) implies that

$$A(-\bar{2}p^\alpha a_0u\bar{x}^2)-k\bar{u}x\equiv 0\pmod{p^{\omega+\kappa_\star}},$$

and so

$$Aa_0u\bar{x}^3p^\alpha-k\bar{u}\equiv -3k\bar{u}\pmod{p^{\omega+\kappa_\star}}.$$

In particular, the left-hand side has order p^{ω} (since $p > 3$). Using (12.6) with $\upsilon = 0$, we are left with a constant sum if $j = 0$ and a nondegenerate quadratic Gauß sum modulo p if $j = 1$; in either case, it follows that

$$\tilde{\Sigma}[A, a, k; p^s] \ll p^{-\frac{j}{2}} \cdot \# \Big\{ x \bmod p^s : A\Big(\overline{1 + p^{\alpha} a_0 u \bar{x}^2}^{1/2} - 1\Big) \equiv k\bar{u}x \bmod p^{\nu+\alpha+\kappa_\star} \Big\} \tag{12.9}$$

if $\operatorname{ord}_p k = \nu + \alpha$, and $\tilde{\Sigma}[A, a, k; p^s] = 0$ otherwise.

We bound the number of solutions of the congruence modulo $p^{\nu+\alpha+\kappa_\star}$ in (12.9) using Lemma 17. Write

$$A = p^{\nu} A_0, \quad k = p^{\nu+\alpha} k_0,$$

with $(A_0, p) = (k_0, p) = 1$. In light of $\overline{1 + p^{\alpha} a_0 u \bar{x}^2}^{1/2} - 1 = -\bar{2} p^{\alpha} a_0 u \bar{x}^2 + \mathbf{M}_{p^{2\alpha}}$, we have that

$$f(x) := p^{-\nu-\alpha} \Big[A\Big(\overline{1 + p^{\alpha} a_0 u \bar{x}^2}^{1/2} - 1\Big) - k\bar{u}x \Big] = A_0 \frac{\overline{1 + p^{\alpha} a_0 u \bar{x}^2}^{1/2} - 1}{p^{\alpha}} - k_0 \bar{u} x$$

is a map $(\mathbb{Z}/p^s\mathbb{Z})^{\times} \to \mathbb{Z}/p^s\mathbb{Z}$. The congruence $f(x) \equiv 0 \pmod p$ implies that

$$\begin{aligned} k_0 \bar{u} x &\equiv -\bar{2} a_0 u A_0 \bar{x}^2 \pmod p, \\ x^3 &\equiv -\bar{2} a_0 u^2 \bar{k}_0 A_0 \pmod p \end{aligned}$$

and hence has $\mathrm{O}(1)$ solutions. Moreover, since, for every $\kappa \geqslant 1$,

$$\begin{aligned} \overline{1 + p^{\alpha} a_0 u \cdot \overline{x + p^{\kappa} t}^2}^{1/2} &= \overline{\big(1 + p^{\alpha} a_0 u \bar{x}^2\big) + p^{\alpha} a_0 u \big(- 2\bar{x}^3 p^{\kappa} t + \mathbf{M}_{p^{2\kappa}}\big)}^{1/2} \\ &= \overline{1 + p^{\alpha} a_0 u \bar{x}^2}^{1/2} + \big(\overline{1 + p^{\alpha} a_0 u \bar{x}^2}^{1/2}\big)^3 p^{\alpha} a_0 u \bar{x}^3 \cdot p^{\kappa} t + \mathbf{M}_{p^{\alpha+2\kappa}}, \end{aligned}$$

we have that

$$f(x + p^{\kappa} t) - f(x) - p^{\kappa} f_1(x) t \in p^{\kappa+1}\mathbb{Z}/p^s\mathbb{Z}$$

for every $\kappa \geqslant 1$, with

$$f_1(x) = A_0 a_0 u \bar{x}^3 - k_0 \bar{u} \equiv -3 k_0 \bar{u} \not\equiv 0 \pmod p$$

for every x such that $f(x) \equiv 0 \pmod p$.

By Lemma 17, we conclude that the congruence $f(x) \equiv 0 \pmod{p^{\kappa_\star}}$ has $\mathrm{O}(1)$ solutions x modulo $p^{\kappa_\star}$ and hence

$$\mathrm{O}(p^{s-\kappa_\star}) = \mathrm{O}\big(p^{(s+j)/2+(\nu+\alpha)/2}\big)$$

solutions in x modulo p^s with the notation as in (12.5). Substituting this bound into (12.9), we conclude for $\nu + \alpha \leqslant s - 2$ that

$$\tilde{\Sigma}[A, a, k; p^s] \ll \begin{cases} p^{\frac{1}{2}s + \frac{1}{2}(\nu+\alpha)}, & \operatorname{ord}_p k = \nu + \alpha, \\ 0, & \text{else.} \end{cases} \tag{12.10}$$

Our second subcase is $\nu + \alpha = s - 1$. Here, we find that $\tilde{f}(A, a, k; x)$ is an even function of x such that $\tilde{f}(A, a, k; x) = p^{s-1} \tilde{f}_1(A, a, k; x)$ with

$$\tilde{f}_1(A, a, k; x) \equiv A_0 a_0 u \bar{x} - k_1 \bar{u} x^2 \pmod p,$$

where $k = p^{s-1} k_1$, $\operatorname{ord}_p k_1 \geqslant 0$. Therefore, by (12.1)–(12.2),

$$\tilde{\Sigma}[A, a, k; p^s] = p^{s-1} \sideset{}{^*}\sum_{x \bmod p} e\left(\frac{\epsilon a_0 u \bar{x} - k_1 \bar{u} x^2}{p}\right).$$

The resulting sum can be estimated by Lemma 16 as $\ll p^{1/2}$ if $\operatorname{ord}_p k_1 = 0$ and becomes the Ramanujan sum (and is hence $\ll 1$) if $\operatorname{ord}_p k_1 \geqslant 1$. Hence, for $\nu + \alpha = s - 1$,

$$\tilde{\Sigma}[A, a, k; p^s] \ll \begin{cases} p^{s-1}, & p^s \mid k, \\ p^{\frac{1}{2}s + \frac{1}{2}(\nu+\alpha)}, & \operatorname{ord}_p k = \nu + \alpha, \\ 0, & \text{else.} \end{cases} \tag{12.11}$$

This completes the analysis of the case $\nu + \alpha \leqslant s - 1$.

In the complementary case when $\nu + \alpha \geqslant s$, we are dealing with a quadratic Gauß sum:

$$\tilde{\Sigma}[A, a, k; p^s] = \frac{1}{2} \sideset{}{^*}\sum_{x \bmod p^s} e\left(-\frac{k\bar{u}x^2}{p^s}\right),$$

which vanishes unless $k = p^{s-1}k_1$ for some $k_1 \in \mathbb{Z}/p\mathbb{Z}$, in which case it is $\ll p^{s-1/2}$ if $\operatorname{ord}_p k_1 = 0$ and $\ll p^s$ if $p \mid k$. Therefore, for $\nu + \alpha \geqslant s$,

$$\tilde{\Sigma}[A, a, k; p^s] \ll \begin{cases} p^s, & p^s \mid k, \\ p^{s-\frac{1}{2}}, & \operatorname{ord}_p k = s - 1, \\ 0, & \text{else.} \end{cases} \tag{12.12}$$

Combining our findings (12.10), (12.11), (12.12) completes the proof of Lemma 19. □

13. Proof of Lemma 20

In this section, we estimate the sum $\Sigma[A, B, a, k; q]$ defined in (11.4) for $q = p^s$ with $s \geqslant 1$, $p \nmid A$ or $p \nmid B$, and either of the following two conditions holds:

(1) $p \nmid a$, or
(2) $p \mid a$ and $A \not\equiv B \pmod p$.

As the final result of this section, we obtain a proof of Lemma 20.

We remark that the argument in the previous section relied heavily on the fact that $p \mid a$ and $A = B$, which results in a specific alignment of the branches of the square-root. This section's argument, which addresses all remaining cases, is different (and harder), in particular due to the possible presence of singular critical points in the stationary phase analysis. Recall the notation (10.9).

