ON MOMENTS OF TWISTED L-FUNCTIONS

VALENTIN BLOMER, ÉTIENNE FOUVRY, EMMANUEL KOWALSKI, PHILIPPE MICHEL,
AND DJORDJE MILIČEVIĆ

ABSTRACT. We study the average of the product of the central values of two L-functions of modular forms twisted by Dirichlet characters to a large prime modulus. When at least one of the modular forms is non-cuspidal, we obtain an asymptotic formula with a power saving error term. In particular, when both are non-cuspidal, this gives a significant improvement on M. Young’s asymptotic evaluation of the fourth moment of Dirichlet L-functions. In the general case the asymptotic formula with power saving is proved under a conjectural estimate for bilinear forms in Kloosterman sums.

1. INTRODUCTION

This paper was motivated by the beautiful work of Matthew Young on the fourth moment of Dirichlet L-functions for prime moduli [24]: for a prime $q > 2$, let

$$M_4(q) := \frac{1}{\varphi^*(q)} \sum_{\chi \mod q, \chi \text{ primitive}} |L(\chi, 1/2)|^4,$$

where $\varphi^*(q)$ is the number of primitive Dirichlet characters modulo $q$, and $L(\chi, s) = \sum_{n \geq 1} \frac{\chi(n)}{n^s}$, $\Re s > 1$, is the Dirichlet L-function. Young obtained the asymptotic formula

$$M_4(q) = P_4(\log q) + O(q^{-\frac{1}{80}(1-2\theta)^+}+\varepsilon)$$

for any $\varepsilon > 0$, where $P_4$ is a polynomial of degree four with leading coefficient $1/(2\pi^2)$, and here and in the following the constant $\theta = 7/64$ is the best known approximation towards the Ramanujan-Petersson conjecture (due to Kim and Sarnak [19]).

The fourth moment of Dirichlet L-functions is a special case of the more general second moment

$$M_{f,g}(q) = \frac{1}{\varphi^*(q)} \sum_{\chi \text{ primitive}} L(f \otimes \chi, 1/2)\overline{L(g \otimes \chi, 1/2)},$$

for any $\varepsilon > 0$, where $P_4$ is a polynomial of degree four with leading coefficient $1/(2\pi^2)$, and here and in the following the constant $\theta = 7/64$ is the best known approximation towards the Ramanujan-Petersson conjecture (due to Kim and Sarnak [19]).

The fourth moment of Dirichlet L-functions is a special case of the more general second moment

$$M_{f,g}(q) = \frac{1}{\varphi^*(q)} \sum_{\chi \text{ primitive}} L(f \otimes \chi, 1/2)\overline{L(g \otimes \chi, 1/2)},$$

for any $\varepsilon > 0$, where $P_4$ is a polynomial of degree four with leading coefficient $1/(2\pi^2)$, and here and in the following the constant $\theta = 7/64$ is the best known approximation towards the Ramanujan-Petersson conjecture (due to Kim and Sarnak [19]).

The fourth moment of Dirichlet L-functions is a special case of the more general second moment

$$M_{f,g}(q) = \frac{1}{\varphi^*(q)} \sum_{\chi \text{ primitive}} L(f \otimes \chi, 1/2)\overline{L(g \otimes \chi, 1/2)},$$
where $q$ is an integer with $q \not\equiv 2 \pmod{4}$ (since otherwise there are no primitive characters modulo $q$), and $f$ and $g$ denote two fixed Hecke eigenforms, not necessarily cuspidal, with respective Hecke eigenvalues $(\lambda_f(n))_{n \geq 1}$, $(\lambda_g(n))_{n \geq 1}$, and

$$L(f \otimes \chi, s) = \sum_{n \geq 1} \frac{\lambda_f(n) \chi(n)}{n^s}, \quad L(g \otimes \chi, s) = \sum_{n \geq 1} \frac{\lambda_g(n) \chi(n)}{n^s} \quad (\Re s > 1)$$

denote the associated twisted $L$-functions: indeed, let $E(z)$ denote the central derivative of the Eisenstein series $E(z, s)$, i.e.,

$$E(z) = \frac{\partial}{\partial s} \bigg|_{s=1/2} E(z, s), \quad \text{with} \quad E(z, s) = \frac{1}{2} \sum_{(c,d)=1} \frac{y^s}{cz + d^{2s}}.$$  

This is a Hecke eigenform of level 1 with Hecke eigenvalues given by the usual divisor function

$$\lambda_E(n) = d(n) = \sum_{ab=n} 1$$

(see [16] §3.4 for instance). We have $L(\chi, s)^2 = L(E \otimes \chi, s)$, and therefore

$$M_4(q) = M_{E,E}(q).$$

Our first main result is a significant improvement of the error term in the fourth moment of Dirichlet $L$-functions [11].

**Theorem 1.1.** Let $q$ be a prime. Then for any $\varepsilon > 0$, we have

$$M_4(q) = P_4(q \log q) + O(q^{-1/32+\varepsilon}).$$

Moreover, under the Ramanujan-Petersson conjecture, the exponent $1/32$ may be replaced by $1/24$.

Our second main result is an asymptotic formula for the “mixed” moment $M_{f,E}(q)$.

**Theorem 1.2.** Let $f$ be a holomorphic cuspidal Hecke eigenform of level 1 and $E$ the Eisenstein series (1.3). Let $q$ be a prime number. Then for any $\varepsilon > 0$, we have

$$M_{f,E}(q) = \frac{1}{\varphi^*(q)} \sum_{\chi \text{ (mod q) primitive}} L(f \otimes \chi, 1/2)L(\chi, 1/2)^2 = \frac{L(f, 1)^2}{\zeta(2)} + O_{f,\varepsilon}(q^{-1/68+\varepsilon}).$$

Our final result establishes an asymptotic formula for the moment $M_{f,g}(q)$, conditionally on a bound for a certain family of algebraic exponential sums.

**Theorem 1.3.** Let $f, g$ be distinct holomorphic cuspidal Hecke eigenforms of level 1 and weights $k_f \equiv k_g \pmod{4}$. Let $q$ be a prime number. Assume that Conjecture 6.5 below holds. Then for any $\varepsilon > 0$, we have

$$M_{f,g}(q) = \frac{1}{\varphi^*(q)} \sum_{\chi \text{ (mod q) primitive}} L(f \otimes \chi, 1/2)L(g \otimes \chi, 1/2) = \frac{2L(f \otimes g, 1)}{\zeta(2)} + O_{f,g,\varepsilon}(q^{-1/144+\varepsilon}),$$

where $L(f \otimes g, 1) \neq 0$ is the value at 1 of the Rankin-Selberg $L$-function of $f$ and $g$, and

$$M_{f,f}(q) = \frac{1}{\varphi^*(q)} \sum_{\chi \text{ (mod q) primitive}} |L(f \otimes \chi, 1/2)|^2 = P_f(q \log q) + O_{f,\varepsilon}(q^{-1/144+\varepsilon}),$$

where $P_f(X)$ is an explicit polynomial of degree 1 with coefficients independent of $q$ and leading coefficient $2L(\text{sym}^2 f, 1)/\zeta(2)^{-1}$, where $L(\text{sym}^2 f, s)$ denotes the symmetric square $L$-function of $f$.  


Remark 1.4. As a rule of thumb, the asymptotic evaluation of $M_{f,g}(q)$ with a good error term gets significantly more challenging as the set \{f, g\} contains more cusp forms. (On the other hand, the main term in the asymptotic expansion of $M_{f,g}(q)$ gets more complicated as the set \{f, g\} contains more Eisenstein series.)

When $f = g$ and $f$ is cuspidal, this problem was studied by Stefanicki [23] and Gao, Khan and Ricotta [13]. In both cases, however, the error term gives only a saving of (at most) a small power of $\log q$ over the main term.

Our results are obtained by combining two new ingredients, each of which is of independent interest:

- Recently, Blomer and Miličević developed in [3] a method to handle shifted convolution sums of Fourier coefficients of cusp forms $f, g$ which, roughly speaking, gives estimates of the same quality as those that would follow from the Ramanujan-Petersson conjecture for the forms appearing in the spectral decomposition. In this paper, we adapt this method (which was designed to treat particularly long shift variables) to obtain similar uniform estimates also when $f$ or $g$ is the Eisenstein series $E$ (see Section 4).

- Being able to efficiently treat shifted convolution sums reduces the problem of an asymptotic formula for $M_{f,g}(q)$ to the question of bounding non-trivially a certain bilinear form in the Kloosterman sums $K_{2}(amn; q)$ when the summation variables $m, n$ are barely beyond the Pólya-Vinogradov range. Such a bound was obtained in [3] when $q$ is suitably composite by a variant of the $q$-van der Corput method. To deal with the case where $q$ is prime, we build on the recent works of Fouvy, Kowalski and Michel [9, 10], who obtained strong bounds for various sums involving general trace functions (in particular, Kloosterman sums) associated to $\ell$-adic sheaves modulo primes. In this paper, we use some of the methods developed there and extend them further. In particular, we develop different techniques to bound non-trivially bilinear forms in $K_{2}(amn; q)$ (according to the shape of the form), which reduce the problem to obtaining a nontrivial estimate for a certain complete sum over $\mathbb{F}_{q}$. Using this approach, we are able to bound non-trivially bilinear forms in $K_{2}(amn; q)$ when one or both summation variables $m, n$ are smooth (that is, appear with smooth coefficients); these bounds, combined with the previous ingredient, lead to Theorems 1.1 and 1.2.

We also show how estimates on a different class of multiple complete sums over $\mathbb{F}_{q}$ would lead to the case of a general bilinear form and hence to Theorem 1.3.

We refer the reader to Theorem 6.1 in Section 6 for a collection of general bounds for bilinear forms in Kloosterman sums.

Remark 1.5. (1) The combination of these arguments leads, in the special case of Young’s Theorem, not only to stronger results, but also to a different and perhaps more streamlined approach.

(2) The mixed asymptotic formula of Theorem 1.2 with some power saving error term could be obtained by combining the arguments of [4] with eitherYoung’s argument or the ones of [5] but it is the combination of the three which makes it eventually possible to reach the saving $q^{1/68}$.

(3) The assumption that the cusp forms in Theorem 1.3 are holomorphic is used to ensure the validity of the Ramanujan-Petersson bound for the Hecke eigenvalues of $f$ (proven by Deligne in the holomorphic case [4]): we have

$$|\lambda_{f}(n)| \leq d(n)$$

for all $n \geq 1$, where $\lambda_{f}(n)$ denotes the (normalized) Hecke eigenvalue of $f$.

We emphasize, however, that Theorem 1.2 is independent of Deligne’s bound (1.5) and uses only a Rankin-Selberg type bound

$$\sum_{n \leq x} |\lambda_{f}(n)|^{2} \ll_{f} x$$
and a weak individual bound

\[ \lambda_f(n) \ll n^{1/4}. \]

One can then show more generally that Theorem 1.2 holds for \( f \) a Hecke eigenform (holomorphic or not) of arbitrary level, while Theorem 1.3 is valid for \( f \) an arbitrary Hecke eigenform whose associated automorphic representation satisfies the Ramanujan–Petersson conjecture (at all places).

The structure of the paper is as follows. We collect some standard facts in Section 2. The following Section 3 presents an outline of the proof, presenting all key steps while omitting technical details. The proof is seen to depend on two crucial ingredients, the treatment of the shifted convolution sum problem and estimates of the bilinear sums of Kloosterman sums; these are the topic of Sections 4 and 6, respectively, while Section 5 recalls briefly M. Young’s method. Finally, Section 7 combines these inputs and presents the formal proofs of Theorems 1.1–1.3.

Acknowledgments. We would like to thank Ian Petrow and Paul Nelson for many discussions on M. Young’s work.

Notation and conventions. In the rest of this paper we will denote generically by \( W \), sometimes with subscripts, some smooth complex-valued functions, compactly supported on \([1/2, 2]\), whose derivatives satisfy

\[ W^{(j)}(x) \ll j, \varepsilon q^{j\varepsilon}, \]

for any \( \varepsilon > 0 \) and any \( j \geq 0 \), the implied constant depending on \( \varepsilon \) and \( j \) (but not on \( q \)). Sometimes even the stronger bounds \( W^{(j)}(x) \ll j \) hold.

From time to time, we will use the \( \varepsilon \)-convention, according to which \( \varepsilon > 0 \) is an arbitrarily small positive number whose value may change from line to line (e.g., the value of \( \varepsilon \) in (1.8) may be different for different functions \( W \)).

We denote \( e(z) = e^{2\pi iz} \), and for \( c \geq 1 \) an integer and \( a \in \mathbb{Z} \), we let \( e_c(a) = e(a/c) \) be the additive character modulo \( c \). We denote by

\[ S(a, b; c) = \sum_{d \equiv 0 \pmod{c}} e_c(ad + b \bar{d}) \]

the usual Kloosterman sum, and we also write

\[ \text{Kl}_2(a; c) = \frac{1}{\sqrt{c}} S(a, 1; c) \]

for the normalized Kloosterman sum.

We will use partitions of unity repeatedly in order to decompose a long sum over integers into smooth localized sums (see, e.g., [7, Lemme 2]):

Lemma 1.6. There exists a smooth non-negative function \( W(x) \) supported on \([1/2, 2]\) and satisfying (1.8) such that

\[ \sum_{k \geq 0} W\left( \frac{x}{2^k} \right) = 1 \]

for any \( x \geq 1 \).

2. Arithmetic and analytic reminders

We collect in this section some known preliminary facts concerning \( L \)-functions and automorphic forms. For many readers, it should be possible to skip this section in a first reading and to begin immediately reading the outline of the proof in Section 3.
2.1. Functional equations. Let $\chi$ be a non-principal character modulo a prime $q > 2$, and let $L(\chi, s)$ be its associated $L$-function. It admits an analytic continuation to $\mathbb{C}$ and satisfies a functional equation which we now recall (see [17, Theorem 4.15] for instance): let
\[
a = a(\chi) = \begin{cases} 
0 & \text{if } \chi(-1) = 1, \\
1 & \text{if } \chi(-1) = -1,
\end{cases}
\]
and let
\[
\Lambda(\chi, s) = q^{s/2} L_\infty(\chi, s) L(\chi, s), \quad L_\infty(\chi, s) = \pi^{-\frac{s}{2}} \Gamma\left(\frac{s + a}{2}\right)
\]
be the completed $L$–function. For $s \in \mathbb{C}$ one has
\[
\Lambda(\chi, s) = i^{-a} \frac{\tau(\chi)}{\sqrt{q}} \Lambda(\chi, 1 - s),
\]
where $\tau(\chi) = \sum_{x \mod q} \chi(x)e(x/q)$ is the associated Gauß sum. Let
\[
L(E \otimes \chi, s) = L(\chi, s)^2, \quad L_\infty(E \otimes \chi, s) = L_\infty(\chi, s)^2, \quad \text{and } \Lambda(E \otimes \chi, s) = \Lambda(\chi, s)^2.
\]
We deduce from the above functional equations that
\[
\Lambda(E \otimes \chi, s) = (-1)^a \frac{\tau(\chi)^2}{q} \Lambda(E \otimes \chi, 1 - s).
\]

We now describe the functional equation when $E$ is replaced by a holomorphic cuspidal Hecke eigenform $f$ for the group $\Gamma_0(1) = \text{SL}(2, \mathbb{Z})$ with even weight $k$. For $\Im z > 0$, the function $f$ has the Fourier expansion (at infinity)
\[
f(z) = \sum_{n \geq 1} \lambda_f(n) n^{k-1} e(nz),
\]
where the numbers $\lambda_f(n)$ are the normalized Hecke eigenvalues of $f$. These are real numbers bounded in absolute value by $d(n)$ [4, Théorème 8.2]. For a primitive Dirichlet character $\chi$ of prime modulus $q$, the sequence $(\lambda_f(n) \chi(n))_{n \geq 1}$ is associated to the Fourier expansion (at infinity) of a primitive Hecke cusp form with weight $k$ on the group $\Gamma_0(q^2)$ with nebentypus $\chi^2$ (see [17, Propositions 14.19 & 14.20], for instance). We denote by $f \otimes \chi$ this Hecke cusp form. We now consider its associated $L$-function, given by
\[
L(f \otimes \chi, s) := \sum_{n \geq 1} \frac{\lambda_f(n) \chi(n)}{n^s} = \prod_p \left(1 - \frac{\lambda_f(p) \chi(p)}{p^s} + \frac{\chi^2(p)}{p^{2s}}\right)^{-1}, \quad \Re s > 1.
\]
Then (see e.g. [17, Theorem 14.17, Proposition 14.20]) $L(f \otimes \chi, s)$ has an analytic continuation to $\mathbb{C}$ and satisfies the functional equation
\[
\Lambda(f \otimes \chi, s) = i_{k, \chi} \Lambda(f \otimes \chi, 1 - s),
\]
where
\[
\Lambda(f \otimes \chi, s) = q^s L_\infty(f \otimes \chi, s) L(f \otimes \chi, s), \quad L_\infty(f \otimes \chi, s) = (2\pi)^{-s} \Gamma\left(\frac{k - 1}{2} + s\right)
\]
and the root number $i_{k, \chi}$ is defined by $i_{k, \chi} = i^k \tau(\chi)^2 q^{-1}$. Consequently one has the following equations:

**Lemma 2.1.** Let $f, g$ be either cuspidal holomorphic Hecke eigenforms of weights $k_f, k_g$ and level $1$ or the Eisenstein series $E$. Then one has
\[
\Lambda(f \otimes \chi, s) \Lambda(g \otimes \chi, s) = \varepsilon(f, g, \chi) \Lambda(f \otimes \chi, 1 - s) \Lambda(g \otimes \chi, 1 - s),
\]
where $\varepsilon(f, g, \chi)$ is the root number of $f \otimes g$ with nebentypus $\chi$. When $f$ and $g$ are both Eisenstein series, $\varepsilon(f, g, \chi)$ is the root number of $E$ with nebentypus $\chi$. For Hecke eigenforms, one has $\varepsilon(f, g, \chi) = 1$ if $f \otimes g$ does not have a non trivial root number, and $\varepsilon(f, g, \chi) = -1$ if $f \otimes g$ has a non trivial root number.
where
\[
\varepsilon(f, g, \chi) = (-1)^{\frac{k_f + k_g}{2}} \text{ for } f \text{ and } g \text{ holomorphic},
\]
\[
\varepsilon(f, E, \chi) = (-1)^{a + \frac{k_f}{2}} \text{ for } f \text{ holomorphic and } E = E',
\]
\[
\varepsilon(E, E, \chi) = 1 \text{ for } f = g = E.
\]

**Remark 2.2.** Observe that the root number \(\varepsilon(f, g, \chi)\) depends on \(\chi\) at most through its parity \(\chi(-1)\) (and does not depend on \(\chi\) at all if \(f\) and \(g\) are both holomorphic or both Eisenstein).