13.1. **Preliminaries.** We start with some useful differencing formulas. Recall our assumption that $p \neq 2$. Note that, for every $\kappa \geqslant 1$ and every $t \in \mathbb{Z}_p$,

$$\begin{aligned} \big(m + p^\kappa t\big)_{1/2} &= m_{1/2} + \bar{2} \cdot \overline{m_{1/2}} p^\kappa t - \bar{8} \cdot \overline{m_{1/2}}^3 \cdot p^{2\kappa} t^2 + \mathbf{M}_{p^{3\kappa}}, \\ \overline{(m + p^\kappa t)_{1/2}} &= \overline{m_{1/2}} - \bar{2} \cdot \overline{m_{1/2}}^3 \cdot p^\kappa t + 3 \cdot \bar{8} \cdot \overline{m_{1/2}}^5 \cdot p^{2\kappa} t^2 + \mathbf{M}_{p^{3\kappa}}, \\ \overline{(m + p^\kappa t)_{1/2}}^3 &= \overline{m_{1/2}}^3 - 3 \cdot \bar{2} \cdot \overline{m_{1/2}}^5 \cdot p^\kappa t + \mathbf{M}_{p^{2\kappa}}. \end{aligned} \tag{13.1}$$

Denote

$$\begin{aligned} g(m, A, B, a) &= Au\overline{\big((m+a)u\big)_{1/2}} - Bu\overline{(mu)_{1/2}}, \\ g_1(m, A, B, a) &= -\bar{2}Au^2\overline{\big((m+a)u\big)_{1/2}}^3 + \bar{2}Bu^2\overline{(mu)_{1/2}}^3, \\ g_2(m, A, B, a) &= 3 \cdot \bar{4} \cdot Au^3\overline{\big((m+a)u\big)_{1/2}}^5 - 3 \cdot \bar{4} \cdot Bu^3\overline{(mu)_{1/2}}^5. \end{aligned}$$

Using (13.1), we thus obtain the following differencing expansions:

$$\begin{aligned} f[m+p^\kappa tA,B,a,k] &= f[m,A,B,a,k] + \big(g(m,A,B,a)-k\big)\cdot p^\kappa t \\ &\qquad + \bar{2}\cdot g_1(m,A,B,a)\cdot p^{2\kappa}t^2 + \mathbf{M}_{p^{3\kappa}}, \\ g(m+p^\kappa t,A,B,a) &= g(m,A,B,a) + g_1(m,A,B,a)\cdot p^\kappa t \\ &\qquad + \bar{2}\cdot g_2(m,A,B,a)\cdot p^{2\kappa}t^2 + \mathbf{M}_{p^{3\kappa}}, \\ g_1(m+p^\kappa t,A,B,a) &= g_1(m,A,B,a) + g_2(m,A,B,a)\cdot p^\kappa t + \mathbf{M}_{p^{2\kappa}}. \end{aligned} \tag{13.2}$$

13.2. **The prime case.** In this subsection, we address the case $s=1$ and prove an estimate for

$$\begin{aligned} \hat{\Sigma} &= \sum_{\boldsymbol{\epsilon}\in\{\pm1\}^2} \Sigma[\epsilon_1 A,\epsilon_2 B,a,k;p] \\ &= \sum_{\boldsymbol{\epsilon}\in\{\pm1\}^2} \sideset{}{^*}\sum_{\substack{m \bmod p \\ m,m+a\in u(\mathbb{Z}/p\mathbb{Z})^{\times 2}}} e\left(\frac{2\epsilon_1 A\big((m+a)u\big)_{1/2} + 2\epsilon_2 B(mu)_{1/2} - km}{p}\right). \end{aligned} \tag{13.3}$$

We first consider the case (1), when $p\nmid a$. Denoting $x=\epsilon_1\big((m+a)u\big)_{1/2}$ and $y=\epsilon_2(mu)_{1/2}$, we have that

$$(x+y)(x-y) = x^2-y^2 = au,$$

so that $v=x+y\in(\mathbb{Z}/p\mathbb{Z})^\times$ and $x-y=au\bar{v}$, as well as $v^2,-(au\bar{v})^2\not\equiv au \pmod p$, that is, $v^2\not\equiv\pm au \pmod p$. Conversely, if $v\in(\mathbb{Z}/p\mathbb{Z})^\times$ is arbitrary such that $v^2\not\equiv\pm au \pmod p$, and if we choose

$$x=\bar{2}(v+au\bar{v}) \quad\text{and}\quad y=\bar{2}(v-au\bar{v})$$

so that $x+y=v$ and $x-y=au\bar{v}$, then $x^2-y^2=au$ and so $x^2=(m+a)u$ and $y^2=mu$ for some $m\in(\mathbb{Z}/p\mathbb{Z})^\times$. In this case, $m,m+a\in u(\mathbb{Z}/p\mathbb{Z})^{\times 2}$ is automatic, and

$$m\equiv\bar{u}y^2\equiv\bar{4}\bar{u}(v-au\bar{v})^2 \pmod p.$$

This discussion shows that

$$\hat{\Sigma} = \sideset{}{^*}\sum_{\substack{v \bmod p \\ v^2\not\equiv\pm au \bmod p}} e\left(\frac{R(v)}{p}\right), \tag{13.4}$$

where $R(v)$ is a rational function given by

$$R(v) = A(v+au\bar{v}) + B(v-au\bar{v}) - \bar{4}k\bar{u}(v-au\bar{v})^2 \tag{13.5}$$

(which can never be constant modulo p). By Lemma 16 we conclude

$$\hat{\Sigma}\ll p^{1/2}. \tag{13.6}$$

In the easier case (2), the inner sum in (13.3) in over $m\in u(\mathbb{Z}/p\mathbb{Z})^{\times 2}$, and, by writing $m=\bar{u}x^2$, we have that

$$\sum_{\epsilon\in\{\pm1\}}\Sigma[\epsilon A,\epsilon B,a,k;p] = \sideset{}{^*}\sum_{x \bmod p} e\left(\frac{2(A-B)x-k\bar{u}x^2}{p}\right) \ll p^{1/2},$$

by the evaluation of the Gauss sum (10.1), or by an application of Lemma 16.

13.3. **Lemmata on Hensel liftings.** Estimating the sum Σ using the method of stationary phase involves solving congruences of the form

$$g(m, A, B, a) \equiv k \pmod{p^\kappa}.$$

The following lemma is concerned with the base case $\kappa = 1$.

Lemma 21. *Let $p \neq 2$, $p \nmid A$ or $p \nmid B$, and either $p \nmid a$, or $p \mid a$ and $A \not\equiv B \pmod p$. Then, the congruence*

$$g(m, A, B, a) \equiv k \pmod p. \tag{13.7}$$

has $\mathrm{O}(1)$ solutions in m modulo p.

Proof. We may rewrite (13.7) as

$$A\overline{\big((m+a)u\big)}_{1/2} \equiv B\overline{(mu)}_{1/2} + k\bar{u} \pmod p, \tag{13.8}$$

If $p \nmid a$, we obtain by repeated squaring from (13.8) that

$$A^2\overline{m+a} \equiv B^2\overline{m} + 2Bk\overline{(mu)}_{1/2} + k^2\bar{u} \pmod p,$$

$$\big(A^2\overline{m+a} - B^2\overline{m} - k^2\bar{u}\big)^2 \equiv 4B^2\overline{mu}k^2 \pmod p.$$

Expanding and multiplying by $m^2(m+a)^2$, we obtain the congruence

$$k^4\bar{u}^2m^2(m+a)^2 - 4B^2k^2\bar{u}m(m+a)^2$$
$$+ 2B^2k^2\bar{u}(m+a)^2 - 2A^2k^2\bar{u}m^2 - 2A^2B^2m(m+a) + B^4(m+a)^2 + A^4m^2 \equiv 0 \pmod p.$$

We immediately see that, if $k \not\equiv 0 \pmod p$, we have a quartic equation, and so it can have at most four solutions mod p.

We next consider the case when $k \equiv 0 \pmod p$. In this case, after squaring the condition (13.7), we have that

$$A^2m \equiv B^2(m+a) \pmod p.$$

This congruence has precisely one solution (for given A, B) in the case when $A^2 - B^2 \not\equiv 0 \pmod p$ and no solutions when $A^2 \equiv B^2 \not\equiv 0 \pmod p$; in all these cases, (13.7) has $\mathrm{O}(1)$ solutions; this completes the proof of our lemma in the case $p \nmid a$.

If $p \mid a$ and $A \not\equiv B \pmod p$, then (13.8) is equivalent to the congruence

$$(A-B)\overline{(mu)}_{1/2} \equiv k\bar{u} \pmod p,$$

which has at most one solution, given by

$$m \equiv (A-B)^2\bar{k}^2u \pmod p.$$

(In particular, there are no solutions if $p \mid k$.) This proves our lemma in the case $p \mid a$, $A \not\equiv B \pmod p$.

Although we will not need this, we remark that, for $p^\alpha \parallel a$, it follows by exactly the same argument that the congruence $g(m, A, B, a) \equiv k \pmod{p^\alpha}$ has at most one solution modulo p^α, given explicitly by $m \equiv (A-B)^2\bar{k}^2u \pmod{p^\alpha}$ (this being a solution of exactly one of the two congruences corresponding to the two pairs $(\epsilon A, \epsilon B)$ entering the statement of Lemma 20). □

An immediate consequence of Lemma 21 and Lemma 17 is the following statement. Note that, when $p \mid a$ and $A \not\equiv B \pmod p$, the congruence $g_1(m, A, B, a) \equiv 0 \pmod p$ has no solutions.