Next, we state a standard approximate functional equation. Let \(f, g\) be either \(E\) or holomorphic cuspidal Hecke eigenforms of level 1 and (even) weights \(\lambda_f(n), \lambda_g(n)\) as in (1.1), (2.1) and Lemma 2.1 and let \(\chi\) be a primitive character modulo \(q\). Then we have (along similar lines to [17, Theorem 5.3]) the formula

\[
(2.2) \quad L(f \otimes \chi, 1/2)L(g \otimes \chi, 1/2) = \sum_{m,n \geq 1} \frac{\lambda_f(m)\lambda_g(n)}{(mn)^{1/2}} \chi(m)\overline{\chi}(n)V_{f,g,\pm 1} \left( \frac{mn}{q^2} \right)
+ \varepsilon(f, g, \chi) \sum_{m,n \geq 1} \frac{\lambda_f(n)\lambda_g(m)}{(mn)^{1/2}} \chi(m)\overline{\chi}(n)V_{f,g,\pm 1} \left( \frac{mn}{q^2} \right)
\]

where \(\pm 1 = \chi(-1)\) and

\[
(2.3) \quad V_{f,g,\pm 1}(x) = \frac{1}{2\pi i} \int_\gamma \frac{L_\infty(f \otimes \chi, 1/2 + s)L_\infty(g \otimes \chi, 1/2 + s)}{L_\infty(f \otimes \chi, 1/2)L_\infty(g \otimes \chi, 1/2)} x^{-s} ds.
\]

Note that this function depends on \(\chi\) at most through its parity \(\chi(-1)\), and does not depend on \(\chi\) at all if \(f\) and \(g\) are both holomorphic.

### 2.2. Voronoi summation and Bessel functions.

The next lemma is a version of the Voronoi formula.

**Lemma 2.3.** [Lemma 2.2] Let \(c\) be a positive integer and \(a\) an integer coprime to \(c\), and let \(W\) be a smooth function compactly supported in \([0, \infty)\). Then, we have:

1. For \(f\) a cuspidal holomorphic Hecke eigenform of level 1 and weight \(k\),
   \[
   \sum_{n \geq 1} \lambda_f(n)W(n)e\left(\frac{an}{c}\right) = \frac{1}{c} \sum_{n \geq 1} \lambda_f(n)\overline{W}(\frac{n}{c^2})e\left(-\frac{na}{c}\right);
   \]

2. For \(d\) the divisor function,
   \[
   \sum_{n \geq 1} d(n)W(n)e\left(\frac{an}{c}\right) = \frac{1}{c} \int_0^{+\infty} (\log x + 2\gamma - 2\log c)W(x)dx + \frac{1}{c} \sum_{n \geq 1} d(n)\overline{W}(\frac{n}{c^2})e\left(\mp\frac{an}{c}\right).
   \]

In the above formulas, \(\gamma\) is Euler’s constant and the transforms \(\overline{W}, \overline{W}_\pm : (0, \infty) \to \mathbb{C}\) of \(W\) are defined by

\[
\overline{W}(y) = 2\pi i^k \int_0^\infty W(u)J_{k-1}(4\pi \sqrt{uy})du
\]

and

\[
\begin{cases}
\overline{W}_+(y) = -2\pi \int_0^\infty W(u)Y_0(4\pi \sqrt{uy})du, \\
\overline{W}_-(y) = 4 \int_0^\infty W(u)K_0(4\pi \sqrt{uy})du.
\end{cases}
\]
analogous formulae for \( \tilde{\phi} \):

\[
K_0(x) \ll x^{-1/2}e^{-x}
\]

for \( x \geq 1 \). At one point we shall need the uniform bounds

\[
J_{it}(x) \ll e^{[t]/2(|t|+x)^{1/2}}, \quad t \in \mathbb{R}, x > 0 \quad \text{and} \quad J_k(x) \ll \min\left(k^{-1/3}, |x^2-k^2|^{-1/4}\right), \quad k > 0, x > 0.
\]

The first bound follows from the power series expansion \([14, 8.402]\) for \( x < t^{1/3} \) (say) and from the uniform expansion \([6, 7.13 \text{ formula (17)}]\) otherwise. The second bound follows also from the power series expansion for \( x < k^{1/3} \) and from Olver's uniform expansion \([22, (4.24)]\).

Integration by parts in combination with \([14, 8.472.3]\) shows the formula

\[
\int_0^\infty W(y)Y_j(4\pi \sqrt{yw+z})dy = \int_0^\infty \left(\frac{j}{4\pi \sqrt{yw+z}} W(y) - \frac{\sqrt{yw+z}}{2\pi w} W'(y)\right) Y_{j+1}(4\pi \sqrt{yw+z})dy
\]

for \( j \in \mathbb{N}_0 \) and any smooth compactly supported function \( W \). Analogous formulae hold for \( J \) and \( K \) in place of \( Y \). We have the well-known asymptotic formula \([14, 8.451.2]\)

\[
Y_0(x) = F_+(x)e^{ix} + F_-(x)e^{-ix} + O(x^{-A})
\]

for \( x \geq 1 \) with smooth, non-oscillating functions \( F_\pm(x) \) satisfying \( x^{3}F_\pm^{(j)}(x) \ll j^{-1/2} \).

Finally, we consider the decay properties of the Bessel transforms \( W, \tilde{W}_\pm \).

**Lemma 2.4.** Let \( W \) be a smooth function compactly supported in \([1/2, 2]\) and satisfying \([18]\). For \( M \geq 1 \) let \( W_M(x) = W(x/M) \). For any \( i, j \geq 0 \) and for all \( y > 0 \), we have

\[
y^j\tilde{W}_M^{(j)}(y) \ll i,j \quad M(1+My)^{j/2}(1+q^j(My)^{-1/2})^i,
\]

\[
y^j(W_M)^{(j)}(y) \ll i,j \quad M(1+My)^{j/2}(1+|\log M|y)(1+q^j(My)^{-1/2})^i.
\]

In particular, the functions \( \tilde{W}_M(y) \), \( (W_M)^\pm(y) \) decay rapidly when \( y \geq q^{2\nu}/M \).

**Proof.** We differentiate \( j \) times under the integral sign, followed by \( i \) applications of \([2.7]\) (or analogous formulae for \( K \) and \( J \)) with \( z = 0 \). Then we estimate trivially, using \( \mathcal{B}(x) = \frac{1}{2}(\pm \mathcal{B}_{\nu-1} - \mathcal{B}_{\nu+1}) \) for \( \mathcal{B} \in \{J, K, B\} \) and the simple bounds

\[
J_\nu(x) \ll \nu \begin{cases} 1, & x \geq 1 \\ x^{\nu} & x < 1 \end{cases}, \quad Y_\nu(x), K_\nu(x) \ll \nu \begin{cases} 1, & x \geq 1 \\ (1 + \log |x|)x^{\nu} & x < 1 \end{cases}.
\]

\[\Box\]

### 2.3. Kuznetsov formula and large sieve.

Next we prepare the scene for the Kuznetsov formula. We follow the notation of \([2]\). 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 \( 0 \leq j \leq 3 \):

\[
\hat{\phi}(k) = 4i^k \int_0^\infty \phi(x) J_{k-1}(x) \frac{dx}{x},
\]

\[
\tilde{\phi}(t) = 2\pi i \int_0^\infty \phi(x) \frac{J_{2it}(x) - J_{-2it}(x)}{\sinh(\pi t)} \frac{dx}{x},
\]

\[
\check{\phi}(t) = 8 \int_0^\infty \phi(x) \cosh(\pi t) K_{2it}(x) \frac{dx}{x}.
\]
We let \( \mathcal{B}_k \) be an orthonormal basis of the space of holomorphic cusp forms of level 1 and weight \( k \), and we write the Fourier expansion of \( f \in \mathcal{B}_k \) as
\[
f(z) = \sum_{n \geq 1} \varrho_f(n)(4\pi n)^{k/2}e(nz).
\]
Similarly, for Maass forms \( f \) of level 1 and spectral parameter \( t \) we write
\[
f(z) = \sum_{n \neq 0} \varrho_f(n)W_{0,it}(4\pi |n|y)e(nx),
\]
where \( W_{0,it}(y) = (y/\pi)^{1/2}K_{it}(y/2) \) is a Whittaker function. We fix an orthonormal basis \( \mathcal{B} \) of Hecke-Maass eigenforms. Finally, we write the Fourier expansion of the (unique) Eisenstein series \( E(z, s) \) of level 1 at \( s = 1/2 + it \) as
\[
E(z, 1/2 + it) = y^{1/2+it} + \varphi(1/2 + it)y^{1/2-it} + \sum_{n \neq 0} \varrho(n, t)W_{0,it}(4\pi |n|y)e(nx).
\]
Then the following spectral sum formula holds (see e.g. \cite{2} Theorem 2).

**Lemma 2.5** (Kuznetsov formula). Let \( \phi \) be as in the previous paragraph, and let \( a, b > 0 \) be integers. Then
\[
\sum_{c \geq 1} \frac{1}{c} S(a, b; c) \phi \left( \frac{4\pi \sqrt{ab}}{c} \right) = \sum_{k \geq 2} \sum_{f \in \mathcal{B}_k} \phi(k) \Gamma(k) \sqrt{ab} \varrho_f(a) \varrho_f(b) + \sum_{f \in \mathcal{B}} \phi(t_f) \frac{\sqrt{ab}}{\cosh(\pi t_f)} \varrho_f(a) \varrho_f(b) + \frac{1}{4\pi} \int_{-\infty}^{\infty} \phi(t) \frac{\sqrt{ab}}{\cosh(\pi t)} \varrho(a, t) \varrho(b, t) dt
\]
and
\[
\sum_{c \geq 1} \frac{1}{c} S(a, -b; c) \phi \left( \frac{4\pi \sqrt{ab}}{c} \right) = \sum_{f \in \mathcal{B}} \phi(t_f) \frac{\sqrt{ab}}{\cosh(\pi t_f)} \varrho_f(a) \varrho_f(-b) + \frac{1}{4\pi} \int_{-\infty}^{\infty} \phi(t) \frac{\sqrt{ab}}{\cosh(\pi t)} \varrho(a, t) \varrho(-b, t) dt.
\]

Often an application of the Kuznetsov formula is followed directly by an application of the large sieve inequalities of Deshouillers-Iwaniec \cite{5} Theorem 2).

**Lemma 2.6** (Spectral large sieve). Let \( T, M \geq 1 \), and let \( (a_m) \), \( M \leq m \leq 2M \), be a sequence of complex numbers. Then all three quantities
\[
\sum_{2 \leq k \leq T \atop k \text{ even}} \Gamma(k) \sum_{f \in \mathcal{B}_k} \left| \sum_m a_m \sqrt{m} \varrho_f(m) \right|^2, \sum_{f \in \mathcal{B}} \frac{1}{\cosh(\pi t_f)} \left| \sum_m a_m \sqrt{m} \varrho_f(\pm m) \right|^2, \int_{-T}^{T} \frac{1}{\cosh(\pi t)} \left| \sum_m a_m \sqrt{m} \varrho(\pm m, t) \right|^2 dt
\]
are bounded by
\[
M^\varepsilon (T^2 + M) \sum_m |a_m|^2.
\]

Finally we quote a special case of \cite{3} Theorem 8] which is an important variant of the preceding inequalities and responsible for making our results independent of the Ramanujan-Petersson conjecture. The main point is that we do not need to factor out the integer \( s \) at the cost of \( s^\theta \).
Lemma 2.7. Let $s \in \mathbb{N}$, $R, T \geq 1$, and let $\alpha(r)$, $R \leq r \leq 2R$, be any sequence of complex numbers with $|\alpha(r)| \leq 1$. Then

$$\sum_{f \in \mathcal{B}} \frac{1}{|t_f|} \left| \sum_{R \leq r \leq 2R} \alpha(r) \sqrt{rs} g_f(rs) \right|^2 \ll (sTR)^\varepsilon (T + s^{1/2})(T + R)R.$$ 

3. OUTLINE OF THE PROOF

To discuss the principle of the proof, it is useful to carry out the argument in a context as general as possible and to specialize only as it becomes necessary. Let $f$ and $g$ be Hecke eigenforms of level one. In this section, we will explain roughly how the proofs of the main results are reduced to estimates of shifted convolution sums and bilinear forms involving Kloosterman sums.

Let $q$ be a prime. The first critical feature is that the root number $\varepsilon(f, g, \chi) = \pm 1$ and the archimedean local factor $L_\infty(f \otimes \chi, s)L_\infty(g \otimes \chi, s)$ of the product of $L$-functions $L(f \otimes \chi, s)L(g \otimes \chi, s)$ both depend on the character $\chi$ only through its parity, i.e., through $\chi(-1) = \pm 1$. We therefore average separately over even or odd characters, and then these quantities are constant for all $\chi$ in the average.

To compute the central values $L(f \otimes \chi, 1/2)L(g \otimes \chi, 1/2)$, we use the approximate functional equation (2.2). This gives an expression of the form

$$L(f \otimes \chi, 1/2)L(g \otimes \chi, 1/2) = \sum_{m, n \geq 1} \sum_{\lambda_f(m), \lambda_g(n)} \frac{\lambda_f(m) \lambda_g(n) \chi(m) \chi(n)}{(mn)^{1/2}} V\left(\frac{mn}{q^2}\right)$$

for some essentially bounded test function $V$ decaying rapidly for $t \geq q^\varepsilon$ (for any fixed $\varepsilon > 0$), which depends only on $f, g$ and the parity of $\chi$. Using the orthogonality of characters with given parity (given by (1.8) below), an easy computation shows that the average is a simple combination of the quantities

$$B_{f, g}^\pm(q) = \sum_{m \equiv \pm n (\text{mod} q)} \frac{\lambda_f(m) \lambda_g(n)}{(mn)^{1/2}} V\left(\frac{mn}{q^2}\right) - \frac{1}{\varphi(q)} \sum_{(mn, q) = 1} \frac{\lambda_f(m) \lambda_g(n)}{(mn)^{1/2}} V\left(\frac{mn}{q^2}\right).$$

A simple main term arises from $B_{f, g}^+(q)$ for $m = n$. Putting this aside and applying a partition of unity reduces the problem to the evaluation of bilinear expressions of the type

$$B_{f, g}^\pm(M, N) = \frac{1}{(MN)^{1/2}} \sum_{m \equiv \pm n (\text{mod} q)} \lambda_f(m) \lambda_g(n) W_1\left(\frac{m}{M}\right) W_2\left(\frac{n}{N}\right)$$

$$- \frac{1}{q(MN)^{1/2}} \sum_{m, n} \lambda_f(m) \lambda_g(n) W_1\left(\frac{m}{M}\right) W_2\left(\frac{n}{N}\right),$$

where $M, N \geq 1$, $MN \leq q^{2 + o(1)}$ and $W_1, W_2$ are test functions satisfying (1.8).