Lemma 22. *Let $p \neq 2$, $p \nmid A$ or $p \nmid B$, and $\kappa \geqslant 1$.*

(1) *If $p \nmid a$, then the congruence*

$$g(m, A, B, a) \equiv k \pmod{p^\kappa} \tag{13.9}$$

has $\mathrm{O}(1)$ solutions in m modulo p^κ such that

$$g_1(m, A, B, a) \not\equiv 0 \pmod p.$$

(2) *If $p \mid a$ and $A \not\equiv B \pmod p$, then (13.9) has $\mathrm{O}(1)$ solutions in m modulo p^κ.*

The remainder of this subsection is concerned with the singular solutions to $g(m, A, B, a) \equiv k \pmod{p^\kappa}$ in the case $p \nmid a$, that is, those solutions for which $g_1(m, A, B, a) \equiv 0 \pmod{p}$. The following lemma, which will ensure non-singularity of certain congruences, is an elementary exercise.

Lemma 23. *Let $p \neq 2$, $p \nmid a$, and $p \nmid A$ or $p \nmid B$. Then, the system of congruences*

$$g(m, A, B, a) \equiv g_1(m, A, B, a) \equiv 0 \pmod{p}$$

has no solutions in m. If additionally $p \neq 3$, then the system of congruences

$$g_1(m, A, B, a) \equiv g_2(m, A, B, a) \equiv 0 \pmod{p}$$

has no solutions in m.

Proof. We consider the first statement; the second is entirely analogous. Assume that

$$A\overline{\big((m+a)u\big)_{1/2}} \equiv B\overline{(mu)_{1/2}} \pmod{p}, \qquad A\overline{\big((m+a)u\big)_{1/2}}^3 \equiv B\overline{(mu)_{1/2}}^3 \pmod{p}.$$

Then $(m+a)u \equiv mu \pmod{p}$, contradicting $p \nmid (au)$. □

We can use the previous simple observation in the proof of the following.

Lemma 24. *Let $p \notin \{2, 3\}$, $p \nmid a$, and $p \nmid A$ or $p \nmid B$. Then the congruence*

$$g_1(m, A, B, a) \equiv 0 \pmod{p}$$

has $\mathrm{O}(1)$ solutions $m_1^\flat, \dots, m_\omega^\flat$. Furthermore, for every $\kappa \geqslant 1$, the congruence

$$g_1(m, A, B, a) \equiv 0 \pmod{p^\kappa}$$

has exactly ω solutions modulo p^κ. In fact, these solutions may be written as $m_1^{[\kappa]}, \dots, m_\omega^{[\kappa]}$ with $m_i^{[\kappa]} \equiv m_i^\flat$ for every $1 \leqslant i \leqslant \omega$.

Proof. We start with the congruence $g_1(m, A, B, a) \equiv 0 \pmod{p}$, which we rewrite as

$$A\overline{\big((m+a)u\big)_{1/2}}^3 \equiv B\overline{(mu)_{1/2}}^3.$$

Squaring both sides and rearranging, it follows that

$$\bar{A}^2(m+a)^3 - \bar{B}^2 m^3 \equiv 0 \pmod{p}.$$

This is at most a cubic congruence in m modulo p, and certainly its leading and constant coefficients cannot both vanish. Therefore it has $\mathrm{O}(1)$ solutions modulo p, say, $m_1^\flat, \dots, m_\omega^\flat$. According to Lemma 23, each of these solutions satisfies $g_2(m, A, B, a) \equiv 0 \pmod{p}$. Thus the remaining claims follow, in light of (13.2), from Lemma 17. □

Applying Lemma 24 with $\kappa = s$, we obtain ω solutions

$$m_1, \dots, m_\omega$$

satisfying $g_1(m, A, B, a) \equiv 0 \pmod{p^s}$; we denote

$$k_i = g(m_i, A, B, a). \tag{13.10}$$

We stress again that, according to Lemma 23, all of these solutions satisfy

$$k_i \not\equiv 0 \pmod{p} \tag{13.11}$$

as well as

$$g_2(m_i, A, B, a) \not\equiv 0 \pmod{p}.$$

We are now ready for the following lemma, which is of key importance in solving our stationary phase problem.

Lemma 25. *Let $p \notin \{2,3\}$, $p \nmid a$, and $p \nmid A$ or $p \nmid B$. Also, let $k \in \mathbb{Z}$ and $1 \leqslant \kappa \leqslant s$. Write $\kappa = 2\kappa_\star + j$ with $j \in \{0,1\}$. The congruence*

$$g(m,A,B,a) \equiv k \pmod{p^\kappa}$$

can have solutions such that

$$g_1(m,A,B,a) \equiv 0 \pmod{p}$$

only if

$$I(k) = \left\{1 \leqslant i \leqslant \omega : k \equiv k_i \bmod p^{\min(\kappa,2)}\right\} \neq \emptyset.$$

For each $i \in I(k)$, let

$$s_i = \begin{cases} p^{\kappa_\star}, & p^\kappa \mid (k-k_i), \\ p^\mu, & p^{2\mu} \parallel (k-k_i) \text{ for some } 1 \leqslant \mu \leqslant \kappa_\star, \\ 0, & \text{else.} \end{cases}$$

Then the congruence $g(m,A,B,a) \equiv a \pmod{p^\kappa}$ has at most

$$\ll \sum_{i \in I(k)} s_i$$

solutions modulo p^κ such that $g_1(m,A,B,a) \equiv 0 \pmod{p}$. In particular, denoting

$$\rho(k) = \max_{1 \leqslant i \leqslant \omega} \operatorname{ord}_p(k-k_i),$$

this number of solutions is

$$\mathrm{O}\left(p^{\lfloor \frac{1}{2}\min(\rho(k),\kappa)\rfloor}\right).$$

Proof. According to Lemma 24, every m such that $g_1(m,A,B,a) \equiv 0 \pmod{p}$ satisfies $m \equiv m_i \pmod{p}$ for exactly one $1 \leqslant i \leqslant \omega$. If $m \equiv m_i \pmod{p^s}$, then according to (13.2) we have that

$$g(m,A,B,a) \equiv k_i \pmod{p^s}, \quad g_1(m,A,B,a) \equiv 0 \pmod{p^s}.$$

Otherwise, write $m = m_i + p^\mu t$ for some $1 \leqslant \mu < s$ and $p \nmid t$. Using (13.2), we find that

$$p^{2\mu} \parallel \big(g(m,A,B,a) - k_i\big), \quad p^\mu \parallel g_1(m,A,B,a).$$

In either case, we see that $p^2 \mid \big(g(m,A,B,a) - k_i\big)$.

If $\kappa = 1$, this shows that solutions of $g(m,A,B,a) \equiv k \pmod{p}$ such that $g_1(m,A,B,a) \equiv 0 \pmod{p}$ exist only if $k \equiv k_i \pmod{p}$ for some $1 \leqslant i \leqslant \omega$ and that each such solution m must satisfy $m \equiv m_i \pmod{p}$ for some $i \in I(k)$; in particular, the number of solutions is $\mathrm{O}(1)$. This completes the proof in the case $\kappa = 1$.

If $2 \leqslant \kappa \leqslant s$, then $k \equiv k_i \pmod{p^2}$ and so $I(k) \neq \emptyset$. We distinguish two cases: $k \equiv k_i \pmod{p^\kappa}$ and $k \not\equiv k_i \pmod{p^\kappa}$.

In the first case, write $\kappa = 2\kappa_\star + j$, $\kappa_\star \geqslant 1$, $j \in \{0,1\}$. We have that $p^\kappa \mid (k-k_i)$, and the congruence to be solved is equivalent to

$$g(m,A,B,a) \equiv k_i \pmod{p^\kappa}, \quad m \equiv m_i \pmod{p}.$$

One solution of this congruence is $m \equiv m_i \pmod{p^\kappa}$. Otherwise, and writing $m = m_i + p^\mu t$ for some $1 \leqslant \mu < s$ and $p \nmid t$, we cannot have $2\mu < \kappa$, that is, we must have $2\mu \geqslant 2\kappa_\star + j$ and hence $\mu \geqslant \kappa_\star + j$. Keeping in mind that we must have $m \equiv m_i \pmod{p^\mu}$, we obtain at most

$$\mathrm{O}(p^{\kappa-\mu}) = \mathrm{O}(p^{\kappa_\star})$$

solutions for m modulo p^κ.

In the second case, let $p^\lambda \parallel (k-k_i)$ for some $2 \leqslant \lambda < \kappa$. In that case, $p^\lambda \parallel \big(g(m,A,B,a) - k_i\big)$, and so we must have $\lambda = 2\mu$ for some $1 \leqslant \mu < \lambda < \kappa$. Therefore, $p^\mu \parallel (m - m_i)$, and $p^\mu \parallel g_1(m,A,B,a)$.

Fix one such solution m_0. We now count the number of solutions of

$$g(m,A,B,a) \equiv k \pmod{p^{\mu+\varsigma}}, \quad m \equiv m_0 \pmod{p^\mu} \tag{13.12}$$

modulo p^ς for every $\mu \leqslant \varsigma \leqslant \kappa$. Note that $\kappa - \mu \geqslant \mu + 1$. For $\varsigma = \mu$, we obviously have exactly one such solution.