At this point, an off-diagonal “main term” appears in the non-cuspidal case $f = g = E$. This is rather complicated, but extracting and estimating this term has been done by Young (see [24]). We denote by $ET_{f, g}(M, N)$ the remaining part of $B_{f, g}^\pm(M, N)$ in all cases (thus $ET_{f, g}(M, N) = B_{f, g}^\pm(M, N)$ unless $f = g = E$).
From this analysis, we know that an asymptotic evaluation of the twisted moment $M_{f,g}(q)$ with power-saving error term follows as soon as one proves that

$$\text{ET}_{f,g}(M, N) \ll q^{-\eta}$$

for some absolute constant $\eta > 0$. As we see quickly, we have a first bound

$$\text{ET}_{f,g}(M, N) \ll \varepsilon (M N)^{1/2} q^{1+\varepsilon},$$

which is however non-trivial in certain cases, since it depends on the assumption that $f$ and $g$ satisfy the Ramanujan-Petersson conjecture. More precisely, only the cusp form associated to the longer variable of $N$ and $M$ needs to satisfy the Ramanujan conjecture. We will return to this subtle point later, when we consider the “mixed moment”. The previous trivial bound means that we may assume that $MN$ is close to $q^2$ when estimating $\text{ET}_{f,g}(M, N)$. At this point, the analysis depends on the relative ranges of $M$ and $N$. We separate two cases.

First, when $M$ and $N$ are close to each other, we interpret the congruence condition as an equality over the integers:

$$0 \neq m \mp n \equiv 0 \pmod{q} \iff m \mp n = qr, \text{ for some } r \neq 0.$$

For each $r$, we face an instance of the shifted convolution problem for Hecke eigenvalues. This problem has now a long history in a variety of contexts (see [20] for an overview). The most powerful methods known today involve the spectral theory of automorphic forms, and therefore they usually depend on bounds towards the Ramanujan-Petersson conjecture.

In our situation, Young showed in [24, Theorem 3.3] that if $f = g = E$, we have the estimate

$$\text{ET}_{f,g}^\pm(M, N) \ll q^{-1/2+\theta+o(1)} \left(\frac{M}{N} + \frac{N}{M}\right)^{1/2}.$$  

for $MN \leq q^{2+o(1)}$. Recall that $\theta$ can, at the current time, be taken to be equal to $7/64$, but not unconditionally smaller. This result, which may be extended to general $f$ and $g$ by different techniques, is quite satisfactory when $M$ and $N$ are close in the logarithmic scale, but (as can be expected) it becomes weaker as $M, N$ get apart from each other. In particular, when $MN = q^{2+o(1)}$, the bound is only non-trivial outside the range

$$\max(M, N) \geq q^{3/2-\theta-\delta}$$

for any fixed $\delta > 0$. Clearly, it would be highly desirable to reduce as much as possible the influence of the parameter $\theta$. This removal of dependence on the Ramanujan conjecture is precisely one of the main achievements in [3], and the method in that paper leads to a bound similar to the above, unconditionally, essentially as if $\theta = 0$. Section 4 contains precise statements on shifted convolution sums and the corresponding estimates on $\text{ET}_{f,g}^\pm(M, N)$.

This means that we can handle $\text{ET}_{f,g}^\pm(M, N)$ as long as $MN = q^{2+o(1)}$, excluding the range

$$\max(M, N) \geq q^{3/2-\delta}$$

for any fixed $\delta > 0$. Up to this point, there are only minor differences (e.g., having to do with the main terms) between all cases of $f$ and $g$. We now explain the second ingredient used to cover this range, which eventually requires us to consider different cases separately.

We assume that $N = \max(M, N)$ is the longest variable, with $N \geq q^{3/2-\delta}$ for some fixed $\delta > 0$. Because it is a rather long variable, we may gain by applying the Voronoi summation formula (followed by a smooth partition of unity) to this variable. This leads to a decomposition of $\text{ET}_{f,g}^\pm(M, N)$ (up to possible main terms that are dealt with separately) into sums of the type

$$C^\pm(M, N') = \frac{1}{(qMN')^{1/2}} \sum_{m, n} \lambda_f(m) \lambda_g(n) K_{L2}(\pm mn; q) W_1 \left(\frac{m}{M}\right) \tilde{W}_2 \left(\frac{n}{N'}\right),$$

10
where the “dual” length $N'$ satisfies
\[ N' \leq N^* := q^2 / N, \]
and $\tilde{W}_2$ is another smooth function satisfying (1.8). (We recall that $\text{Kl}_2(a; q)$ is the Kloosterman sum modulo $q$, normalized so that $|\text{Kl}_2(a; q)| \leq 2$ is the Weil bound.) Thus the goal is now to prove that
\[ C^\pm(M, N') \ll q^{-\eta} \]
for some absolute constant $\eta > 0$. It turns out that the main difficulty is when $N' = N^* = q^2 / N$, which we now assume.

First, as above, we investigate the “trivial” bound, which in this case (from Weil’s bound) is
\[ C^\pm(M, N') \ll (MN^*/q)^{1/2} \ll q^{3/2 + o(1)}/N, \]
which is satisfactory as soon as $N \geq q^{3/2 + o(1)}$. Hence, we are left with a single critical range
\[ M = q^{1/2 + o(1)}, \quad N^* = q^{1/2 + o(1)}, \]
which is the Pólya-Vinogradov range for the bilinear form in Kloosterman sums
\[ \sum_{m, n} \lambda_f(m)\lambda_g(n)W_1 \left( \frac{m}{M} \right) \tilde{W}_2 \left( \frac{n}{N^*} \right) \text{Kl}_2(\pm mn; q). \]

For general $f$ and $g$ (namely, when $f$ and $g$ are both cuspidal), the range of the variables is so short that we don’t see any way to exploit the fact that $\lambda_f(m)$ and $\lambda_g(n)$ are Hecke eigenvalues. Thus we try to handle the sum as simply a special value of a general bilinear form
\[ \sum_{m, n} \alpha_m \beta_n \text{Kl}_2(\pm mn; q) \]
in Kloosterman sums. Based on earlier work of Fouvry and Michel [12], we prove, in Section 6, the following bound which is conditional on a suitable bound for a complete multivariable sum of products of Kloosterman sums.

**Proposition 3.1** (Bilinear forms of Kloosterman sums). Let $q$ be a prime, $(a, q) = 1$, $U, V \in [1, q]$, $V \subset \mathbb{R}$ an interval of length $V$ and $(\alpha_u)_u$, $(\beta_v)_v$ be two sequences of complex numbers supported respectively on $[1, U]$ and $V$ and with $\ell_2$-norms given by
\[ \|\alpha\|_2^2 = \sum_{u \leq U} |\alpha_u|^2, \quad \|\beta\|_2^2 = \sum_{v \in V} |\beta_v|^2. \]

Assuming that
\[ q^{1/4} \leq UV \leq q^{5/4} \quad \text{and} \quad U \leq q^{1/4}V, \]
and that Conjecture [6.5] holds, one has
\[ \sum_{u \leq U, v \in V} \alpha_u \beta_v \text{Kl}_2(auv; q) \ll (qUV)^{\varepsilon} \|\alpha\|_2 \|\beta\|_2 (UV)^{1/2} (U^{-1/4} + q^{11/64}(UV)^{-16/11}) \]
for any $\varepsilon > 0$, uniformly in $a$.

In the critical range $U \asymp V \asymp q^{1/2}$, the above bound saves a factor $q^{1/64}$ over the trivial bound. This, as we will see, leads to Theorem 1.3.

For some special coefficients $\alpha$ and $\beta$, we can handle (3.2) (without requiring Conjecture 6.5), and this will lead to our unconditional results.
First, if (say) \( g = E \) is the Eisenstein series, we exploit the decomposition of the Hecke eigenvalues 
\( d(n) = (1 \ast 1)(n) \) as a Dirichlet convolution. Inserting this expression, the bilinear form (3.2) 
transforms into trilinear forms with two smooth variables of the type 
\[
\sum_{m \geq M} \sum_{n_1 \geq N_1} \sum_{n_2 \geq N_2} \lambda_f(m) \text{Kl}_2(\pm mn_1n_2; q) \quad \text{with } N_1N_2 = N^*.
\]

Now we can group the variables differently to bring this into either beyond the Pólya-Vinogradov range (which can be handled in great generality, as shown in [10]), or into a situation corresponding to bilinear forms in the Pólya-Vinogradov range but with one smooth variable. We will show in Theorem 6.1 how to get an analogue of (3.1) in that case: this is the main new result of this paper concerning Kloosterman sums. In combination with arguments from [24] this will eventually lead to the exponent 1/68 in the error term of Theorem 1.2.

Finally, in the case \( f = g = E \) of Young’s Theorem, we may now decompose combinatorially both variables \( m \) and \( n \). Thus we reduce to quadrilinear forms
\[
\sum_{m_1, m_2, n_1, n_2} \ldots \sum_{m_i \equiv \pm n_i (\mod q)} \lambda_f(m_1m_2n_1n_2; q),
\]
where
\[
M \asymp q^{1/2+o(1)}, \quad N^* \asymp q^{1/2+o(1)}.
\]

We have now more possibilities for grouping variables. Especially when two of the variables (say \( M_1 \) and \( N_2 \)) are small, we can exploit a special case of some other general results of [10] that provide quite strong (unconditional) bounds for (3.2) when both variables are smooth.

The precise statements concerning Kloosterman sums are found in Section 6.1.

### 4. Shifted convolution sums

#### 4.1. Statements of results

We begin by stating the results that we use concerning the shifted convolution problem. We will then prove the new cases that we require.

For fixed modular forms \( f \) and \( g \) as in the introduction, for test functions \( W_1 \) and \( W_2 \) compactly supported in \([1/2, 2]\) and satisfying (1.8), and for \( M, N \geq 1 \), we denote
\[
\text{ET}^\pm_{f,g}(M, N) = \frac{1}{(MN)^{1/2}} \sum_{m \equiv \pm n (\mod q)} \lambda_f(m)\lambda_g(n)W_1(\frac{m}{M})W_2(\frac{n}{N}) - \frac{1}{q(MN)^{1/2}} \sum_{(mn, q) = 1} \lambda_f(m)\lambda_g(n)W_1(\frac{m}{M})W_2(\frac{n}{N}) - \delta_{f=g=E} \text{MT}^{\text{od,} \pm}_{E,E}(M, N),
\]
where \( \text{MT}^{\text{od,} \pm}_{E,E}(M, N) \) is the off-diagonal main term discussed by Young in [24], §6, §8. We have the following simple bound which follows from the Ramanujan-Petersson conjecture together with a bound for \( \text{MT}^{\text{od,} \pm}_{E,E}(M, N) \) given in [24] Lemma 6.1.

**Proposition 4.1.** Let \( f, g \) be either \( E \) or holomorphic cuspidal Hecke eigenforms of level 1, with weights \( k_f \equiv k_g (\mod 4) \) if both are cuspidal. Let \( q \) be a prime and assume that \( W_1, W_2 \) satisfy (1.8). For any \( \varepsilon > 0 \), we have
\[
\text{ET}^\pm_{f,g}(M, N) \ll \varepsilon \left( \frac{(MN)^{1/2}}{q} + \frac{\delta_{f=g=E}}{\min(M, N)} \left( \frac{\max(M, N)}{\min(M, N)} \right)^{1/2} \right) (MNq)\varepsilon.
\]

12
For later purposes we state the following variants: using only (1.6), but no individual bound on \( \lambda_f(n) \), we have

\[
ET_{f,E}^\pm(M,N) \ll \frac{(qMN)^\varepsilon}{(MN)^{1/2}} \left( \frac{N}{q} + 1 \right) M,
\]

and using the weak individual bound (1.7) we have

\[
ET_{f,E}^\pm(M,N) \ll \frac{M^{1/4}(M^N)^{1/2}}{q}(qMN)^\varepsilon.
\]

Our main result in this section is the following improvement.

**Proposition 4.2.** Let \( f,g \) be either \( E \) or holomorphic cuspidal Hecke eigenforms of level 1, with weights \( k_f \equiv k_g \pmod{4} \) if both are cuspidal; let \( q \) be a prime and assume that \( W_1, W_2 \) satisfy (1.8). For any \( \varepsilon > 0 \), there exists \( \varepsilon' > 0 \) such that for \( M,N \geq 1 \) and \( MN \leq q^{2+\varepsilon'} \), one has

\[
ET_{f,g}^\pm(M,N) \ll q^\varepsilon \left( \frac{1}{q} \left( \frac{M}{N} + \frac{N}{M} \right) + \frac{1}{q^{1/2}} \left( \frac{M}{N} + \frac{N}{M} \right)^{1/2} \right)^{1/2}
\]

**Remark 4.3.** It is a very pleasing feature that the same bound holds for cuspidal and non-cuspidal automorphic forms, even though the methods are – at least on the surface – rather different. We note that in the case \( f = g = E \) the bound (4.4) improves on [24, Theorem 3.3].

The remaining part of this section is devoted to the proof of Proposition 4.2.

### 4.2. Preliminaries.

We start with some general remarks. We denote

\[
S_{f,g}^\pm(M,N) := \frac{1}{(MN)^{1/2}} \sum_{\substack{m,n \pmod{q} \text{ (mod } q) \atop m \equiv \pm n}} \lambda_f(m)\lambda_g(n)W_1\left( \frac{m}{M} \right)W_2\left( \frac{n}{N} \right)
\]

We first observe that, by applying the Mellin inversion formula to \( W_1 \) and \( W_2 \) together with suitable contour shifts, we have

\[
\frac{1}{q(MN)^{1/2}} \sum_{m,n} \lambda_f(m)\lambda_g(n)W_1\left( \frac{m}{M} \right)W_2\left( \frac{n}{N} \right) = \frac{1}{q(MN)^{1/2}} \left( \text{res}_{s=1} L(f,s) \hat{W}_1(s)M^s + O_f(A(M^{-A})) \right) \left( \text{res}_{s=1} L(g,s) \hat{W}_2(s)N^s + O_g(A(N^{-A})) \right)
\]

for any \( A \geq 0 \).

If both \( f, g \) are cuspidal, or if \( f \) is cuspidal and \( M \geq q^\varepsilon \), this term is negligible. In particular, if both \( f \) and \( g \) are cuspidal, it is enough to obtain the stated bound for the quantity \( S_{f,g}^\pm(M,N) \) in place of \( ET_{f,g}^\pm(M,N) \). In addition, at the cost of an additional error \( O(q^{-1+\varepsilon}) \), which is admissible, it suffices to estimate

\[
\frac{1}{(MN)^{1/2}} \sum_{\substack{m \equiv \pm n \pmod{q} \atop m \neq n}} \lambda_f(m)\lambda_g(n)W_1\left( \frac{m}{M} \right)W_2\left( \frac{n}{N} \right).
\]

Then the required estimate for this last quantity is contained in [23, (3.4), (3.11)] if \( \max(N,M) \geq 20\min(N,M) \); in the complementary case \( N \asymp M \) it follows in a stronger form from the bound \( ET_{f,g}^\pm(M,N) \ll q^{1+\varepsilon}(N + M)^{1/2+\theta} \) contained in [1, Theorem 1.3]. This completes the proof of (4.4) in the case \( f, g \) cuspidal.

We prepare similarly for the proof of (4.4) in the case \( f \) cuspidal, \( g = E \) Eisenstein to which we devote the most work of the section. For \( M \leq q^{1/4} \) the right-hand side of (4.4) is larger than
$q^{1/8+\varepsilon}$, which dominates the bound (4.1). We may therefore assume that $M \geq q^{1/4}$, in which case it suffices (by (4.5) again) to estimate $S_{f,q}^\pm(M,N)$.

The following argument will feature a lot of separating variables by integral transforms, but this is only a technical necessity and has little to do with the core of the argument. In this context we will frequently use Lemma 1.6.