We use the second relationship from (13.2):

$$g(m + p^\iota t, A, B, a) = g(m, A, B, a) + g_1(m, A, B, a) \cdot p^\iota t + \bar{2} \cdot g_2(m, A, B, a) \cdot p^{2\iota} t^2 + \mathbf{M}_{p^{3\iota}}. \tag{13.13}$$

Using (13.13) with $\iota = \mu$, we see that the congruence

$$g(m_0 + p^\mu t, A, B, a) \equiv k \pmod{p^{2\mu+1}}$$

is equivalent to

$$\bar{2} \cdot g_2(m_0, A, B, a) \cdot t^2 + \frac{g_1(m_0, A, B, a)}{p^\mu} t + \frac{g(m_0, A, B, a) - k}{p^{2\mu}} \equiv 0 \pmod{p}.$$

This is a nontrivial quadratic congruence in t, and so it has $\mathrm{O}(1)$ solutions in t modulo p. Corresponding to this are $\mathrm{O}(1)$ solutions m modulo $p^{\mu+1}$ of (13.12) with $\varsigma = \mu + 1$.

We now prove that, given a $\varsigma \geqslant \mu + 1$ and a solution of m_1 of

$$g(m, A, B, a) \equiv k \pmod{p^{\mu+\varsigma}},$$

there exists a unique m_2 modulo $p^{\varsigma+1}$ such that

$$g(m, A, B, a) \equiv k \pmod{p^{\mu+\varsigma+1}}, \quad m_2 \equiv m_1 \pmod{p^\varsigma}.$$

Indeed, writing $m_2 = m_1 + p^\varsigma t$ and using (13.13) with $\iota = \varsigma$ (and noting that $2\varsigma \geqslant \mu + \varsigma + 1$), the congruence $g(m_1 + p^\varsigma t, A, B, a) \equiv k \pmod{p^{\mu+\varsigma}}$ is equivalent to

$$\frac{g_1(m, A, B, a)}{p^\mu} t + \frac{g(m, A, B, a) - k}{p^{\mu+\varsigma}} \equiv 0 \pmod{p}.$$

This is a nontrivial linear congruence in t, and so it has a unique solution in t modulo p. Corresponding to this is a unique solution m_2 modulo $p^{\varsigma+1}$ of $g(m, A, B, a) \equiv k \pmod{p^{\mu+\varsigma+1}}$ such that $m_2 \equiv m_1 \pmod{p^\varsigma}$.

Putting everything together, we have proved that, for every $\varsigma \geqslant \mu$, the system (13.12) has $\mathrm{O}(1)$ solutions modulo p^ς. In particular, there are $\mathrm{O}(1)$ solutions modulo $p^{\kappa-\mu}$ of

$$g(m, A, B, a) \equiv k \pmod{p^\kappa} \quad m \equiv m_0 \pmod{p^\mu}.$$

Adding over all $\mathrm{O}(1)$ values of m_0, we finally obtain

$$\mathrm{O}(p^\mu)$$

solutions of $g(m, A, B, a) \equiv k \pmod{p^\kappa}$ modulo p^κ. □

We see from Lemma 25 that the numbers k_i play a central role in counting the solutions to $g(m, A, B, a) \equiv k \pmod{p^\kappa}$ such that $g_1(m, A, B, a) \equiv 0 \pmod{p}$. In the following lemma, we make the dependence of these special values on the parameter a a bit more explicit.

Lemma 26. *Let $p \neq 2$, $p \nmid a$, and $p \nmid A$ or $p \nmid B$, and let $k_1, \dots, k_\omega$ be defined as in (13.10). There exists a finite set $T \subseteq \mathbb{Z} \setminus p\mathbb{Z}$ of absolutely bounded cardinality whose elements depend on p^s, A, and B only, such that for every $k \in \mathbb{Z}$ and every $1 \leqslant \lambda \leqslant k$, the congruence $k \equiv k_i \pmod{p^\lambda}$ for some $1 \leqslant i \leqslant \omega$ implies that*

$$k^2 a \equiv t \pmod{p^\lambda}$$

for some $t \in T$.

Proof. Let $\{v_1, v_2\} \in (\mathbb{Z}/p\mathbb{Z})^\times$ be fixed representatives of the two cosets of the subgroup $(\mathbb{Z}/p\mathbb{Z})^{\times 2}$, and let $\bar{v}_j v_j \equiv 1 \pmod{p^s}$. According to Lemma 24, each of the four congruences

$$g_1(m, A, \epsilon B, v_j) \equiv 0 \pmod{p^s},$$

where $\epsilon \in \{\pm 1\}$ and $j \in \{1, 2\}$, has $\mathrm{O}(1)$ solutions modulo p^s, which we denote as

$$m_1^{\epsilon, v_j}, \dots, m_{\omega(\epsilon, v_j)}^{\epsilon, v_j}.$$

Let

$$k_r^{\epsilon,v_j} = g\big(m_r^{\epsilon,v_j}, A, \epsilon B, v_j\big).$$

Recall that $k_r^{\epsilon,v_j} \not\equiv 0 \pmod p$ by (13.11). We claim that the set

$$T = \Big\{ \big(k_r^{\epsilon,v_j}\big)^2 \bar v_j : \epsilon \in \{\pm 1\},\ j \in \{1,2\},\ 1 \leqslant r \leqslant \omega(\epsilon, v_j) \Big\}$$

satisfies all our properties.

Clearly, it suffices to prove that, for every $p \nmid a$, and with $k_1, \dots, k_\omega$ defined as in (13.10), we have that, for every $1 \leqslant i \leqslant \omega$, there exists a $t \in T$ such that

$$k_i^2 a \equiv g(m_i, A, B, a)^2 a \equiv t \pmod{p^s}.$$

Indeed, let $v_j \in \{v_1, v_2\}$ be the chosen representative of the coset $a(\mathbb{Z}/p\mathbb{Z})^{\times 2}$. The values $m = m_i$ are solutions of the congruence

$$A\overline{\big((m+a)u\big)_{1/2}}^3 - B\overline{(mu)_{1/2}}^3 \equiv 0 \pmod{p^s}.$$

Write

$$m \equiv a\bar v_j x \pmod{p^s},$$

and let $\epsilon_1 = \epsilon_1(x, v_j, a)$ and $\epsilon_2 = \epsilon_2(x, v_j, a)$ be such that

$$\big((x+v_j)u(a\bar v_j)\big)_{1/2} \equiv \epsilon_1 \cdot \big((x+v_j)u\big)_{1/2}(a\bar v_j)_{1/2} \pmod{p^s},$$
$$\big((xu)(a\bar v_j)\big)_{1/2} = \epsilon_2 \cdot (xu)_{1/2}(a\bar v_j)_{1/2} \pmod{p^s}.$$

We stress that ϵ_1 and ϵ_2 may depend on x, and that the definition of the set T is such that this causes no problem. With this change of variables, the above congruence is equivalent to the following congruence in x such that $(x+v_j)u$, $xu \in (\mathbb{Z}/p\mathbb{Z})^{\times 2}$:

$$g_1(x, A, \epsilon_1\epsilon_2 B, v_j) = A\overline{\big((x+v_j)u\big)_{1/2}}^3 - \epsilon_1\epsilon_2 B\overline{(xu)_{1/2}}^3 \equiv 0 \pmod{p^s}.$$

According to Lemma 24, this means that

$$x \equiv m_r^{\epsilon_1\epsilon_2, v_j} \pmod{p^s}$$

for some $1 \leqslant r \leqslant \omega(\epsilon_1\epsilon_2, v_j)$. Consequently, we find that

$$\begin{aligned} k_i = g(m_i, A, B, a) &= A\overline{\big((m+a)u\big)_{1/2}} - B\overline{(mu)_{1/2}} \\ &\equiv \epsilon_1 A\overline{\big((x+v_j)u\big)_{1/2}} \cdot \overline{(a\bar v_j)_{1/2}} - \epsilon_2 B\overline{(xu)_{1/2}} \cdot \overline{(a\bar v_j)_{1/2}} \\ &\equiv \epsilon_1 g\big(m_r^{\epsilon_1\epsilon_2, v_j}, A, \epsilon_1\epsilon_2 B, v_j\big)\overline{(a\bar v_j)_{1/2}} \pmod{p^s}. \end{aligned}$$

This final congruence implies that

$$k_i^2 a \equiv \big(k^{\epsilon_1\epsilon_2, v_j}, r\big)^2 v_j \pmod{p^s},$$

and the right-hand side is an element of the set T by construction. □

As a consequence of Lemmas 25 and 26, we obtain the following compact statement.