To begin with, we make no assumption about the size of $M,N,q$ and write $P := MNq$. For simplicity, we denote $\lambda(n) = \lambda_f(n)$. We open the divisor function, getting

\begin{equation}
(MN)^{1/2} S_{f,E}^\pm(M,N) = \sum_{r \neq 0} \sum_{a,b,m \geq 1 \atop m \equiv ab - rq} \lambda(m) W_1 \left( \frac{ab}{N} \right) W_2 \left( \frac{m}{M} \right).
\end{equation}

We localize the variable $a$ by attaching a weight function $W_3(a/A)$ where (by symmetry)

\begin{equation}
A \leq N^{1/2}
\end{equation}

and $W_3$ is a fixed smooth weight function with support in $[1/2,2]$. Hence it suffices to estimate

\begin{equation}
S(M,N,q,A) = \sum_{r \neq 0} \sum_{a} \sum_{m \equiv rq(a)} \lambda(m) W_2 \left( \frac{m}{M} \right) W_3 \left( \frac{a}{A} \right) W_1 \left( \pm \frac{m-rq}{A} \right).
\end{equation}

This expression is not symmetric in $M$ and $N$, and therefore we will now distinguish two cases according as whether $NP^\varepsilon \geq M$ or not (the reason why it is convenient to include a $P^\varepsilon$-power will be clear when we treat the second case.)

4.3. First case. We first assume that $NP^\varepsilon \geq M$. This condition implies $|rq| \leq N_0 := 4NP^\varepsilon$. We separate variables by Fourier inversion

\begin{equation}
S(M,N,q,A) = \int_{-\infty}^\infty W_1^\dagger(x) \sum_{1 \leq |r| \leq N_0/q} e \left( \pm \frac{r qx}{N} \right) \sum_{a} \sum_{m \equiv rq(a)} \lambda(m) W_2 \left( \frac{m}{M} \right) e \left( \pm \frac{mx}{N} \right) W_3 \left( \frac{a}{A} \right) dx
\end{equation}

where $W_1^\dagger$ denotes the Fourier transform. We can truncate the integral at $|x| \leq P^{2\varepsilon}$ at the cost of a negligible error. We write

$$V(z) = V_2(z) = W_2(z) e \left( \mp \frac{xM}{N} \right),$$

so that $V$ has compact support in $[1/2,2]$ and satisfies $V^{(j)} \ll P^{3j\varepsilon}$, uniformly in $|x| \leq P^{2\varepsilon}$, and it remains to estimate

\begin{equation}
S_x(M,N,q,A) = \sum_{1 \leq |r| \leq N_0/q} e \left( \pm \frac{r qx}{N} \right) \sum_{a} W_3 \left( \frac{a}{A} \right) \sum_{m \equiv rq(a)} \lambda(m) V \left( \frac{m}{M} \right).
\end{equation}

For later purposes, we also need to separate variables $r$ and $q$. Let $W_4$ be smooth with support in $[0,3]$ and constantly 1 on $[0,2]$, and write $V^*(y) = V_x^*(y) = W_4(y)e(\pm yxP^\varepsilon)$. Then by Mellin inversion we have

\begin{equation}
S_x(M,N,q,A) = \sum_{1 \leq |r| \leq N_0/q} V^* \left( \frac{|r|q}{NP^\varepsilon} \right) \sum_{a} W_3 \left( \frac{a}{A} \right) \sum_{m \equiv rq(a)} \lambda(m) V \left( \frac{m}{M} \right)
\end{equation}

\begin{equation}
= \int_{(\varepsilon)} \tilde{V}^* (u) \sum_{1 \leq |r| \leq N_0/q} \left( \frac{|r|q}{NP^\varepsilon} \right)^{-u} \sum_{a} W_3 \left( \frac{a}{A} \right) \sum_{m \equiv rq(a)} \lambda(m) V \left( \frac{m}{M} \right) \frac{du}{2\pi i}.
\end{equation}

We can truncate the $u$-integral at $|3mu| \leq P^4\varepsilon$, and hence it remains to estimate

\begin{equation}
S_u(M,N,q,A) = \sum_{1 \leq |r| \leq N_0/q} |r|^{-u} \sum_{a} W_3 \left( \frac{a}{A} \right) \sum_{m \equiv rq(a)} \lambda(m) V \left( \frac{m}{M} \right)
\end{equation}
uniformly in $\Re u = \varepsilon$ and $|\Im u| \leq P^{4\varepsilon}$. We detect the congruence with primitive additive characters modulo $d$ for $d \mid a$. By the Voronoi summation formula (Lemma 2.3), the $m$-sum equals

$$\sum_{d | a} \sum_{m} \lambda(m) S(rq, m; d) \tilde{V} \left( \frac{mM}{d^2} \right).$$

By Lemma 2.4 we see that $\tilde{V}$ is again a Schwartz class function satisfying

$$y^j \tilde{V}(y) \ll_k P^{4j\varepsilon} \left(1 + \frac{\sqrt{y}}{P^{3\varepsilon}}\right)^{-k}$$

for any $k \geq 0$. This gives

$$\tilde{S}_u(M, N, q, A) = \sum_{1 \leq |r| \leq N_0/q} \left| r \right|^{-u} \sum_{a} W_3 \left( \frac{a}{A} \right) \sum_{d | a} \sum_{m} \lambda(m) S(rq, m; d) \tilde{V} \left( \frac{mM}{d^2} \right)$$

$$= \frac{M}{A} \sum_{0 \neq |r| \leq N_0/q} \left| r \right|^{-u} \sum_{b} \sum_{d} \sum_{m} W_5 \left( \frac{db}{A} \right) \lambda(m) S(rq, m; d) \tilde{V} \left( \frac{mM}{d^2} \right)$$

where $W_5(z) = W_3(z)/z$. We localize $R \leq |r| \leq 2R$ and $M^* \leq m \leq 2M^*$ with

$$(4.12) \quad 1 \leq R \leq \frac{4NP^\varepsilon}{q}, \quad 1 \leq M^* \ll \frac{P^{4\varepsilon}A^2}{Mb^2},$$

up to a negligible error. Then we are left with

$$\tilde{S}_u(M, N, q, A, R, M^*) = \frac{M}{A} \sum_{b \leq PR \leq |r| \leq 2R} \left| r \right|^{-u} \sum_{M^* \leq m \leq 2M^*} \lambda(m) \sum_{d} \frac{1}{d} S(rq, m; d) \Omega \left( \frac{4\pi \sqrt{|r|qm}}{d} \right)$$

where

$$\Omega(z) = \Omega_{m,b,r}(z) = W_5 \left( \frac{4\pi \sqrt{|r|qm}}{zA} \right) \tilde{V} \left( \frac{z^2M}{(4\pi)^2|r|q} \right).$$

The support of $W_5$ restricts the support of $\Omega$ to

$$\frac{2\pi \sqrt{M^*Rqb}}{A} \leq z \leq \frac{16\pi \sqrt{M^*Rqb}}{A}.$$

Let $W_6$ be a smooth weight function that is constantly 1 on $[2\pi, 16\pi]$ and vanishes outside $[\pi, 17\pi]$. Then we have by double Mellin inversion

$$\Omega(z) = W_6 \left( \frac{zA}{\sqrt{M^*Rqb}} \right) W_5 \left( \frac{4\pi \sqrt{|r|qm}}{zA} \right) \tilde{V} \left( \frac{z^2M}{(4\pi)^2|r|q} \right)$$

$$= W_6 \left( \frac{zA}{\sqrt{M^*Rqb}} \right) \int_{(\varepsilon)} \int_{(\varepsilon)} \left( \frac{4\pi \sqrt{|r|qm}}{zA} \right)^{-s} \left( \frac{z^2M}{(4\pi)^2|r|q} \right)^{-t} \tilde{W}_5(s) \tilde{V}(t) \frac{dt \, ds}{(2\pi i)^2}$$

$$= \int_{(\varepsilon)} \int_{(\varepsilon)} \left( \frac{4\pi \sqrt{|r|}}{\sqrt{M^*R}} \right)^{-s} \left( \frac{MM^*Rb^2}{(4\pi A)^2|r|} \right)^{-t} \left( \frac{zA}{\sqrt{M^*Rqb}} \right)^{s-2t} \tilde{W}_5(s) \tilde{V}(t) W_6 \left( \frac{zA}{\sqrt{M^*Rqb}} \right) \frac{dt \, ds}{(2\pi i)^2}$$

The integrals can be truncated at $|\Im s|, |\Im t| \leq P^{4\varepsilon}$ at the cost of a negligible error. Writing

$$\Theta(z) = \Theta_{s,t}(z; b) = W_6 \left( \frac{zA}{\sqrt{M^*Rqb}} \right) \left( \frac{zA}{\sqrt{M^*Rqb}} \right)^{s-2t}$$
which depends on \( b \), but not on \( r \) or \( m \), and satisfies \( z^j \Theta^{(j)}(z) \ll j P^{12\varepsilon j} \), we are now left with bounding

\[
S_{u,s,t}(M, N, q, A, R, M^*) = \frac{M}{A} \sum_{b} \left| \sum_{R \leq |r| \leq 2R} |r|^{l-u} \sum_{M^* \leq m \leq 2M^*} \lambda(m) m^{-\frac{2}{3}} \Sigma(rq, m) \right|
\]

where

\[
\Sigma(rq, m) = \sum_d \frac{1}{d} S(rq, m; d) \Theta \left( \frac{4\pi \sqrt{|r|qm}}{d} \right)
\]

and \( \Re t = \Re u = \varepsilon \), \( \Re s = 0 \), \( |\Im t|, |\Im u|, |\Im s| \leq P^{\varepsilon} \). This is in a form to apply the Kuznetsov formula (Lemma 2.5). We treat in detail the case \( r > 0 \), the other case is essentially identical. We get

\[
\Sigma(rq, m) = \sum_{k \geq 2} \sum_{f \in \mathbb{F}_k} \hat{\Theta}(k) \Lambda(k) \sqrt{r q m} \varphi_f(rq) \varphi_f(m)
\]

\[
+ \sum_{f \in \mathbb{F}} \hat{\Theta}(t_f) \sqrt{r q m} \varphi_f(rq) \varphi_f(m) + \frac{1}{4\pi} \int_{-\infty}^{\infty} \hat{\Theta}(t) \sqrt{r q m} \varphi(t) \varphi(m) dt.
\]

By [2] Lemma 2.1] we have

\[
\hat{\Theta}(k) \ll \frac{P^{12\varepsilon}}{\mathcal{J}} \left( 1 + \frac{k}{P^{13\varepsilon}} \right)^{-B}, \quad \hat{\Theta}(t) \ll \frac{P^{12\varepsilon}}{\mathcal{J}} \left( 1 + \frac{|t|}{P^{13\varepsilon}} \right)^{-B},
\]

where

\[
\mathcal{J} = 1 + \frac{\sqrt{M^* R q b}}{A}
\]

(this uses, for simplicity, the Selberg eigenvalue conjecture, known in the present case of level 1.)

From now on, we use \( \varepsilon \)-convention. By the Cauchy-Schwarz inequality we find that the contribution of the holomorphic spectrum is at most

\[
\sum_{b \in P} \frac{P^\varepsilon M}{\sqrt{M^* R q b}} \left( \sum_{2 \leq k \leq P^\varepsilon \mathcal{J}} \Lambda(k) \sum_{f \in \mathbb{F}_k} \sum_{R \leq |r| \leq 2R} |r|^{l-u} \sqrt{r q \varphi_f(rq)} \right)^2 \left( \sum_{2 \leq k \leq P^\varepsilon \mathcal{J}} \Lambda(k) \sum_{f \in \mathbb{F}_k} \sum_{M^* \leq m \leq 2M^*} \lambda(m) m^{-\frac{2}{3}} \varphi_f(m) \right)^2.
\]

Using the Ramanujan conjecture for \( \sqrt{q} \varphi_f(q) \) and the spectral large sieve Lemma 2.6 (separately for odd and even forms), this is (recall (4.7), (4.12))

\[
\sum_{b \in P} \frac{P^\varepsilon M}{\sqrt{M^* R q b}} \left( \frac{RM^* q b^2}{A^2} + R \right)^{1/2} \left( \frac{RM^* q b^2}{A^2} + M^* \right)^{1/2}
\]

\[
\ll P^\varepsilon \sum_{b \in P} \frac{M}{\sqrt{q b}} \left( \frac{N}{M} + \frac{N}{q} \right)^{1/2} \left( \frac{N}{M} + \frac{A^2}{Mb^2} \right)^{1/2} \ll P^\varepsilon \left( \frac{N}{q^{1/2}} + \frac{N \sqrt{M}}{q} \right).
\]

The same argument works for the Eisenstein spectrum. For the Maaß spectrum, we need to argue differently in order to avoid the Ramanujan conjecture. Here we use Lemma 2.7 with \( s = q \).
(estimating trivially the terms with \((r, q) > 1\) that only occur in the case \(R \geq q\) to conclude that the total contribution of the Maaβ spectrum is

\[
\sum_{b \leq p} \frac{P^e M}{\sqrt{M^2 R q b}} \left[ \left( (\sigma^2 + R^2) R^2 \right)^{1/4} (\sigma^2 + q)^{1/4} + (\sigma^2 R^4)^{1/2} \right] \left( (\sigma^2 + M^*) M^* \right)^{1/2}
\]

\[
\ll P^e \sum_{b \leq p} \frac{M}{\sqrt{q b}} \left( \frac{N^2}{N M + q^2} \right)^{1/4} \left( \frac{N}{M} + q \right)^{1/4} \left( \frac{N + N^2}{M + M q b^2} \right)^{1/2}
\]

\[
\ll P^e \left( \frac{N}{q^{1/4}} + \frac{M^{1/4} N^{3/4}}{q^{1/4}} \right) \left( 1 + (MN)^{1/4} \right).
\]

Note that \((4.14)\) is larger than \((4.13)\) when \(M \ll NP^e\). This completes the analysis of the contribution of \(\Sigma(rq, m)\).

4.4 Second case. We now assume that \(NP^e \ll M\). We return to \((4.8)\) and begin with some preliminary transformations. We write

\[
V(z) = V_{rq}(z) := W_2 \left( \frac{N z + rq}{M} \right) = \int_{-\infty}^{\infty} W_2^1(x) e \left( \frac{N z + rq}{M} - x \right) dx.
\]

The integral can be truncated at \(|x| \leq P^e\) at the cost of a negligible error. Since \(W_2(m/M) = V((m - rq)/N)\), putting \(W_4(z) = W_1(z)e(\pm z N x/M)\), we get

\[
W_2 \left( \frac{m}{M} \right) W_1 \left( \pm \frac{m - rq}{N} \right) = V \left( \frac{m - rq}{N} \right) W_1 \left( \pm \frac{m - rq}{N} \right)
\]

\[
= \int_{-\infty}^{\infty} W_2^1(x)e \left( \frac{r qx}{M} \right) W_4 \left( \pm \frac{m - rq}{N} \right) dx.
\]

Note that \(W_4\) has support in \([1/2, 2]\) and satisfies \(W_4^{(j)}(x) \ll \varepsilon\) uniformly in \(|x| \leq P^e\). Hence we are left with

\[
S_x(M, N, q, A) = \sum_{r \approx M/q} e \left( \frac{r qx}{M} \right) \sum_{a} \sum_{m \equiv rq(a)} \lambda(m) W_3 \left( \frac{a}{A} \right) W_4 \left( \pm \frac{m - rq}{N} \right)
\]

where \(r \approx M/q\) is short for \(r \in [c_1 M/q, c_2 M/q]\) for suitable constants \(c_1, c_2\). As in \((4.9)\) – \((4.11)\) we may separate the variables \(r\) and \(q\), and need to bound

\[
\tilde{S}_u(M, N, q, A) = \sum_{r \approx M/q} r^{-u} \sum_{a} \sum_{m \equiv rq(a)} \lambda(m) W_3 \left( \frac{a}{A} \right) W_4 \left( \pm \frac{m - rq}{N} \right)
\]

with \(\Re u = \varepsilon, \ |\Im u| \leq P^e\). Again we detect the congruence with primitive additive characters modulo \(d\) for \(d \mid a\) and apply Voronoi summation (Lemma\([2,3]\)) to the \(m\)-sum getting

\[
(4.15)
\]

\[
\tilde{S}_u(M, N, q, A) = \sum_{r \approx M/q} r^{-u} \sum_{a} W_3 \left( \frac{a}{A} \right) \sum_{d \mid a} \sum_{\epsilon \in \{\pm\}} \sum_{m} \lambda(m) S(rq, \epsilon m; d) W_4 \left( \frac{mrq}{d^2}, \pm \frac{m N}{d^2} \right)
\]

where

\[
W_4^\pm(z, w) = \int_{-\infty}^{\infty} W_4(y) \mathcal{J}^\pm(4\pi \sqrt{z + wy}) dy, \quad 4|w| \leq z,
\]

with

\[
\mathcal{J}^+(x) = -2\pi Y_0(x), \quad \mathcal{J}^-(x) = 4K_0(x).
\]

Note that by our current size assumption \(NP^e \ll M\), the first argument in \(W_4^\pm(z, w)\) is substantially larger than the second. We follow the argument of \([3]\) Lemma 9 & Remark after Corollary 10].
As
\[ \frac{mrq}{d^2} - 2 \frac{mN}{d^2} \gg \frac{M - O(N)}{A^2} \gg \frac{M}{N} \gg P^\epsilon, \]
the case \( \epsilon = -1 \) contributes negligibly due to the rapid decay of the Bessel-K-function, cf. (2.4). Hence it suffices to consider only the case \( \epsilon = 1 \). For later purposes (see (4.17) below) it is convenient to insert into (4.15) a smooth, redundant weight function \( W_0(mrq/d^2, \pm mN/d^2) \) such that \( W_0(z, w) = 0 \) for \( z \leq 1 \) or \( 3|w| \geq z \), and \( W_0(z, w) = 1 \) for \( z \geq 2 \) and \( 4|w| \leq z \). We write \( W_0(z, w) = W_0(z, w)W_4^2(z, w) \), so that
\[ \widehat{S}_u(M, N, q, A) = \sum_{r \leq M/q} r^{-u} \sum_{a} W_3 \left( \frac{a}{A} \right) \sum_{d/a} \frac{N}{d} \sum_{m} \lambda(m)S(rq, m; d)W_5 \left( \frac{mrq}{d^2}, \pm \frac{mN}{d^2} \right) \]
up to a negligible error coming from \( \epsilon = -1 \). An integral transform similar to \( W_5(z, w) \) was analyzed in [3, Lemma 9]. Our general assumption in the forthcoming analysis is
\[ z \asymp z + wy \gg P^\epsilon. \]
Repeated application of the formula (2.7) yields the preliminary bound
\[ W_5(z, w) \ll_k \left( \frac{\sqrt{z}}{w} \right)^k \]
for any \( k \geq 0 \). In particular, up to a negligible error of \( O(P^{-k}) \), we can assume that
\[ \sqrt{z} \geq wP^{-\epsilon}. \]
In this range we use the asymptotic formula (2.8), so that
\[ W_5(z, w) = W_+(z, w)e(2\sqrt{z}) + W_-(z, w)e(-2\sqrt{z}) + O(P^{-k}) \]
where
\[ z^4|w|^j \frac{\partial}{\partial z} \frac{\partial^j}{\partial w^j} W_\pm(z, w) \ll P^{\epsilon(i+j)}z^{-1/4}. \]
It is now easy to see (cf. [3 Corollary 10]) that its double Mellin transform
\[ \widehat{W}_\pm(s, t) = \int_0^\infty \int_0^\infty W_\pm(z, \pm w)z^{s-1}w^{t-1}dz dw \]
is rapidly decaying on vertical lines (i.e. \( \ll_k \ell, \epsilon \) \( P^\epsilon |s|^{-k} |t|^{-\ell} \) for \( |s|, |t| \geq 1 \)) and absolutely convergent in \( \Re t > 0, \Re s + \Re t/2 < 1/4 \).
We can restrict \( m \) to a dyadic range \( M^* \leq m \leq 2M^* \), and (4.16) implies
\[ M^* \ll \frac{P^2 \epsilon MA^2}{(bN)^2}. \]
This leaves us with bounding
\[ \widehat{S}_u(M, N, q, A, M^*) \]
\[ = \sum_{r \leq M/q} r^{-u} \sum_{b \leq P} b^{-u} \sum_{d} W_3 \left( \frac{bd}{A} \right) N \sum_{M^* \leq m \leq 2M^*} \lambda(m)S(rq, m; d)e\left( \pm \frac{2\sqrt{mrq}}{d} \right) W_\pm \left( \frac{mrq}{d^2}, \pm \frac{mN}{d^2} \right) \]
\[ = \sum_{b \leq P} \frac{N}{b} \sum_{r \leq M/q} r^{-u} \sum_{M^* \leq m \leq 2M^*} \lambda(m) \sum_{d} \frac{1}{d} S(rq, m; d)\Omega \left( \frac{4\pi \sqrt{mrq}}{d} \right) \]
where
\[ \Omega(z) = W_3 \left( \frac{4\pi b \sqrt{mrq}}{A z} \right) \frac{z}{4\pi \sqrt{mrq}} W_\pm \left( \frac{z^2}{4\pi^2}, \frac{z^2}{4\pi^2} \right) \frac{N}{r^2} \exp(\pm iz) \]
\[ ^1 \text{This is another reason why we separate the two cases in the somewhat artificial way } NP^\epsilon \geq M \text{ and } NP^\epsilon \leq M. \]
has support contained in

\[ z \times Z := \frac{b \sqrt{M^* M}}{A} \gg 1. \]

Once again we add a redundant weight function \( W_6(z/Z) \) of compact support (to remember the original size condition) that is constantly 1 on a sufficiently large (fixed) interval, and we separate variables by Mellin inversion:

\[ \Omega(z) = W_6\left( \frac{z}{Z} \right) \exp(\pm iz) \]

\[ \times \int_{(0)} \int_{(0)} \int_{(1/4-\varepsilon)} \widehat{W}_3(v) \widehat{W}_\pm(s, t) \left( \frac{4\pi b \sqrt{mrq}}{Az} \right)^{-v} \frac{z}{4\pi \sqrt{mrq}} \left( \frac{z}{4\pi} \right)^{-2s - 2t} \left( \frac{N}{rq} \right)^{-t} \frac{ds \, dt \, dv}{2\pi i 2\pi i 2\pi i} \]

\[ = \int_{(0)} \int_{(0)} \int_{(1/4-\varepsilon)} \widehat{W}_3(v) \widehat{W}_\pm(s, t) W_6\left( \frac{z}{Z} \right) \exp(\pm iz) \left( \frac{b}{A} \right)^{-v} \times (M^*)^{-\varepsilon} \frac{M^*}{Z}^{1 - 2s - 2t} N^{-t} (\frac{m}{M^*})^{1 - \varepsilon} \frac{N}{rq}^{1 + \varepsilon} \left( \frac{z}{Z} \right)^{1 - 2s - 2t} \frac{ds \, dt \, dv}{2\pi i 2\pi i 2\pi i}. \]

We can truncate the integrals at \(|\text{Im } s|, |\text{Im } t|, |\text{Im } v| \leq P^{2\varepsilon}\) at the cost a negligible error. Hence we need to bound

\[ S_{u, s, t, v}(M, N, q, A, M^*) = \sum_{b \leq P} \frac{NZ^{1/2}}{b \sqrt{M^* M}} \left| \sum_{r \sim M/q} r^{-\alpha} \sum_{M^* \leq m \leq 2M^*} \lambda(m) \left( \frac{m}{M^*} \right)^{-\varepsilon} \times \sum_{d} \frac{1}{d} S(rq, m; d) \Theta \left( \frac{4\pi \sqrt{mrq}}{d} \right) \right| \]

where \( \alpha = \varepsilon - t + u \) and

\[ \Theta(z) = \Theta_{s, t}(z) = W_6\left( \frac{z}{Z} \right) \exp(\pm iz) \left( \frac{z}{Z} \right)^{1 - s - t} \]

We apply the Kuznetsov formula (Lemma 2.5) to the \( d \)-sum. By [3, Lemma 8], the spectral sum can be truncated (with a negligible error) at spectral parameter \( P^{3\varepsilon} Z^{1/2} \), and we obtain

\[ S_{u, s, t, v}(M, N, q, A, M^*) = \sum_{b \leq P} \frac{NZ^{1/2}}{b \sqrt{M^* M}} \left| \sum_{r \sim M/q} r^{-\alpha} \sum_{M^* \leq m \leq 2M^*} \lambda(m) \left( \frac{m}{M^*} \right)^{-\varepsilon} \left( \mathcal{H} + \mathcal{M} + \mathcal{E} \right) \right| \]

(up to a negligible error) where

\[ \mathcal{H} = \sum_{2 \leq k \ll P^{3\varepsilon} Z^{1/2}} \sum_{f \ll k} 4i^k \Gamma(k) \int_{\text{even}} \int_0^\infty J_{k-1}(z) \Theta(z) \frac{dz}{z} \sqrt{mrq} \phi_f(m) \phi_f(rq), \]

\[ \mathcal{M} = \sum_{f \ll P^{3\varepsilon} Z^{1/2}} 2\pi i \int_0^\infty \frac{J_{2it_f}(z) - J_{-2it_f}(z)}{\sinh(\pi t_f)} \Theta(z) \frac{dz}{z} \sqrt{mrq} \phi_f(m) \phi_f(rq) \cosh(\pi t_f), \]

\[ \mathcal{E} = \int_{|t| \ll P^{3\varepsilon} Z^{1/2}} \frac{i}{2} \int_0^\infty \frac{J_{2it}(z) - J_{-2it}(z)}{\sinh(\pi t)} \Theta(z) \frac{dz}{z} \sqrt{mrq} \phi(m, t) \phi(rq, t) \cosh(\pi t) \]
(Indeed, if \( k \asymp Z \), then \( Z \ll P^{6e} \) and \( k^{-1/3} \asymp Z^{1/6} Z^{-1/2} \ll P^e Z^{-1/2} \).) We estimate the \( z \)-integral trivially. From now on we use \( \varepsilon \)-convention. The Maaß contribution is at most

\[
\sum_{b \leq P} \frac{P^e N}{b \sqrt{M!*M}} \sum_{t_f \leq P^e Z^{1/2}} \left| \sum_{r \asymp M/q} r^{-\alpha} \sum_{M^* \leq m \leq 2M^*} \lambda(m) \left( \frac{m}{M^*} \right)^{-\frac{1}{2}} \sqrt{m \varrho_f(m) \varrho_f(q)} \cosh(\pi t_f) \right|
\]

and similar expressions hold for the holomorphic and Eisenstein contribution. By the Cauchy-Schwarz inequality this is at most

\[
\sum_{b \leq P} \frac{P^e N}{b \sqrt{M!*M}} \left( \sum_{t_f \leq P^e Z^{1/2}} \left| \sum_{r \asymp M/q} r^{-\alpha} \sqrt{q \varrho_f(q)} \cosh(\pi t_f) \right|^2 \right)^{1/2}
\times \left( \sum_{t_f \leq P^e Z^{1/2}} \sum_{M^* \leq m \leq 2M^*} \lambda(m) \left( \frac{m}{M^*} \right)^{-\frac{1}{2}} \sqrt{m \varrho_f(m) \cosh(\pi t_f)} \right)^{1/2}
\]

Using Lemmas 2.6 -- 2.7 as in the previous case, this is

\[
\ll \sum_{b \leq P} \frac{P^e N}{b \sqrt{M!*M}} \left[ (Z + q)^{1/4} \left( (Z + \frac{M^2}{q^2}) \left( \frac{M^2}{q} \right)^{1/4} \right) \left( \frac{Z + M}{q} \right)^{1/2} \right] \left( \frac{Z + M^*}{M^*} \right)^{1/2}
\]

\[
\ll \sum_{b \leq P} \frac{P^e N}{b \sqrt{M}} \left[ \left( \frac{M}{N} + q \right)^{1/4} \left( \left( \frac{M}{N} + \frac{M^2}{q^2} \right) \left( \frac{M^2}{q^2} \right)^{1/4} \right) + \left( \frac{M}{N} + \frac{M}{q} \right)^{1/2} \right] \left( \frac{M}{N} + \frac{M^2}{(bN)^2} \right)^{1/2}
\]

By (4.7) this

\[
\ll P^e N^{1/2} \left[ \left( \frac{M}{N^{1/2} q^{1/2}} + \frac{M^{5/4}}{qN^{1/4}} + \frac{M^{3/4}}{N^{1/4} q^{1/4}} + \frac{M^{3/4}}{q^{3/4}} \right) \left( 1 + \left( \frac{MN}{q^{1/2}} \right)^{1/4} \right) \right]
\]

(4.18)

Combining (4.14) and (4.18), and recalling the extra factor \((MN)^{1/2}\) in (4.6), we complete the proof of (4.4) in the case \( f \) cuspidal, \( g = E \) Eisenstein.

4.5. The case \( f = g = E \). Here we merely indicate the points where the proof of [24, Thm. 3.3] needs some modifications. We will freely borrow the notations of that paper, where the error term \( E_{E,E}(M,N) \) is denoted by \( E_{M,N} \).

In [24 §9], this term is further decomposed as a sum of two terms \( E_{M,N} = E_+ + E_- \) and each of these two terms is decomposed into cuspidal holomorphic, cuspidal non-holomorphic and Eisenstein contributions denoted by \( E_{h \pm} + E_{m \pm} + E_{c \pm} \) in that paper. The term \( E_{m-} \) is the most complicated one, and it is here that we insert some modifications. This term decomposes further as a sum of terms denoted by \( \tilde{E}_K \) where \( K \) is a parameter around which the Laplace eigenvalues of the Maaß forms are localized.

20
In [24 (9.12)], we apply Hölder’s inequality with exponents \((1/4, 1/4, 1/2)\), and obtain
\[
\left| \sum_{K \leq \kappa < 2K} \frac{|g_j(1)|^2}{\cosh(\pi \kappa_j)} \lambda_j(q)L_j(1/2 + s_1)L_j(1/2 + s_2)^2 \right| \leq \left| \sum_{K \leq \kappa < 2K} \frac{|g_j(1)|^2}{\cosh(\pi \kappa_j)} |\lambda_j(q)|^4 \right|^{1/4}
\]
\[
\cdot \left| \sum_{K \leq \kappa < 2K} \frac{|g_j(1)|^2}{\cosh(\pi \kappa_j)} |L_j(1/2 + s_1)|^{4/4} \right| \cdot \left| \sum_{K \leq \kappa < 2K} \frac{|g_j(1)|^2}{\cosh(\pi \kappa_j)} |L_j(1/2)|^{4/2} \right|^{1/2}.
\]

Exactly as in [24 (9.13)], we bound the last two factors by \(K^{3/2+\varepsilon}\). For the first factor we write
\[
|\lambda_j(q)|^4 \leq 2(|\lambda_j(1)| + |\lambda_j(q^2)|^2)
\]
and use [21 Lemma 2.4] to estimate this factor by \((qK)^\varepsilon(K^2 + q)^{1/4}\). In this way, [24 (9.15)] becomes
\[
E_{m \pm} \ll q^{-1/2+\varepsilon} \left( \frac{N}{M} \right)^{1/2} + q^{-1/4+\varepsilon} \left( \frac{N}{M} \right)^{1/4}.
\]

Then [24 Proposition 9.3] and the last bound of [24 Section 9.6] give two additional error terms
\[
q^{-1/2+\varepsilon} \left( \frac{N}{M} \right)^{1/4} + q^{-1/2+\theta+\varepsilon} \left( \frac{M}{N} \right)^{1/2}
\]
both of which are dominated by (4.19), at least for \(\theta < 1/4\) and under the general assumption \(N \geq M\). Hence we get an improved version of [24 Theorem 3.4]:
\[
(4.20) \quad ET_{E, E}(M, N) \ll q^{-1/2+\varepsilon} \left( \frac{N}{M} \right)^{1/2} + q^{-1/4+\varepsilon} \left( \frac{N}{M} \right)^{1/4}
\]
under the assumption \(MN \ll q^{2+\varepsilon}\) and \(N \geq M\). \(\square\)

**Remark 4.4.** Inserting (4.20) into the subsequent analysis of the piecewise linear function at the end of Section 3 in [24], we obtain the exponent \(-1/82\) in the error term of [24 Theorem 1] (with no \(\theta\)-dependence), the maximum being taken at \(a = 21/41\) and \(b = 60/41\).

5. **Young’s Method**

In this section we prove the following small variation of [24 Lemma 4.1, 4.2].