Lemma 27. *Let $p > 3$, $p \nmid a$, and $p \nmid A$ or $p \nmid B$, and let $1 \leqslant \kappa \leqslant s$. There exists a finite set $T \subseteq \mathbb{Z} \setminus p\mathbb{Z}$ of absolutely bounded cardinality whose elements depend on p^s, A, and B only, such that the number of solutions of the congruence*

$$g(m, A, B, a) \equiv k \pmod{p^\kappa}$$

such that

$$g_1(m, A, B, a) \equiv 0 \pmod p$$

is at most

$$\mathrm{O}\left(p^{\lfloor \frac{1}{2}\min(\tilde\rho(k^2 a), \kappa)\rfloor}\right),$$

where

$$\tilde{\rho}(\ell) = \max_{t \in T} \operatorname{ord}_p(\ell - t).$$

13.4. **Stationary phase estimates.** In this subsection, we use the facts from subsection 13.3 about the number of solutions to the stationary phase congruence (13.15), below, to estimate the sum

$$\tilde{\Sigma} := \Sigma[A, B, a, k; p^s] = \sideset{}{^*}\sum_{\substack{m \bmod p^s \\ m, m+a \in n_1(\mathbb{Z}/p\mathbb{Z})^{\times 2}}} e\left(\frac{f[m, A, B, a, k]}{p^s}\right).$$

We recall our general assumptions

$$s \geqslant 2, \ p > 3, \ \text{and } p \nmid A \text{ or } p \nmid B,$$

as well as

$$\text{either} \quad p \nmid a \quad \text{or} \quad p \mid a, \ A \not\equiv B \pmod{p}.$$

Write $s = 2\kappa + j$, $\kappa \geqslant 1$, $j \in \{0, 1\}$. Applying the usual stationary phase argument and the first equality in (13.2), we find that

$$\begin{aligned} \tilde{\Sigma} = p^{\kappa} \sideset{}{^*}\sum_{\substack{m \bmod p^{\kappa},\ m, m+a \in n_1(\mathbb{Z}/p\mathbb{Z})^{\times 2} \\ g(m, A, B, a) \equiv k \bmod p^{\kappa}}} e\left(\frac{f[m, A, B, a, k]}{p^s}\right) \\ \times \sum_{t \bmod p^j} e\left(\frac{\left[\big(g(m, A, B, a) - k\big)/p^{\kappa}\right] \cdot t + \bar{2} g_1(m, A, B, a) \cdot t^2}{p^j}\right). \end{aligned} \tag{13.14}$$

The outer sum is indexed by solutions of the congruence

$$g(m, A, B, a) \equiv k \pmod{p^{\kappa}} \tag{13.15}$$

modulo p^{κ}. We write

$$\tilde{\Sigma} = \tilde{\Sigma}_0 + \tilde{\Sigma}_1,$$

where $\tilde{\Sigma}_0$ and $\tilde{\Sigma}_1$ denote the contributions to the right-hand side of (13.14) from those solutions to (13.15) for which $g_1(m, A, B, a) \not\equiv 0 \pmod{p}$ and those for which $g_1(m, A, B, a) \equiv 0 \pmod{p}$, respectively.

We first consider $\tilde{\Sigma}_0$. According to Lemma 22, the congruence (13.15) has $\mathrm{O}(1)$ solutions such that $g_1(m, A, B, a) \not\equiv 0 \pmod{p}$. Moreover, in this case, the inner sum in (13.14) is a non-trivial quadratic Gauß sum and is $\mathrm{O}(p^{j/2})$. Combining everything, we find that

$$\tilde{\Sigma}_0 \ll p^{\kappa + (j/2)} = p^{s/2}.$$

We next consider the sum $\tilde{\Sigma}_1$; note that this sum can only be nonempty if $p \nmid a$. Let the finite set T and $\tilde{\rho}(\ell)$ be as in Lemma 27. The number of solutions of (13.15) such that $g_1(m, A, B, a) \equiv 0 \pmod{p}$ is

$$\mathrm{O}\left(p^{\lfloor \frac{1}{2} \min(\tilde{\rho}(k^2 a), \kappa) \rfloor}\right).$$

If $j = 0$, then this shows that

$$\tilde{\Sigma}_1 \ll p^{s/2 + \lfloor \frac{1}{2} \min(\tilde{\rho}(k^2 a), \kappa) \rfloor}.$$

If $j = 1$, then the inner sum in (13.14) is actually a complete exponential sum with a linear phase, so that only the terms with $g(m, A, B, a) \equiv k \pmod{p^{\kappa+1}}$ contribute. We find that, in this case,

$$\tilde{\Sigma}_1 \ll p^{\kappa + \lfloor \frac{1}{2} \min(\tilde{\rho}(k^2 a), \kappa+1) \rfloor} = p^{s/2 + \lfloor \frac{1}{2} \min(\tilde{\rho}(k^2 a), \kappa+1) \rfloor - \frac{1}{2}}.$$

Putting everything together, we have proved the following estimate.

Lemma 28. *Let $p > 3$, and $p \nmid A$ or $p \nmid B$, and $s \geqslant 2$.*

(1) *If $p \nmid a$, then, letting the finite set $T \subset \mathbb{Z} \setminus p\mathbb{Z}$ and $\tilde{\rho}(\ell)$ be as in Lemma 27, we have that*

$$\tilde{\Sigma} \ll p^{s/2} + p^{\lfloor \frac{1}{2}s \rfloor + \min\left(\lfloor \frac{1}{2}\tilde{\rho}(k^2a)\rfloor, \lfloor \frac{1}{4}(s+1)\rfloor\right)}.$$

In particular, writing $q = q_\square^2 q_1$ with $q_1 \in \{1, p\}$, we have that

$$\tilde{\Sigma} \ll q^{1/2}\big(k^2a - T, q_\square\big)^{1/2},$$

as well as $\tilde{\Sigma} \ll q^{1/2}$ if $\tilde{\rho}(k^2a) \leqslant 2$ or if $s = 2$.

(2) *If $p \mid a$ and $A \not\equiv B \pmod p$, then*

$$\tilde{\Sigma} \ll q^{1/2}.$$

Combining Lemma 28 and the bound (13.6), which covers the case $s = 1$, we obtain Lemma 20. □

14. Proof of Theorem 2

In this section we indicate the necessary changes if f_1 and f_2 are Maaß forms. The Voronoi formula, Lemma 1, reads as follows (see e.g. [HM, Proposition 1]).

Lemma 29. *Let $c \in \mathbb{N}$, $b \in \mathbb{Z}$, and assume $(b, c) = 1$. Let V be a smooth compactly supported function, and let $N > 0$. Let $\lambda(n)$ denote the normalized Hecke eigenvalues of a cuspidal Maaß newform with spectral parameter t for $\mathrm{SL}_2(\mathbb{Z})$. Then*

$$\sum_n \lambda(n) e\left(\frac{bn}{c}\right) V\left(\frac{n}{N}\right) = \frac{N}{c} \sum_{\pm} \sum_n \lambda(n) e\left(\mp\frac{\bar{b}n}{c}\right) \mathring{V}^{\pm}\left(\frac{n}{c^2/N}\right)$$

where

$$\mathring{V}^{\pm}(y) = \int_0^\infty V(x) \mathcal{J}_{2it}^{\pm}(4\pi\sqrt{xy}) dx$$

with the notation as in (6.1).

Note that $\mathring{V}^{\pm}(y)$ is again a Schwartz class function. The Gamma factors in the Mellin transform of the weight function (3.2) depend on the parity of χ, so we sum over odd and even characters separately[2]. Note that the root number of $L(s, f_1 \otimes \chi)\overline{L(s, f_2 \otimes \chi)}$ is the product of the signs of f_1 and f_2, and in particular independent of χ. This yields a congruence condition $n \equiv \pm m \pmod d$ for various divisors $d \mid q$. The treatment of the diagonal term $n = m$ remains unchanged, but in (3.4) we define

$$S_{N,M,d,q} := \frac{d}{(NM)^{1/2}} \sum_{\substack{n \equiv \pm m \bmod d \\ (nm,q)=1 \\ n \neq m}} \lambda_1(m)\lambda_2(n) V_1\left(\frac{m}{M}\right) V_2\left(\frac{n}{N}\right).$$

Correspondingly, the definition of $\mathcal{D}(\ell_1, \ell_2, h, N, M)$ in (3.7) is changed into

$$\mathcal{D}(\ell_1, \ell_2, h, N, M) = \sum_{\ell_1 n \mp \ell_2 m = h} \lambda_1(m)\lambda_2(n) V_1\left(\frac{\ell_2 m}{M}\right) V_2\left(\frac{\ell_1 n}{N}\right). \tag{14.1}$$

As remarked in [Bl1], the results in this paper hold for Maaß forms as well, and they are also insensitive to a change of sign in the summation condition. The proof of Proposition 6 in Section 4 requires only some extra signs at the appropriate places.

[2]The dependence of the Gamma factors and the root number on the parity of χ is missing in [St, p. 3-4].

In Sections 7 and 8, we need to keep track of various extra signs, which arise from two principal sources while following the arguments in Subsections 7.1 and 7.2. One source of extra signs comes from (14.1), so that the analogue of (7.7) is

$$\begin{aligned}&\mathcal{D}_{z,\eta}(\ell_1,\ell_2,h,N,M)\\&=\frac{1}{\Lambda}\sum_{\ell_1\ell_2|c}w_0\left(\frac{c}{C}\right)\sideset{}{^*}\sum_{d \bmod c}\sum_{n,m}\lambda_1(m)\lambda_2(n)e\left(\frac{d}{c}(\ell_1 n\mp\ell_2 m-h)\right)W_{\eta M}\left(\pm\frac{\ell_1 n-h}{M}\right)V_{\pm z,\eta M}\left(\frac{\ell_2 m}{M}\right)\end{aligned}$$

The other source of extra signs are the two applications of Lemma 29 in the situation of (7.8) and (7.9). In (7.9), we encounter integral transforms of the shape

$$W^*_{\eta M}\left(\frac{h\ell_1 n}{c^2},\pm\frac{M\ell_1 n}{c^2}\right)$$

where

$$W^*_{\eta M}(z,w)=\int_0^\infty W_{\eta M}(y)\mathcal{J}^{\pm}_{2it}(4\pi\sqrt{yw+z})dy.$$

Here w can be negative, but by (7.2) we can guarantee $5|w|\leqslant z$, and we always have $z\geqslant(4C^2)^{-1}$. In particular, we can add a smooth redundant weight function $W_0(h\ell_1 nc^{-2},\pm M\ell_1 nc^{-2})$ with $z_0=(4C^2)^{-1}$ as in the remark after Corollary 10 without changing the expression.

Now Lemma 9 applies to the relevant integral transform with $\mathcal{J}^+_{2it}$ in place of $J_{\kappa-1}$ (here we assume the Selberg eigenvalue conjecture, i.e. $t\in\mathbb{R}$, for convenience). For $\mathcal{J}^-_{2it}$ one can simply use the rapid decay of the Bessel-K-function to obtain a trivial decomposition of the type (6.14) with

$$W_+(z,w)=W^*(z,w)e(-2\sqrt{z})=\int_0^\infty W(y)\mathcal{J}^-_{2it}(4\pi\sqrt{yw+z})dy\,e(-2\sqrt{z})$$

and $W_-(z,w)=0$. This satisfies the stronger bound

$$z^i|w|^j\frac{\partial^i}{\partial z^i}\frac{\partial^j}{\partial w^j}W_\pm(z,w)\begin{cases}=0, & z\geqslant C^\varepsilon,\\ \ll C^{\varepsilon(i+j)}, & \text{otherwise.}\end{cases}\tag{14.2}$$

for any $i,j\in\mathbb{N}_0$.

Hence in the case of terms involving $\mathcal{J}^+_{2it}$ in the application of Lemma 29 to (7.9), the ranges in (7.11) remain the same. In the case of terms involving $\mathcal{J}^-_{2it}$, we have even stronger conditions

$$\ell_1 n\leqslant\mathcal{N}_0^-:=\frac{C^{2+\varepsilon}}{N},\quad \ell_2 m\leqslant\mathcal{M}_0\tag{14.3}$$

from (14.2). At the end of subsection 7.2, we thus end up with the spectral analysis of terms involving six types of Kloosterman sums:

- *Case I:* $S(\ell_1 n-\ell_2 m,h,c)$, $\ell_1 n>\ell_2 m$. This is the case of Σ_+.
- *Case II:* $S(\ell_1 n-\ell_2 m,h,c)$, $\ell_1 n<\ell_2 m$. This is the case of Σ_-.
- *Case III:* $S(-\ell_1 n-\ell_2 m,h,c)$ under the size constraint (14.3).
- *Case IV:* $S(\ell_1 n+\ell_2 m,h,c)$.
- *Case V:* $S(-\ell_1 n+\ell_2 m,h,c)$, $\ell_1 n>\ell_2 m$ under the size constraint (14.3).
- *Case VI:* $S(-\ell_1 n+\ell_2 m,h,c)$, $\ell_1 n<\ell_2 m$ under the size constraint (14.3).

Case IV is identical to Case I with minor sign changes. Cases III, V and VI are much simpler than Cases I and II because of the stronger size conditions (14.3) (coming from the rapid decay of the Bessel K-function), but formally one can treat Case VI as Case I using the same sign Kuznetsov formula, and Cases III and V as Case II using the opposite sign formula. Note that $\mathcal{N}_0^-\leqslant\mathcal{M}_0$, so that in the notation of Sections 7 and 8 we automatically have $\mathcal{M},\mathcal{K},\mathcal{N}\leqslant\mathcal{M}_0$ if (14.3) holds. □

15. Proof of Theorem 4

Let V be a fixed smooth function that is 1 on $[0,1]$ and vanishes on $[2,\infty)$. Let

$$X := q^{1/1000}.$$

Define

$$A(\chi) := \sum_{a,b} \frac{\lambda_1(a)\lambda_2(b)(\chi(a)\bar\chi(b)+\bar\chi(a)\chi(b))}{\sqrt{ab}} V\left(\frac{ab}{X}\right).$$

By the Cauchy-Schwarz inequality, we have

$$\Big|\sideset{}{^*}\sum_{\chi \bmod q} L(1/2,f_1\otimes\chi)\overline{L(1/2,f_2\otimes\chi)}A(\chi)\Big|^2 \leqslant \sideset{}{^*}\sum_{\chi \bmod q} \big(L(1/2,f_1\otimes\chi)\overline{L(1/2,f_2\otimes\chi)}\big)^2 \sideset{}{^*}\sum_{\chi \bmod q} A(\chi)^2.$$

Note that both $A(\chi)$ and $L(1/2,f_1\otimes\chi)\overline{L(1/2,f_2\otimes\chi)}$ are real (cf. (3.1)), so that we do not need absolute values on the right hand side. We conclude

$$\sideset{}{^*}\sum_{\chi \bmod q} \Big(L(1/2,f_1\otimes\chi)\overline{L(1/2,f_2\otimes\chi)}\Big)^2 \geqslant \frac{|S_1|^2}{S_2}, \tag{15.1}$$

where

$$S_1 := \sideset{}{^*}\sum_{\chi \bmod q} L(1/2,f_1\otimes\chi)\overline{L(1/2,f_2\otimes\chi)}A(\chi), \quad S_2 := \sideset{}{^*}\sum_{\chi \bmod q} A(\chi)^2.$$

Clearly,

$$S_2 = 2\sum_{d\mid q}\phi(d)\mu(q/d)\Big(\sum_{\substack{a_1a_2\equiv b_1b_2 \bmod d\\ (a_1a_2b_1b_2,p)=1}} + \sum_{\substack{a_1b_2\equiv a_2b_1 \bmod d\\ (a_1a_2b_1b_2,p)=1}}\Big)\frac{\lambda_1(a_1)\lambda_1(a_2)\lambda_2(b_1)\lambda_2(b_2)}{\sqrt{a_1a_2b_1b_2}} V\left(\frac{a_1b_1}{X}\right)V\left(\frac{a_2b_2}{X}\right).$$

Here $d\in\{q,q/p\}$, and the support of V implies that the congruences are equalities, so that

$$S_2 = 2\psi(q)\Big(\sum_{\substack{a_1a_2=b_1b_2\\ (a_1a_2b_1b_2,p)=1}} + \sum_{\substack{a_1b_2=a_2b_1\\ (a_1a_2b_1b_2,p)=1}}\Big)\frac{\lambda_1(a_1)\lambda_1(a_2)\lambda_2(b_1)\lambda_2(b_2)}{\sqrt{a_1a_2b_1b_2}} V\left(\frac{a_1b_1}{X}\right)V\left(\frac{a_2b_2}{X}\right) = S_{21}+S_{22},$$

say. By Mellin inversion we have

$$S_{21} = 2\psi(q)\int_{(1)}\int_{(1)} \widehat V(s)\widehat V(t)X^{s+t} \sum_{\substack{a_1a_2=b_1b_2\\ (a_1a_2b_1b_2,p)=1}} \frac{\lambda_1(a_1)\lambda_1(a_2)\lambda_2(b_1)\lambda_2(b_2)}{(a_1b_1)^{s+\frac12}(a_2b_2)^{t+\frac12}}\frac{ds\,dt}{(2\pi i)^2}.$$

In $\Re s,\Re t>-1/10$, say, the double Dirichlet series can be expanded into an Euler product:

$$\prod_{p\nmid\ell}\Big(1+\frac{2\lambda_1(\ell)\lambda_2(\ell)}{\ell^{1+s+t}}+\frac{\lambda_1(\ell)\lambda_2(\ell)}{\ell^{1+2s}}+\frac{\lambda_1(\ell)\lambda_2(\ell)}{\ell^{1+2t}}+O\big(\frac{1}{\ell^{3/2}}\big)\Big)$$
$$= L(1+s+t,f_1\times f_2)^2L(1+2s,f_1\times f_2)L(1+2t,f_1\times f_2)H_{21}(s,t)$$

with a holomorphic Euler product $H_{21}(s,t)$ that converges absolutely in $\Re s,\Re t>-1/10$ and is uniformly bounded (from above and beyond) in q in the same vertical strip. Shifting contours, we find that

$$S_{21} = 2\psi(q)L(1,f_1\times f_2)^4H(0,0)+O(\psi(q)X^{-1/10}).$$

Similarly, we have

$$S_{22} = 2\psi(q)\int_{(1)}\int_{(1)} L(1+s+t,f_1\times f_1)L(1+s+t,f_2\times f_2)L(1+2s,f_1\times f_2)L(1+2t,f_1\times f_2)$$
$$\times\widehat V(s)\widehat V(t)X^{s+t}H_{22}(s,t)\frac{ds\,dt}{(2\pi i)^2}.$$