**Proposition 5.1.** Let \(\lambda(m)\) denote either the (normalized) Fourier coefficients of a holomorphic cuspidal Hecke eigenform \(f\) of level 1, or the divisor function. Let \(N, N_1, N_2, M \geq 1\) be parameters with \(N_1 N_2 = N\), \(N_1 \leq N_2\), and let \(q \in \mathbb{N}\). Let \(W_1, W_2, W_3\) be smooth compactly supported weight functions satisfying (1.8). Then
\[
\frac{1}{\sqrt{MN}} \sum_{n_1 n_2 = \pm m, q} \lambda(m) W_1(n_1/N_1) W_2(n_2/N_2) W_3(m/M)
\]
is bounded by the following two quantities:
\[
(MNq)^\varepsilon \cdot \min \left\{ \frac{\sqrt{MN}}{q^{1/4}}, \frac{\sqrt{MN}}{q^{1/4}}, \frac{\sqrt{MN}}{q^{1/4}}, \frac{\sqrt{MN}}{q^{1/4}} \right\}
\]
\[
\left( \frac{M^{1/2} N^{1/2}}{q^{1/2}} + \frac{N_1 M^{1/2}}{q^{1/2}} + \frac{N_1 M^{1/2}}{q^{1/2}} + \frac{N_1 M^{1/2}}{q^{1/2}} \right),
\]
\[
\left( \frac{N_2^{1/2} N_2^{1/2}}{q^{1/2}} + \frac{N_2^{1/2} N_2^{1/2}}{q^{1/2}} + \frac{N_2^{1/2} N_2^{1/2}}{q^{1/2}} \right).
\]

**Remark 5.2.** The proof uses only the “soft” bounds (1.6) and (1.7) for \(\lambda\) in the cuspidal case, not Deligne’s bound (1.5). In particular Proposition 5.1 holds also if \(\lambda(m)\) are Fourier coefficients of Hecke-Maaß forms (cf. Remark 1.5 (3)).
Remark 5.3. As will become clear from the proof, the starting point is to apply Poisson summation in \( n_2 \). This is a very different strategy compared to the outline in Section 3 which dualizes the variables \( n_1, n_2 \) simultaneously in the form of Voronoi summation.

Proof. We follow closely the argument in [24, Lemma 4.1, 4.2] and keep track of the following two differences. We drop the assumption \( MN \leq q^{2+\varepsilon} \) and we allow that \( \lambda \) can be the divisor function or the sequence of Hecke eigenvalues (and we make sure to use only bounds of the type (1.6) and (1.7)).

The first bound is the analogue of [24, Lemma 4.1]. We can exclude the terms \( n_1 n_2 \equiv 0 \pmod{q} \) at the cost of an error \( (MNq)^{\varepsilon} (MN)q^{-7/4} \). We apply Poisson summation to the \( n_2 \)-sum. The central term contributes an error of \( O((\sqrt{MN}/q)^2) \), and we bound the quantity \( R \) in [24, (4.6)] by

\[
R \ll (MNq)^{\varepsilon} \frac{N_2}{q \sqrt{MN}} S(N_1, Mq/N_2, q)
\]

where

\[
S(K, L, q) = \sum_{l \leq L} |(\lambda * 1)(l)| : \sum_{(k, q) = 1} e\left(\frac{\overline{k}}{q}\right) W\left(\frac{k}{K}\right)
\]

is analogous to [24, (4.7)]. The proof of [24, Proposition 4.3] provides bounds for \( S(K, L, q) \) in the situation where \( \lambda \) is the divisor function and under the additional assumption \( L, K \ll q^{1+\varepsilon} \). In order to also include Fourier coefficients, we notice that the proof of [24, Proposition 4.3] uses only \( \infty \)-norms or 2-norms for the \( k \)-sum, so a Rankin-Selberg-type bound for \( \lambda * 1 \) suffices. Without the condition \( L, K \ll q^{1+\varepsilon} \) we obtain

\[
S(K, L, q) \ll (KLq)^{\varepsilon} \min\left((Lq^{1/2} + KL/q, (Lq^{3/2} + L^2K + L^2K^2/q)^{1/2}, LK\right)
\]

where the first bound is the analogue of [24, (4.11)], the second bound is the analogue of the last display in [24, Section 4.2] and the last bound is the trivial bound. In this way we arrive at the first bound of our proposition.

The second bound is the analogue of [24, Lemma 4.2], and again we only indicate the changes in Young’s proof. The error term in the second display of [24, Section 4.3] is (recall Young’s notation \( H = q/N_2 \))

\[
(MNq)^{\varepsilon} \left(\frac{MN_1}{q \sqrt{MN}} + \frac{M^2}{N_2 \sqrt{MN}}\right)
\]

If \( \lambda = d \) is the divisor function, then the pole in [24, (4.13)] contributes

\[
\ll (MNq)^{\varepsilon} \frac{M}{\sqrt{MN}}
\]

In either case, after shifting the contours, we apply Voronoi summation to the term \( U(h, m, n_1) \) in the last line on [24, p. 22] and arrive at a quantity analogous to [24, (4.15)]. Finally, the pointwise bound on the quantity \( V(h, k) \) defined under [24, (4.17)] and proved above [24, (4.18)] allows us to reach the analogue of [24, (4.16)–(4.17)], which yields the second bound of our proposition (and we notice that the assumption \( N \gg q^{1+\varepsilon} \) in [24, p. 24, line 8] can be assumed in our case, too, since otherwise the term \( (M/N)^{1/2} \) is worse than the trivial bound).

For later purposes, we will also need the following immediate corollary, which we state here for easy reference.

Corollary 5.4. Let \( \lambda(m) \) denote either the (normalized) Fourier coefficients of a holomorphic cuspidal Hecke eigenform \( f \) of level 1, or the divisor function. Let

\[
1 \leq N' \leq N^*, \quad N_1 N_2 = N', \quad N_2 \geq N_1 \geq 1, \quad 1 \leq M' \leq M^*
\]

For later purposes, we will also need the following immediate corollary, which we state here for easy reference.
be parameters and let \( q \in \mathbb{N} \). Let \( W_1, W_2, W_3 \) be smooth compactly supported weight functions satisfying (1.8). Then

\[
\frac{1}{\sqrt{M^*N^*}} \sum_{n_1n_2 \equiv \pm m(q)} \lambda(m)W_1(n_1/N_1)W_2(n_2/N_2)W_3(m/M')
\]

is bounded by the following two quantities:

\[
(M^*N^*q)^\varepsilon \left\{ \frac{\sqrt{M^*N^*}}{q^{1/4}} + \min \left( \frac{M'q^{1/2}}{\sqrt{M^*N^*}} + \frac{N_1M'}{q\sqrt{M^*N^*}}, \frac{\sqrt{M^*N^*}}{q^{1/4}} + \frac{M'N^{1/2}}{\sqrt{M^*N^*}} + \frac{N_1M'}{q\sqrt{M^*N^*}}, \frac{M'N_1^{1/2}}{\sqrt{M^*N^*}} \right) + \frac{N_1M'}{q\sqrt{M^*N^*}} + \frac{M'N^{1/2}}{\sqrt{M^*N^*}} + \frac{N_1M'}{q\sqrt{M^*N^*}} + \sqrt{(M')^2} \right\},
\]

6. Bilinear forms in Kloosterman sums

6.1. Statements of results. We begin by stating the precise results we obtain concerning the bilinear forms of type (3.2), i.e.

\[
\sum_m \sum_n \alpha_m \beta_n \text{Kl}_2(ann; q),
\]

where the normalized Kloosterman sum is defined in (1.9).

We recall from Section 3 that we will be especially interested in cases where \( m \) and \( n \) range over intervals of size close to \( q^{1/2} \). Our results in this section go in the direction of the conditional estimate in Proposition 3.1 being of similar (or better) quality for special coefficients \((\alpha_m)\) and \((\beta_n)\). Proposition 3.1 itself will be proved in Subsection 6.5 depending on a conjecture on certain complete sums over finite fields.

**Theorem 6.1.** Let \( q \) a prime number, let \( a \) be an integer coprime with \( q \), \( M, N \geq 1 \), \( N \) an interval of length \( N \), and \((\alpha_m)_m, (\beta_n)_n \) two sequences supported respectively on \([1, M]\) and \( N \).

1. If \( M, N \leq q \), we have

\[
\sum_{m \leq M, n \in N} \alpha_m \beta_n \text{Kl}_2(ann; q) \ll (qMN)^\varepsilon (MN)^{1/2} ||\alpha||_2 ||\beta||_2 (M^{-1/2} + q^{1/4}N^{-1/2})
\]

for any \( \varepsilon > 0 \).

2. If the conditions

\[
M, N \leq q, \quad MN \leq q^{3/2}, \quad M \leq N^2,
\]

are satisfied, then we have

\[
\sum_{m \leq M, n \in N} \alpha_m \text{Kl}_2(ann; q) \ll (qMN)^\varepsilon (||\alpha||_1 ||\alpha||_2)^{1/2} M^{1/4}N^{1/4} (q^{1/2}M^{-1/8}N^{-3/16}).
\]

3. Let \( W_i \), for \( 1 \leq i \leq 2 \), be smooth, compactly supported functions satisfying

\[
W_i^{(j)}(x) \ll_j Q^j, \quad i = 1, 2
\]

for some \( Q \geq 1 \) and for all \( j \geq 0 \). There exists an absolute constant \( A \geq 0 \) such that for any \( \varepsilon > 0 \), we have

\[
\sum_m \sum_n \frac{W_1\left(\frac{m}{M}\right)W_2\left(\frac{n}{N}\right)\text{Kl}_2(ann; q) \ll \varepsilon q^\varepsilon Q^A MN \left(\frac{1}{q^{1/8}} + \frac{q^{3/8}}{(MN)^{1/2}}\right)}.
\]

All bounds are uniform in \( a \), and we write as usual

\[
||\alpha||_1 = \sum_m |\alpha_m|, \quad ||\alpha||_2 = \left(\sum_m |\alpha_m|^2\right)^{1/2}.
\]
The first part (6.1) and the third (6.5) have been proven by Fouvry, Kowalski and Michel in Theorems 1.17 and 1.16 (respectively) of [10] (building on [9]), in considerably greater generality. Thus it remains to prove the second part.

6.2. General setup. Some of our arguments are valid for bilinear forms involving a more general kernel $K(mn)$ modulo $q$ than the Kloosterman sums $K_l(arm; q)$. It is therefore useful to consider first the general problem of bounding a general “type I” sum

$$S(K, M, N; q) := \sum_{m \leq M} \alpha_m \sum_{n \in \mathbb{N}} K(mn),$$

where $K : \mathbb{F}_q \rightarrow \mathbb{C}$ is an arbitrary function. We assume that $|K(x)| \leq 1$ with an absolute implied constant.

The proof is a slight generalization of the method of [12]: given $A, B \geq 1$ such that

$$AB \leq N, \quad AM < q,$$

we have

$$S(K, M, N; q) = \frac{1}{AB} \sum_{A < a \leq 2A} \sum_{m \leq M} \alpha_m \sum_{n+ab \in \mathbb{N}} K(m(n + ab))$$

$$= \frac{1}{AB} \sum_{A < a \leq 2A} \sum_{m \leq M} \alpha_m \sum_{n+ab \in \mathbb{N}} K(am(\overline{a}n + b)),$$

where, as usual, $a\overline{a} \equiv 1 \pmod{q}$. By the method of [12, p. 116], we get

$$S(K, M, N; q) \ll_{H, \varepsilon} \frac{q^\varepsilon}{AB} \sum_{(r \pmod{q}, s \leq AM)} \nu(r, s) \left| \sum_{B < b \leq 2B} \eta_b K(s(r + b)) \right|,$$

where

$$\nu(r, s) = \sum_{A < a \leq 2A, m \leq M, n \in \mathbb{N}, \ am = s, \ \overline{a}n \equiv r \pmod{q}} |\alpha_m|,$$

and $(\eta_b)_{B < b \leq 2B}$ are some complex numbers such that $|\eta_b| \leq 1$. We have the bounds

$$\sum_{r, s} \nu(r, s) \ll AN \sum_{m \leq M} |\alpha_m|$$

and

$$\sum_{r, s} \nu(r, s)^2 = \sum_{a, m, n, a', m', n'} |\alpha_m||\alpha_{m'}| \ll \sum_{a, m} |\alpha_m|^2 \sum_{a', m', n'} \sum_{a'' = \overline{a}'n} 1 \ll q^\varepsilon AN \sum_{m} |\alpha_m|^2.$$

Here, we have used the inequality $|\alpha_m||\alpha_{m'}| \leq |\alpha_m|^2 + |\alpha_{m'}|^2$ and the fact that, once $a$ and $m$ are given, the equation $am = a'm'$ determines $a'$ and $m'$ up to $O(q^\varepsilon)$ possibilities, and, for each such $a'$, $m'$ and each $n \in \mathbb{N}$, the congruence $a'n = an' \pmod{q}$ determines $n'$ uniquely since $n'$ varies over an interval of length $\leq q$ (cf. [12, p. 116]).

From these bounds and from Hölder’s inequality, we obtain that

$$(6.8) \ A B \cdot S(K, M, N; q) \ll q^\varepsilon (AN)^{3/4} (||\alpha||_1||\alpha||_2)^{1/2} \left( \sum_{r \pmod{q}, 1 \leq s \leq AM} \sum_{B < b \leq 2B} \eta_b K(s(r + b)) \right)^{1/4}.$$

24
Expanding the fourth power, the inner term of the second factor can be written as
\[
\sum_{b \in \mathcal{B}} \eta(b) \Sigma(K, b; q),
\]
where \( \mathcal{B} \) denotes the set of quadruples \( b = (b_1, b_2, b_1', b_2') \) of integers satisfying \( B < b_i, b_i' \leq 2B \) \( (i = 1, 2) \), the coefficients \( \eta(b) \) satisfy \( |\eta(b)| = 1 \) for all \( b \in \mathcal{B} \), and we denote
\[
(6.9) \quad \Sigma(K, b; q) = \sum_{r \pmod{q}} \sum_{1 \leq s \leq AM} K(s(r + b_1))K(s(r + b_2))\overline{K(s(r + b_1'))K(s(r + b_2'))}.
\]

Let \( \mathcal{B}^\Delta \) be the subset of \( \mathcal{B} \) admitting a subset of two entries matching the entries of the complement (for instance such that \( b_1 = b_1' \) and \( b_2 = b_2' \)); one has \( |\mathcal{B}^\Delta| = O(B^2) \). For such \( b \), we use the trivial estimate \( \Sigma(K, b; q) \), getting
\[
(6.10) \quad \sum_{b \in \mathcal{B}^\Delta} |\Sigma(K, b; q)| \ll AB^2Mq,
\]
where the implied constant depends only on \( H \).

To bound the contribution of the \( b \notin \mathcal{B}^\Delta \), we complete the \( s \)-sum using additive characters and obtain
\[
\Sigma(K, b; q) \ll (\log q) \max_{h \pmod{q}} |\Sigma(K, b, h; q)|,
\]
where
\[
(6.11) \quad \Sigma(K, b, h; q) := \sum_{r,s \pmod{q}} K(s(r + b_1))K(s(r + b_2))\overline{K(s(r + b_1'))K(s(r + b_2'))}e_q(hs).
\]

The procedure we have described gives a general scheme for estimating special bilinear forms \((6.6)\) with a general uniformly bounded kernel \( K \): the sum \( S(K, M, N; q) \) is estimated as in \((6.8)\), where the contribution of the diagonal quadruples \( b \in \mathcal{B}^\Delta \) is bounded in \((6.10)\), and the contributions of off-diagonal quadruples \( b \notin \mathcal{B}^\Delta \) are estimated in terms of the complete sums \( \Sigma(K, b, h; q) \) given by \((6.11)\).

If we now insert the trivial bound \( \Sigma(K, b, h; q) \ll q^2 \), we obtain a bound that is never better than the trivial estimate \( S(K, M, N; q) \ll MN \). We must improve on this by exhibiting cancellation in the complete sum \( \Sigma(K, b, h; q) \) by exploiting the structure of \( K \).

In Section \( 6.3 \) we will further show how, for kernels \( K \) that are themselves given by a complete exponential sum of a specific shape, the estimation of \( \Sigma(K, b, h; q) \) reduces to the estimation (with square-root cancellation) of certain auxiliary additive character sums in two variables, which can in turn sometimes be treated using the Riemann Hypothesis over finite fields.

In particular, we will prove:

**Proposition 6.2.** Let \( q \) be a prime and define \( K(a) = K_{12}(a; q) \). With notation as above, for all \( b \in \mathcal{B} - \mathcal{B}^\Delta \) and all \( h \in \mathbb{F}_q \), we have
\[
(6.12) \quad |\Sigma(K, b, h; q)| \ll q
\]
where the implied constant is absolute.

Combining this bound and the contribution from \( \mathcal{B}^\Delta \), we obtain (in the case of Kloosterman sums) by \((6.8)\), \((6.10)\), and \((6.12)\) that
\[
S(K, M, N; q) \ll q^\varepsilon (AB)^{-1}(AN)^{3/4}(\|\alpha\|_1\|\alpha\|_2)^{1/2}(AB^2Mq + B^4q)^{1/4}.
\]
We then finish the proof of the second part of Theorem \( 6.1 \) by choosing
\[
A = M^{-\frac{1}{4}}N^{\frac{1}{4}}, \quad B = (MN)^{\frac{1}{4}};
\]

25
Theorem 6.4. Let \((K, \mathbf{b}; h; q)\). A weaker bound \(\Sigma(K, \mathbf{b}; h; q) \ll q^{3/2}\) can be proved easily and quite generally directly (by fixing one of the variables) from, say, [11], but this yields a power saving over the trivial bound for \(S(K, M, N; q)\) (say, if \(\alpha_m = 1\) only if \(N > q^{1/2+\delta}\) and \(MN^{5/2} > q^{2+\delta}\) for some \(\delta > 0\). This shows that we do require a stronger bound in the critical range \(M, N \asymp q^{1/2}\).

6.3. Reduction to two-variable character sums. We will now study the sums \(\Sigma(K, \mathbf{b}; q)\) for special kernels \(K\). Precisely, we assume that there exists a rational function \(f \in \mathbf{F}_q(T)\), not a linear polynomial, such that

\[
K(x) = q^{-1/2} \sum_{u (\text{mod } q)}' e_q(f(u) + xu)
\]

where the asterisk denotes that the values of \(u\) where \(f\) has a pole are excluded. In that case, Weil’s theory shows that \(K\) is bounded by some constant \(H\) depending only on the degrees of the numerator and denominator of \(f\), and we can attempt to estimate the corresponding bilinear form as in the previous section.

Replacing \(K\) in (6.9) by this formula and performing the averaging over \(s\), we obtain

\[
\Sigma(K, \mathbf{b}; h; q) = q^{-1} \sum_{r (\text{mod } q)} \sum_{(u, v, u', v') \in V_r(\mathbf{F}_q)} e_q(f(u) + f(v) - f(u') - f(v'))
\]

where \(V_r(\mathbf{F}_q)\) is set of solutions \((u, v, u', v')\) of the equation

\[
r(u + v - u' - v') + b_1u + b_2v - (b'_1u' + b'_2v') + h = 0.
\]

The sum further decomposes into two sums, depending on whether \((u, v, u', v')\) satisfies the additional equation

\[
u + v - u' - v' = 0
\]

or not. If \(u + v - u' - v' \neq 0\), there is, for a given \((u, v, u', v')\), only one possible \(r\) such that \((u, v, u', v') \in V_r(\mathbf{F}_q)\), and therefore the contribution \(\Sigma_1\) of these terms to \(\Sigma(K, \mathbf{b}; h; q)\) of these terms is equal to

\[
\Sigma_1 = q^{-1} \sum_{(u, v, u', v') (\text{mod } q)} \sum_{u + v - u' - v' \neq 0} e_q(f(u) + f(v) - f(u') - f(v')) = q^{-1}\left(q^2|K(0)|^4 - q \sum_{r (\text{mod } q)} |K(r)|^4\right) \ll q.
\]

We are left with the following 2-dimensional exponential sum over \(\mathbf{F}_q^4\)

\[
S(f, h, \mathbf{b}; q) := \sum_{(u, v, u', v') \in W(\mathbf{F}_q)} e_q(f(u) + f(v) - f(u') - f(v'))
\]

where \(W(\mathbf{F}_q)\) is the set of quadruples \((u, v, u', v') \in \mathbf{F}_q^4\) satisfying

\[
\begin{align*}
u + v &= u' + v', \\
b_1u + b_2v &= b'_1u' + b'_2v' - h.
\end{align*}
\]

We will prove the following estimate for these sums:

Theorem 6.4. Let \(f(T) = 1/T \in \mathbf{F}_q(T)\) and consider four non-zero linear forms in two variables

\[
l_1(u, v) := u, \ l_2(u, v) := v, \ l_3(u, v) = \alpha u + \beta v, \ l_4(u, v) = \gamma u + \delta v.
\]

If

\[
\{l_3, l_4\} \neq \{l_1, l_2\},
\]
then, for all \( h \in \mathbb{F}_q \), we have
\[
\sum_{u,v} e_q(f(u) + f(v) - f(l_3(u,v) + h) - f(l_4(u,v) - h)) \ll q,
\]
where the implied constant is absolute.

We will prove this in the next section. Assuming the result, we conclude the proof of Proposition 6.2 (and hence of Theorem 6.1) as follows: if \( b_1 \neq b_2 \), we can write
\[
S(f, h, b; q) = \sum_{u,v} e_q(f(u) + f(v) - f(l_3(u,v) + h') - f(l_4(u,v) - h'))
\]
where
\[
l_3(u,v) = \frac{b_1 - b_2'}{b_1' - b_2'} u + \frac{b_2 - b_2'}{b_1' - b_2'} v, \quad l_4(u,v) = \frac{b_1' - b_1}{b_1' - b_2'} u + \frac{b_1' - b_2}{b_1' - b_2'} v,
\]
\[
h' = \frac{h}{b_1' - b_2'}.
\]
Simple checks show that the sets \( \{l_1, l_2\} \) and \( \{l_3, l_4\} \) thus defined coincide only if \( b \in \mathbb{B}^\Delta \). Hence we then get
\[
S(f, h, b; q) \ll q
\]
for all \( b \in \mathbb{B} - \mathbb{B}^\Delta \) and all \( h \in \mathbb{F}_q \).

If \( b_1' = b_2' \) but \( b_1 \neq b_2 \), we can proceed in a similar way, exchanging the roles of \((u,v)\) and \((u',v')\).

This gives the desired bounds except when \( b_1 = b_2 \) and \( b_1' = b_2' \). But such quadruples \( b \) are also in \( \mathbb{B}^\Delta \). \qed

6.4. Estimate of two-variable character sums. We prove Theorem 6.4 in this section. By a general criterion (due to Hooley [15, Theorem 5] and Katz [18, Cor. 4], see also [12, Prop. 2.1]), the desired estimate follows from the Riemann Hypothesis over finite fields of Deligne for any \( h \in \mathbb{F}_q \) such that the rational function
\[
F(U,V) = f(U) + f(V) - f(l_3(U,V) + h) - f(l_4(U,V) - h) \in \mathbb{F}_q(U,V)
\]
is not composed, which means that it is not of the shape
\[
F = Q \circ P
\]
where \( P \in \mathbb{F}_q(U,V) \) and where \( Q \in \mathbb{F}_q(T) \) is a rational function which is not a fractional linear transformation \((aT + b)/(cT + d)\) (in particular \( F(U,V) \) is not constant).

This is a purely geometric question and we will show this more generally for \( h, \alpha, \beta, \gamma, \delta \in \mathbb{F}_q \), under the assumption that \( \{l_1, l_2\} \neq \{l_3, l_4\} \).

In the following, we denote by \( C \in \mathbb{F}_q \) a non–zero constant, the value of which may change from one line to another. We follow closely the method of [12, Proposition 2.3] but first we make the birational change of variables
\[
X = U/V, \quad Y = V,
\]
so that
\[
F(U,V) = f(XY) + f(Y) - f(Yl_3(X,1) + h) - f(Yl_4(X,1) - h)
\]
\[
= \frac{1}{XY} + \frac{1}{Y} - \frac{1}{Yl_3(X,1) + h} - \frac{1}{Yl_4(X,1) - h}.
\]

We then need to prove that \( F(XY,Y) \) is not of the shape
\[
\frac{Q_1(P_1(X,Y)/P_2(X,Y))}{Q_2(P_1(X,Y)/P_2(X,Y))}
\]
where \( P_1, P_2 \) are polynomials.
where \( P_1(X, Y), P_2(X, Y) \in \mathbf{F}_q[X, Y] \) are coprime polynomials in two variables and

\[
Q_1(T) = C \prod_\lambda (T - \lambda)^{m(\lambda)}, \quad Q_2(T) = \prod_\mu (T - \mu)^{m(\mu)},
\]

are coprime polynomials in one variable (here \( m(\lambda) \) and \( m(\mu) \) denote the multiplicity of the zeros \( \lambda \) and \( \mu \)). Moreover, up to changing the variable \( T \) by a fractional linear transformation, we may assume that the degrees \( q_1(= \sum_\lambda m(\lambda)) \) and \( q_2(= \sum_\mu m(\mu)) \) of \( Q_1 \) and \( Q_2 \) satisfy the inequality

(6.13) \quad q_1 > q_2,

which means that \( \infty \) is a pole of \( Q \). Our objective is then to show that

(6.14) \quad q_1 = 1.

We have the identity

\[
F(XY, Y) = \frac{C \prod_\lambda (P_1(X, Y) - \lambda P_2(X, Y))^{m(\lambda)}}{P_2(X, Y)^{q_1 - q_2} \prod_\mu (P_1(X, Y) - \mu P_2(X, Y))^{m(\mu)}} =: \frac{\text{NUM}(X, Y)}{\text{DEN}(X, Y)}.
\]

In this latter expression, the numerator and denominator, \( \text{NUM}(X, Y) \) and \( \text{DEN}(X, Y) \), are coprime.

We also have

(6.15) \quad F(XY, Y) = \frac{(Yl_3(X, 1) + h)(Yl_4(X, 1) - h)(1 + X) - XY^2(l_3(X, 1) + l_4(X, 1))}{XY(Yl_3(X, 1) + h)(Yl_4(X, 1) - h)}.

By the assumption (6.13), we deduce that \( P_2(X, Y) \) is not a constant polynomial (it suffices to compare the differences of the total degrees of the numerator and of the denominator of the two above expressions of \( F(XY, Y) \)). We distinguish two cases to finish the proof.

(1) Assume first that \( h \neq 0 \). If \( l_3 + l_4 \neq 0 \) then the numerator and denominator of (6.15) are coprime and are equal to \( C \cdot \text{NUM}(X, Y) \) and \( C \cdot \text{DEN}(X, Y) \) respectively. Since the factors \( X, Y, Yl_3(X, 1) + h, Yl_4(X, 1) - h \) are simple and coprime and since \( P_2(X, Y) \) is not constant, we have

(6.16) \quad q_1 - q_2 = 1,

and if \( q_2 \neq 0 \), we necessarily have \( m(\mu) = 1 \) for any \( \mu \).

In particular if \( q_2 = 0 \), we obtain (6.14) and we are done.

Suppose now that \( q_2 \geq 1 \). If \( Y \) does not divide \( P_2 \), it divides some \( P_1 - \mu P_2 \) (and then \( m(\mu) = 1 \)) and up to the change of variable \( T \mapsto T + \mu \) (which does not change the condition \( q_1 - q_2 > 0 \)) we may assume that \( \mu = 0 \): hence \( Y \mid P_1 \) and all the zeros \( \lambda \) of \( Q_1 \) are non-zero. Hence, in all the cases, we have \( Y \mid P_1 P_2 \) from which we deduce the equality

\[
\text{NUM}(X, 0) = CP_1(X, 0)^{q_1} \quad \text{or} \quad CP_2(X, 0)^{q_1}
\]

but \( \text{NUM}(X, 0) = -h^2(1 + X) \) and therefore \( q_1 = 1 \). This contradicts the equality (6.16) and the assumption \( q_2 \geq 1 \).

The proof when \( l_3 + l_4 = 0 \) is identical except that the fraction \( F(XY, Y) \) simplifies to the reduced fraction

\[
F(XY, Y) = \frac{1 + X}{XY}.
\]

(2) Assume now finally that \( h = 0 \). In this case we have

(6.17) \quad F(XY, Y) = \frac{l_3(X, 1)l_4(X, 1)(1 + X) - X(l_3(X, 1) + l_4(X, 1))}{XYl_3(X, 1)l_4(X, 1)}.

Let us assume that \( F(XY, Y) \neq 0 \). The polynomials \( \text{NUM}(X, Y) \) and \( \text{DEN}(X, Y) \) divide the numerator and denominator of the right-hand side of (6.17), in particular \( \text{NUM}(X, Y) \) does not depend on \( Y \). Suppose that \( q_2 \geq 1 \); by arguing as above and up possibly to making a change of
variable $T \mapsto T + \mu$, we may assume that $0 \not\in \{\lambda \mid Q_1(\lambda) = 0\}$ and that either $Y$ divides $P_1(X,Y)$ or $P_2(X,Y)$ (but not both); in either cases, this is not compatible with the equality

$$\text{NUM}(X,Y) = C \prod_{\lambda} (P_1(X,Y) - \lambda P_2(X,Y))^{m(\lambda)},$$

since the left-hand side only depends on $X$ and the $\lambda$ are $\not= 0$. Therefore $q_2 = 0$ and $Y$ divides $P_2(X,Y)^{q_1}$ to order 1 so that $q_1 = 1$. The only remaining case is when

$$l_3(X,1)l_4(X,1)(1 + X) - X(l_3(X,1) + l_4(X,1)) = 0.$$

By the explicit expressions $l_3(X,1) = \alpha X + \beta$ and $l_4(X,1) = \gamma X + \delta$ and by the fact that $l_3$ and $l_4$ are not zero, the above equality is equivalent to $\{l_3, l_4\} = \{l_1, l_2\}$. \hfill \Box

6.5. Conjectural bounds for bilinear forms in Kloosterman sums. In this section, which is not needed for the proof of the unconditional results of this paper, we prove Proposition 3.1 which provides a bound for the bilinear forms

$$\sum_{m \in M, n \in N} \alpha_m \beta_n K(mn)$$

where $(\alpha_m)$, $(\beta_n)$ are sequences of complex numbers, and we assume that $M, N \in [1, q]$ satisfy

$$M \leq Nq^{1/4}, \quad q^{1/4} \leq MN \leq q^{5/4},$$

that $N$ is an interval of length $N$ and that

$$K(x) = \text{KL}_2(ax; q)$$

for some $(a, q) = 1$. More precisely, we show how this bound follows from Conjecture 6.5 below on a certain complete sum of products of Kloosterman sums, which is the content of Proposition 3.1.

In [12, §VII], Fouvry and Michel studied bilinear forms as in (6.18) for the kernels $K(x)$ given by

$$K(x) = e\left(\frac{x^k + x}{q}\right)$$

for some fixed integer $k$, such that either $k < 0$ or $k \geq 3$ is odd. They proved the bound

$$\sum_{M < m \leq 2M} \sum_{N < n \leq 2N} \alpha_m \beta_n K(mn) \ll_{\epsilon, k} q^\epsilon \|\alpha\|_2 \|\beta\|_2 (MN)^{1/2} (M^{-1/2} + q^{11/24}(MN)^{-\epsilon/24})$$

for any $\epsilon > 0$. This bound follows from bounds for families of multivariable complete algebraic exponential sums, using the work of Deligne and Katz. One is led to believe that a corresponding bound holds for $K(x) = \text{KL}_2(ax; q)$, uniformly in $a$ coprime with $q$, and this exactly leads to the statement of Proposition 3.1.