The integrand in this double integral has a double pole at $s+t=0$ and two simple poles at $s=0$ and $t=0$. We first shift to $\Re s, \Re t = 1/20$. Then we shift to $\Re s = -1/10$, picking up two poles at $s=0$ at $s=-t$, and then to $\Re t = -1/10$ picking up one pole at $t=0$. In this way we obtain $S_{22} \asymp \psi(q)(\log X)^2 \asymp \psi(q)(\log q)^2$, and we conclude

$$S_2 \ll \psi(q)(\log q)^2. \tag{15.2}$$

Next we turn to the analysis of S_1. Here we use the approximate functional equation (3.1) to write

$$S_1 = 2\sum_{d|q}\phi(d)\mu(q/d)\Big(\sum_{\substack{a_1a_2\equiv b_1b_2 \bmod d\\(a_1a_2b_1b_2,p)=1}} + \sum_{\substack{a_1b_2\equiv a_2b_1 \bmod d\\(a_1a_2b_1b_2,p)=1}}\Big)\frac{\lambda_1(a_1)\lambda_1(a_2)\lambda_2(b_1)\lambda_2(b_2)}{\sqrt{a_1a_2b_1b_2}}W\left(\frac{a_1b_1}{q^2}\right)V\left(\frac{a_2b_2}{X}\right).$$

Note that this has (by design) the same shape as S_2, except that the range of summation of the a_1, b_1 variables is much longer. We decompose $S_1 = M_1 + E_1$ where M_1 represents the diagonal contributions and E_1 is the rest. By the same argument as before, we find that

$$M_1 \asymp \psi(q)\log X \log q \asymp \psi(q)(\log q)^2. \tag{15.3}$$

For the error term, we first estimate trivially (using (2.7) for convenience)

$$E_1 \ll q\sum_{d\in\{q,q/p\}}\sum_{\substack{a_2,b_2\ll X\\(a_2b_2,p)=1}}\frac{1}{(a_2b_2)^{1/2-\varepsilon}}\Big|\Big(\sum_{\substack{a_1a_2\equiv b_1b_2 \bmod d\\(a_1b_1,p)=1\\a_1a_2\neq b_1b_2}} + \sum_{\substack{a_1b_2\equiv a_2b_1 \bmod d\\(a_1b_1,p)=1\\a_1b_2\neq a_2b_1}}\Big)\frac{\lambda_1(a_1)\lambda_2(b_1)}{\sqrt{a_1b_1}}W\left(\frac{a_1b_1}{q^2}\right)\Big|.$$

We show in detail how to treat the first term, since the second one is very similar. Injecting a smooth partition of unity and arguing as in the beginning of Section 3.2, we need to estimate

$$E(A,B) := \frac{q}{\sqrt{AB}}\sum_{d\in\{q,q/p\}}\sum_{\substack{a_2,b_2\ll X\\(a_2b_2,p)=1}}\frac{1}{(a_2b_2)^{1/2-\varepsilon}}\Big|\sum_{\substack{a_1a_2\equiv b_1b_2 \bmod d\\(a_1b_1,p)=1\\a_1a_2\neq b_1b_2}}\lambda_1(a_1)\lambda_2(b_1)V_1\left(\frac{a_1}{A}\right)V_2\left(\frac{b_1}{B}\right)\Big|$$

for smooth compactly supported weight functions V_1, V_2 satisfying (3.5), and $AB \ll q^{2+\varepsilon}$. Without loss of generality, consider the case $B \geqslant A$. We estimate $E(A,B)$ in two ways. First, we remove the coprimality condition by Möbius inversion, getting

$$\begin{aligned}E(A,B) &\ll \frac{q}{\sqrt{AB}}\sum_{d\in\{q,q/p\}}\sum_{\substack{a_2,b_2\ll X\\(a_2b_2,p)=1}}\sum_{f|g|p}\frac{|\lambda_1(g/f)|}{(a_2b_2)^{1/2-\varepsilon}}\Big|\sum_{\substack{fga_1a_2\equiv b_1b_2 \bmod d\\fga_1a_2\neq b_1b_2}}\lambda_1(a_1)\lambda_2(b_1)V_1\left(\frac{fga_1}{A}\right)V_2\left(\frac{b_1}{B}\right)\Big|\\ &\ll \frac{q}{\sqrt{AB}}\sum_{d\in\{q,q/p\}}\sum_{\substack{a_2,b_2\ll X\\(a_2b_2,p)=1}}\sum_{f|g|p}\frac{1}{(a_2b_2)^{1/2-\varepsilon}}|\mathcal{S}(fga_2,b_2,d,Aa_2,Bb_2)|,\end{aligned}$$

using the notation (3.8). Provided $A \gg BX$ or $B \gg AX$ with a sufficiently large implied constant, we find by Proposition 7 that

$$\begin{aligned}E(A,B) &\ll \frac{q^{1+\varepsilon}X}{\sqrt{AB}}\left(\frac{BX}{q^{1/2}} + \frac{B^{5/4}A^{1/4}X^{3/2}}{q} + \frac{B^{3/4}A^{1/4}X}{q^{1/4}} + \frac{BA^{1/2}X^{3/2}}{q^{3/4}}\right)\\ &\ll \frac{q^{1+\varepsilon}}{\sqrt{AB}}X^{5/2}\left(\frac{A^{1/4}B^{3/4}}{q^{1/4}} + \frac{B}{q^{1/2}}\right)\end{aligned} \tag{15.4}$$

since $AB \leqslant q^{2+\varepsilon}$. If $A \ll BX \ll AX^2$, we have the individual bound

$$E(A,B) \ll \frac{q^{1+\varepsilon}X}{\sqrt{AB}}X^{5/2}B^{1/2+\theta}. \tag{15.5}$$

Note that, up to powers of X, this is comparable to (3.11) and (3.12) with N and M replaced with B and A respectively.

Alternatively, we write

$$E(A,B)=\frac{q}{\sqrt{AB}}\sum_{d\in\{q,q/p\}}\sum_{\substack{a_2,b_2\ll X\\(a_2b_2,p)=1}}\frac{1}{(a_2b_2)^{1/2-\varepsilon}}\Big|\sum_{p\nmid a_1}\lambda_1(a_1)V_1\left(\frac{a_1}{A}\right)\sum_{b_1\equiv a_2\overline{b_2}a_1 \bmod d}\lambda_2(b_1)V_2\left(\frac{b_1}{B}\right)\Big|.$$

Arguing as in Section 4, the innermost sum equals

$$\frac{1}{d}\sum_{r|d}\frac{B}{r}\sum_b S(a_2\overline{b_2}a_1,b,r)\lambda_2(b)\mathring{V}_2\left(\frac{bB}{r^2}\right),$$

and hence, by an application of the Cauchy-Schwarz inequality and (2.7),

$$E(A,B)\ll q\sum_{d\in\{q,q/p\}}\sum_{\substack{a_2,b_2\ll X\\(a_2b_2,p)=1}}\frac{1}{(a_2b_2)^{1/2-\varepsilon}}\sum_{r|d}\frac{B^{1/2}}{dr}\Big(\sum_{n_1,n_2\ll r^2q^\varepsilon/B}|\mathcal{S}_A(a_2\overline{b_2}n_1,a_2\overline{b_2}n_2,r)|\Big)^{1/2},$$

where, as in (4.2), we write

$$\mathcal{S}_A(a_2\overline{b_2}n_1,a_2\overline{b_2}n_2,r)=\sum_{\substack{m\asymp A\\(m,p)=1}}S(m,a_2\overline{b_2}n_1,r)S(m,a_2\overline{b_2}n_2,r).$$

By Theorem 5, we conclude as in the proof of Proposition 6 that

$$E(A,B)\ll\frac{q^{1+\varepsilon}}{B^{1/2}}X\left(A^{1/4}q^{7/12}+A^{1/2}q^{5/12}+q^{2/3}\right). \tag{15.6}$$

Combining (15.4), (15.5), and (15.6), we conclude as in Section 3.3 that

$$E(A,B)\ll q^{65/66+\varepsilon}X^{5/2}.$$

Together with (15.3), this estimate shows that

$$S_1\gg\psi(q)(\log q)^2. \tag{15.7}$$

Combining (15.1), (15.2), and (15.7), we complete the proof of Theorem 4. □

References

[Ak] A. Akbary, *Simultaneous non-vanishing of twists*, Proc. Amer. Math. Soc. **134** (2006), 3143-3151

[Bl1] V. Blomer, *Shifted convolution sums and subconvexity bounds for automorphic L-functions*, Int. Math. Res. Not. **2004**, 3905-3926

[Bl2] V. Blomer, *Non-vanishing of class group L-functions at the central point*, Annales de l'Institut Fourier **54** (2004), 831-847

[BFKMM] V. Blomer, E. Fouvry, E. Kowalski, P. Michel, D. Milićević, *On the fourth moment of Dirichlet L-functions*, preprint

[BH1] V. Blomer, G. Harcos, *Spectral decomposition of shifted convolution sums*, Duke Math. J. **144** (2008), 321-339

[BH2] V. Blomer, G. Harcos, *Hybrid bounds for twisted L-functions*, J. Reine Angew. Math. **621** (2008), 53-79

[BHM] V. Blomer, G. Harcos, P. Michel, *A Burgess-like subconvex bound for twisted L-functions* (with appendix 2 by Z. Mao), Forum Math. **19** (2007), 61-105

[BKY] V. Blomer, R. Khan, M. Young, *Mass distribution of holomorphic cusp forms*, Duke Math. J. **162** (2013), 2609-2644

[BM] V. Blomer, D. Milićević, *p-adic analytic twists and strong subconvexity*, preprint

[Bo] E. Bombieri, *On exponential sums in finite fields*, Amer. J. Math. **88** (1966), 71-105

[Br] F. Brumley, *Effective multiplicity one on* GL_N *and narrow zero-free regions for Rankin-Selberg L-functions*, Amer. J. Math. **128** (2006), 1455-1474

[Ch] G. Chinta, *Analytic ranks of elliptic curves over cyclotomic fields*, J. reine angew. Math. **544** (2002), 13-24

[DF] R. Dabrowski, B. Fisher, *A stationary phase formula for exponential sums over* $\mathbb{Z}/p^m\mathbb{Z}$ *and applications to* GL(3)*-Kloosterman sums*, Acta Arith. **80** (1997), 1-48

[De] P. Deligne, *La conjecture de Weil. I.* Inst. Hautes Études Sci. Publ. Math. **43** (1974), 273-307

[DI] J.-M. Deshouillers, H. Iwaniec, *Kloosterman sums and Fourier coefficients of cusp forms*, Invent. Math. **70** (1982/83) 219-288

[EMOT] A. Erdélyi, W. Magnus, F. Oberhettinger, F. Tricomi, *Higher transcendental functions II*, McGraw-Hill 1953

[FGKM] E. Fouvry, S. Ganguly, E. Kowalski, P. Michel, *Gaussian distribution for the divisor function and Hecke eigenvalues in arithmetic progressions*, Comm. Math. Helv., to appear

[FKM] E. Fouvry, E. Kowalski, P. Michel, *A study in sums of products*, preprint

[FMRS] E. Fouvry, P. Michel, J. Rivat, A. Sárkőzy, *On the pseudorandomness of the signs of Kloosterman sums*, J. Aust. Math. Soc. **77** (2004), 425-436

[GKR] P. Gao, R. Khan, G. Ricotta, *The second moment of Dirichlet twists of Hecke L-functions*, Acta Arith. **140** (2009), 57-65.

[Go] A. Good, *The mean square of Dirichlet series associated with cusp forms*, Mathematika **29** (1982), 278-295

[GR] I.S. Gradshteyn, I.M. Ryzhik, *Table of integrals, series, and products*, sixth edition, Academic Press, Inc., San Diego, CA, 2000

[HM] G. Harcos, P. Michel, *The subconvexity problem for Rankin-Selberg L-functions and equidistribution of Heegner points. II*, Invent. Math. **163** (2006), 581-655

[HB] D. R. Heath-Brown, *Hybrid bounds for Dirichlet L-functions*, Invent. Math. **47** (1978), 149-170

[HL] J. Hoffstein, M. Lee, *Second moments and simultaneous non-vanishing of* GL(2) *automorphic L-series*, `arXiv:1308.5980`

[HLo] J. Hoffstein, P. Lockhart, *Coefficients of Maass forms and the Siegel zero*, with an appendix by D. Goldfeld, J. Hoffstein, and D. Lieman, Ann. of Math. (2) **140** (1994), 161-181

[IK] H. Iwaniec, E. Kowalski, *Analytic Number Theory*, Colloquium Publication **53** (2004), AMS, Providence, RI

[ILS] H. Iwaniec, W. Luo, P. Sarnak, *Low lying zeros of families of L-functions*, Inst. Hautes Études Sci. Publ. Math. **91** (2000), 55-131

[IS] H. Iwaniec, P. Sarnak, *The non-vanishing of central values of automorphic L-functions and Landau-Siegel zeros*, Israel J. Math. **120** (2000), part A, 155-177.

[IM] A. Ivić, Y. Motohashi, *On the fourth power moment of the Riemann zeta function*, J. Number Theory **51** (1995), 16-45

[J1] M. Jutila, *Transformations of exponential sums*, Proceedings of the Amalfi Conference on Analytic Number Theory (Maiori 1989). Univ. Salerno, Salerno 1992, 263-270

[J2] M. Jutila, *A variant of the circle method*, in: Sieve methods, exponential sums and their applications in number theory, 245-254. Cambridge University Press, 1996

[J3] M. Jutila, *Convolutions of Fourier coefficients of cusp forms*, Publ. Inst. Math. (Beograd) **65 (79)** (1999), 31-51

[Ka] N. Katz, *Gauss sums, Kloosterman sums and monodromy groups*, Ann. of Math. Stud. **116**, Princeton University Press, 1988

[Kh] R. Khan, *Simultaneous non-vanishing of* GL(3) $\times$ GL(2) *and* GL(2) *L-functions*, Math. Proc. Cambridge Philos. Soc. **152** (2012), 535-553.

[KS] H. Kim, *Functoriality for the exterior square of* GL(4) *and symmetric fourth of* GL(2), Appendix 1 by Dinakar Ramakrishnan; Appendix 2 by Henry H. Kim and Peter Sarnak, J. Amer. Math. Soc. **16** (2003), 139-183.

[KMV] E. Kowalski, P. Michel, J. VanderKam, *Mollification of the fourth moment of automorphic L-functions and arithmetic applications*, Invent. math. **142** (2000), 95-151

[Li] X. Li, *The central value of the Rankin-Selberg L-functions*, Geom. Funct. Anal. **18** (2009), 1660-1695

[Mi] D. Milićević, *Sub-Weyl subconvexity for Dirichlet L-functions to prime power moduli*, preprint

[Mot] Y. Motohashi, *Spectral theory of the Riemann zeta-function*, Cambridge tracts in mathematics **127**, Cambridge 1997

[PM] D. H. J. Polymath, *New equidistribution estimates of Zhang type and bounded gaps between primes*, arXiv:1402.0811

[Po] A. G. Postnikov, *On the sum of characters with respect to a modulus equal to a power of a prime number*, Izv. Akad. Nauk. SSSR Ser. Mat. **19** (1955), 11-16

[Ra] R. A. Rankin, *The vanishing of Poincaré series*, Proc. Edin. Math. Soc. **23** (1980), 151-161

[Roh] D. Rohrlich, *On L-functions of elliptic curves and cyclotomic towers*, Invent. Math. **75** (1984), 409-423

[Ro] D. Rouymi, *Formules de trace et non-annulation de fonctions L automorphes au niveau* $\mathfrak{p}^\nu$, Acta Arith. **147** (2011), 1-32

[RS] Z. Rudnick, K. Soundararajan, *Lower bounds for moments of L-functions*, Proc. Nat. Acad. Sci. **102** (19), 6837-6838

[RS1] Z. Rudnick, K. Soundararajan, *Lower bounds for moments of L-functions: symplectic and orthogonal examples*, in: Proceedings of the Bretton Woods Workshop on Multiple Dirichlet Series, Proceedings of Symposia in Pure Mathematics **75**, American Mathematical Society 2006.

[RW] D. Ramakrishnan, S. Wang, *On the exceptional zeros of Rankin-Selberg L-functions*, Compositio Math. **135** (2003), 211-244

[Sa] P. Sarnak, *Estimates for Rankin-Selberg L-functions and quantum unique ergodicity*, J. Funct. Anal. **184** (2001), 419-453

[Sh] G. Shimura, *The special values of the zeta functions associated with cusp forms*, Comm. Pure Appl. Math. 29 (1976), 783-804.

[St] T. Stefanicki, *Non-vanishing of L-functions attached to automorphic representations of* GL(2) *over* $\mathbb{Q}$, J. Reine Angew. Math. **474** (1996), 1-24

[Y] M. Young, *The fourth moment of Dirichlet L-functions*, Ann. of Math. (2) **173** (2011), 1-50.

[Za] N. I. Zavorotnyi, *On the fourth moment of the Riemann zeta-function*, in: Automorphic Functions and Number Theory 2, Computation Center of the Far East Branch of the Science Academy of USSR 1989, 69-125 (in Russian)

Mathematisches Institut, Bunsenstr. 3-5, D-37073 Göttingen, Germany
E-mail address: blomer@uni-math.gwdg.de

Bryn Mawr College, Department of Mathematics, 101 North Merion Avenue, Bryn Mawr, PA 19010, U.S.A.
E-mail address: dmilicevic@brynmawr.edu