We now establish this proposition by repeating the argument of Fouvry and Michel. With $K(x) = \text{KL}_2(ax; q)$, the Cauchy-Schwarz inequality gives

$$\left| \sum_{m \in M, n \in N} \alpha_m \beta_n K(mn) \right|^2 \leq \|\beta\|_2^2 \sum_{m_1, m_2 \in M} \sum_{n \in N} \alpha_{m_1} \overline{\alpha}_{m_2} \sum_{n \in N} K(m_1 n) \overline{K}(m_2 n) =: \|\beta\|_2^2 (\Sigma = + \Sigma^\#)$$

where $\Sigma = \epsilon$ is the contribution of the diagonal terms $m_1 \equiv m_2 (\text{mod} \ q)$ and $\Sigma^\#$ is the remaining off-diagonal contribution. The diagonal term is bounded by $O(\|\alpha\|_2^2 N)$. For the remaining terms we

---

\footnote{Actually they give a slightly weaker bound with $\|\alpha\|_2 \|\beta\|_2$ replaced by $(MN)^{1/2}$ under the assumption that $|\alpha_m|, |\beta_n|$ are bounded by 1; as we show below the method of [12] yields the slightly stronger bound presented here.}
apply again Vinogradov’s “shift by \(ab\)” trick: given \(A, B \geq 1\) satisfying the conditions (6.7) (these will be satisfied by (6.10) after a suitable choice of \(A, B\)), the off-diagonal term \(\Sigma^\neq\) is bounded by

\[
\frac{1}{AB} \sum_{\substack{A < a \leq 2A \\ B < b \leq 2B}} \sum_{m_1, m_2 \leq M} \sum_{n + ab \in \mathbb{N}} \alpha_{m_1} \overline{\alpha_{m_2}} \sum_{\nu_0} K(am_1(\nu_0 + b)) \overline{K}(am_2(\nu_0 + b))
\]

which is itself bounded by

\[
\ll \frac{q^e}{AB} \sum_{r \pmod{q}} \sum_{1 \leq s_1, s_2 \leq AM} \nu(r, s_1, s_2) \sum_{B < b \leq 2B} \eta_b K(s_1(r + b)) \overline{K}(s_2(r + b))
\]

where \(|\eta_b| \leq 1\) and

\[
\nu(r, s_1, s_2) = \sum \cdots \sum_{A < a \leq 2A, m_1, m_2 \leq M, n \in \mathbb{N}} |\alpha_{m_1}| |\alpha_{m_2}|
\]

By the same reasoning as above we have

\[
\sum_{r, s_1, s_2} \nu(r, s_1, s_2) \ll AN||\alpha||^2_1 \leq AMN||\alpha||^2_2
\]

and

\[
\sum_{r, s_1, s_2} \nu(r, s_1, s_2)^2 \ll q^e AN||\alpha||^4_2.
\]

From these bounds and Hölder’s inequality (see [12, Lemma 7.1]) and (6.7), we obtain that \(\Sigma^\neq\) is bounded by

\[
\frac{q^e}{AB} (AN)^{3/4} M^{1/2} ||\alpha||^2_2
\]

\[
\left( \sum_{b} \sum_{r \pmod{q}} \sum_{1 \leq s_1, s_2 \leq AM} \prod_{i=1}^{2} K(s_1(r + b_i)) \overline{K}(s_2(r + b_i)) \overline{K}(s_1(r + b_{i+2})) \overline{K}(s_2(r + b_{i+2})) \right)^{1/4}
\]

for \(b\) running over the set \(\mathcal{B}\) of quadruples \((b_1, b_2, b_3, b_4)\) satisfying \(B < b_i \leq 2B\). We bound the inner triple sum over \(r, s_1, s_2\) depending on the value taken by \(b\): let

\[
\mathcal{B}^\Delta \subset \mathcal{B}
\]

be the “diagonal” set of elements \(b\) for which some pair \((b_i, b_j)\) \((i, j \leq 4\) with distinct indices equals a pair having complementary indices (for instance \((b_1, b_4) = (b_3, b_2)\)). We have \(|\mathcal{B}^\Delta| = O(B^2)\) and for \(b \in \mathcal{B}^\Delta\) we use the trivial bound to obtain

\[
\sum_{b \in \mathcal{B}^\Delta} \sum_{r \pmod{q}} \sum_{s_1, s_2} \cdots \ll qA^2B^2M^2.
\]

For the \(O(B^4)\) elements not contained in \(\mathcal{B}^\Delta\) we detect the condition \(s_1 \not\equiv s_2 \pmod{q}\) via additive characters, writing

\[
\delta(s_1 \not\equiv s_2 \pmod{q}) = 1 - \frac{1}{q} \sum_{\lambda \pmod{q}} e_q(\lambda(s_1 - s_2)).
\]

We then complete the \(s_1, s_2\) sums, also using additive characters: for \(\lambda \in \mathbb{Z}/q\mathbb{Z}\) let

\[
S(r, \lambda; q) = \sum_{s \pmod{q}} K(s(r + b_1)) K(s(r + b_2)) \overline{K}(s(r + b_3)) \overline{K}(s(r + b_4)) e_q(\lambda s),
\]

30
\[
\mathcal{R}(\mu_1, \mu_2; q) := \sum_{r \mod q} S(r, \mu_1; q) \overline{S(r, \mu_2; q)}
\]
and
\[
\Sigma(b, \mu_1, \mu_2; q) = \mathcal{R}(\mu_1, \mu_2; q) - \frac{1}{q} \sum_{\lambda \mod q} \mathcal{R}(\mu_1 + \lambda, \mu_2 + \lambda; q).
\]

Now we see that we are led to make the following auxiliary conjecture:

**Conjecture 6.5.** There exists a constant \(C\) such that for any prime \(q\), every integer \(a\) coprime with \(q\), every \(\mu_1, \mu_2 \in F_q\) and every \(b \in B^\text{gen} := B \setminus B\Delta\) we have
\[
|\Sigma(b, \mu_1, \mu_2; q)| \leq Cq^{3/2};
\]
here the sum \(\Sigma\) is the sum relative to the function \(K\) defined in (6.20).

If we assume Conjecture 6.5 we obtain that
\[
\sum_{b \in B^\text{gen}} \left| \sum_{r \mod q} \sum_{s_1, s_2, \ldots} \cdots \right| \ll q^2 B^2 q^{3/2}.
\]

Hence, under this assumption, we have
\[
\Sigma^\neq \ll \frac{q^2}{AB} (AN)^{3/4} M^{1/2} ||\alpha||^2_2 (A^2 B^2 M^2 q^3 + B^4 q^{3/2})^{1/4}.
\]

We may choose (see [12, p.128])
\[
A = q^{1/8} M^{-1/2} N^{1/2}, \quad B = q^{-1/8} (MN)^{1/2}
\]
for which (6.7) as well as \(A, B \geq 1\) are satisfied by (6.19). Combining this bound with that for \(\Sigma^=\), we conclude that Proposition 3.1 follows from Conjecture 6.5.

### 7. Evaluation of Moments of \(L\)-functions

In this section, we implement the strategy sketched in Section 3 to prove Theorems 1.1, 1.2 and 1.3.

#### 7.1. First steps.
Let \(f, g\) be either holomorphic Hecke cusp forms of level 1, or the Eisenstein series \(E\) defined in (1.3). Let \(q\) be a prime number. We decompose the second moment (1.2) into the moments of twists by even and odd characters separately:
\[
M_{f,g}(q) = M_{f,g,1}(q) + M_{f,g,-1}(q),
\]
where, for \(\sigma \in \{-1, 1\}\), we put
\[
M_{f,g,\sigma}(q) = \frac{1}{\varphi(q)} \sum_{\chi \text{ primitive}} L(f \otimes \chi, 1/2) L(g \otimes \overline{\chi}, 1/2).
\]

Using the computation of the root number in Lemma 2.1 and the invariance of the parity \(\chi(-1)\) under complex conjugation, we find that
\[
M_{f,g,\sigma}(q) = \frac{1 + \varepsilon(f, g, \sigma)}{2} M_{f,g,\sigma}(q),
\]
where \(\varepsilon(f, g, \sigma)\) is the root number \(\varepsilon(f, g, \chi)\) for any primitive character \(\chi\) with parity \(\chi(-1) = \sigma\). Thus \(M_{f,g,\sigma}(q) = 0\) unless
\[
\varepsilon(f, g, \sigma) = 1,
\]
which we henceforth assume. By the approximate functional equation (2.2), we have
\[
L(f \otimes \chi, 1/2) L(g \otimes \chi, 1/2) = 2 \sum_{m,n \geq 1} \frac{\lambda_f(m) \lambda_g(n)}{(mn)^{1/2}} \chi(m) \chi(n) V_{f,g,\sigma} \left( \frac{mn}{q^2} \right)
\]
where the function \( V_{f,g,\sigma} \) is given by (2.3).

We now average over \( \chi \) of parity \( \sigma \). The orthogonality relation for these characters is
\[
\left( \frac{q}{\chi(1)} \right) \sum_{\chi \equiv \chi^* (mod q)} \chi(m) \chi(n) = \delta_{m \equiv n (mod q)} + \sigma \delta_{m \equiv -n (mod q)}
\]
for \( q \) prime and any integers \( m \) and \( n \) such that \( (mn, q) = 1 \). Inserted in the above formula, it yields
\[
M_{f,g,+1}(q) = B_{f,g,+1}^+(q) + B_{f,g,+1}^-(q), \quad M_{f,g,-1}(q) = B_{f,g,-1}^+(q) - B_{f,g,-1}^-(q),
\]
where
\[
B_{f,g,\sigma}^\pm(q) = \sum_{m \equiv \pm n (mod q)} \frac{\lambda_f(m) \lambda_g(n)}{(mn)^{1/2}} V_{f,g,\sigma} \left( \frac{mn}{q^2} \right)
\]
(indeed, the second term in (7.2) is canceled in the right hand side of \( M_{f,g,-1}(q) \), and for \( M_{f,g,+1}(q) \) it compensates the missing trivial character).

A diagonal main term \( MT_{f,g,\sigma}^d(q) \) is given by the contribution of \( n = m \) in \( B_{f,g,\sigma}^+(q) \). By Mellin inversion and a contour shift, we can compute explicitly:
\[
MT_{f,g,\sigma}^d(q) = \sum_{m \geq 1 \atop (m,q) = 1} \frac{\lambda_f(m) \lambda_g(m)}{m} V_{f,g,\sigma} \left( \frac{m^2}{q^2} \right)
\]
\[
= \text{res}_{s=0} \frac{L_\infty(f \otimes \chi, 1/2 + s) L_\infty(g \otimes \chi, 1/2 + s) L(q)(f \otimes g, 1 + 2s)}{\zeta(q)(2 + 4s)} \frac{q^{2s}}{s} + O(q^{-1/2 + \varepsilon}),
\]
for any \( \varepsilon > 0 \), where \( \chi \) denotes any primitive character of modulus \( q \) of parity \( \chi(-1) = \sigma \), \( L(f \otimes g, s) \) denotes the Rankin-Selberg \( L \)-function of \( f \) and \( g \), including
\[
L(f \otimes E, s) = L(f, s)^2, \quad L(E \otimes E, s) = \zeta(s)^4,
\]
and the superscript \( (q) \) denotes omission of the Euler factor at \( q \).

Computing the residue explicitly, we find that
\[
MT_{f,g,\sigma}^d(q) = MT_{f,g,\sigma}^0(q) + O(q^{\varepsilon-1/2})
\]
where
\[
MT_{f,g,\sigma}^0(q) = \begin{cases} P_{f,\sigma}(\log q) & \text{for } P_{f,\sigma}(X) \text{ a degree 1 polynomial if } f = g \text{ is cuspidal,} \\ L(f \otimes g, 1) & \text{if } f \neq g \text{ are both cuspidal,} \\ \frac{\zeta(2)}{\zeta(2)} & \text{if } f \text{ is cuspidal and } g = E, \\ \frac{L(f, 1)^2}{\zeta(2)} & \text{if } f \text{ is cuspidal and } g = E, \\ P_{g,\sigma}(\log q) & \text{for } P_{g,\sigma}(X) \text{ a polynomial of degree } 4 \text{ if } f = g = E. \end{cases}
\]

We note also that, by Lemma 2.1, the root number \( \varepsilon(f, g, \pm 1) \) is always 1 or always \(-1 \) if \( f \) and \( g \) are cuspidal, but it is 1 for exactly one choice of sign if \( f \) is cuspidal and \( g = E \). This explains the additional factor of 2 in Theorem 1.3.
The contribution of the terms \( n = m \) in \( B_{f,g} \) is easily seen to be \( O(q^{-3/4+\varepsilon}) \) using only (1.7). For the remaining terms in (7.2) (that is, those with \( m \neq n \)), we apply a partition of unity to the \( m,n \) variables and are led to evaluate the dyadic sums

\[
\sum_{M,N \geq 1} \sum_{m \equiv \pm n \mod q \atop m \neq n} \frac{\lambda_f(m)\lambda_g(n)}{(mn)^{1/2}} V_{f,g,\pm 1} \left( \frac{mn}{q^2} \right) W_1 \left( \frac{m}{M} \right) W_2 \left( \frac{n}{N} \right)
\]

up to an error of size \( O(q^{-1+\varepsilon}) \), for any \( \varepsilon > 0 \), that arises from removing the condition \( (mn,q) = 1 \) and replacing \( \varphi^*(q) \) by \( q \). In this expression, the symbol \( \sum_{dy}^{\langle M,N \rangle} \) indicates that \( M,N \geq 1 \) range over powers of 2, and \( W_1, W_2 \) are smooth compactly supported on \([1/2,2]\) satisfying \( W^{(j)}_1(x) \ll_j 1 \) for \( i = 1, 2 \) and all \( j \geq 0 \). Using the rapid decay of \( V_{f,g,\pm 1}(x) \), we may moreover, up to an admissible error term, assume that \( M, N \) satisfy

\[
1 \leq MN \leq q^{2+\varepsilon}.
\]

In order to evaluate the remaining \( O(\log^2 q) \) sums with \( M \) and \( N \) fixed, we first separate the variables \( m \) and \( n \). We proceed by Mellin inversion (as in [23]): using the definition (2.3) of \( V_{f,g,\pm 1}(x) \) as a Mellin transform, we shift the line of integration to \( \Re s = \varepsilon \) and approximate

\[
V_{f,g,\pm 1}(x) = \frac{1}{2\pi i} \int_{(\varepsilon),|s| \leq \log^2 q} \frac{L_\infty(f \otimes \chi, 1/2 + s) L_\infty(g \otimes \chi, 1/2 + s)}{L_\infty(f \otimes \chi, 1/2) L_\infty(g \otimes \chi, 1/2)} x^{-s} \frac{ds}{s} + O(q^{-100})
\]

due to the exponential decay of \( L_\infty(f \otimes \chi, 1/2 + s) L_\infty(g \otimes \chi, 1/2 + s) \) as \( |\Im s| \to \infty \). We exchange summation and integration and, up to replacing \( W_1(x), W_2(x) \) by \( x^{-1/2-\varepsilon} W_1(x), x^{-1/2-\varepsilon} W_2(x) \), we are led to evaluating bilinear sums of the shape

\[
B_{f,g}^\pm(M, N) = \frac{1}{(MN)^{1/2}} \sum_{m \equiv \pm n \mod q \atop m \neq n} \lambda_f(m)\lambda_g(n) W_1 \left( \frac{m}{M} \right) W_2 \left( \frac{n}{N} \right)
\]

\[
- \frac{1}{q(MN)^{1/2}} \sum_{m,n} \lambda_f(m)\lambda_g(n) W_1 \left( \frac{m}{M} \right) W_2 \left( \frac{n}{N} \right),
\]

with new test functions \( W_1, W_2 \) (which depend on \( s \)) satisfying (1.8), since \( s = \varepsilon + it \) and \( |t| < \log^2 q \).

As explained in Section 3, our objective is then to show that

\[
B_{f,g}^\pm(M, N) = \delta_{f=g=EMT_{E,E}^\pm}(M, N) + O(q^{-\eta+\varepsilon})
\]

for any \( \varepsilon > 0 \), with

\[
\begin{cases}
\eta = 1/32, & f = g = E, \\
\eta = 1/68, & f \text{ cuspidal, } g = E, \\
\eta = 1/144, & f, g \text{ both cuspidal}.
\end{cases}
\]

Once this is done (uniformly in terms of \( W_1 \) and \( W_2 \)), we can perform the last integration over \( s \) and finish the proof of the theorems.

We will now begin the proof of this estimate. To ease notation, we define the exponents \( \mu, \nu, \mu^*, \nu^* \) by

\[
M = q^\mu, \quad N = q^\nu, \quad \mu^* := 2 - \mu, \quad \nu^* = 2 - \nu.
\]

By (7.3) we have

\[
0 \leq \mu + \nu \leq 2 + \varepsilon.
\]
We consider the three main results in turn.

7.2. The case \( f \) and \( g \) cuspidal. Let \( \eta = 1/144 \). We attempt to prove (7.5). First of all, using the bound (4.11) (which depends on the Ramanujan-Petersson conjecture), we obtain (7.5) immediately if \( \mu + \nu \leq 2 - 2\eta \). We can therefore assume that the exponents satisfy
\[
-2\eta \leq \mu - \nu^* = \nu - \mu^* \leq \varepsilon.
\]

By symmetry, we may assume that \( \nu \geq \mu \) (up to exchanging the roles of \( f \) and \( g \)). From (4.4), we obtain that (7.5) holds unless
\[
1 - 4\eta \leq \nu - \mu,
\]
which we now assume.

By (7.6), we then have \( \nu \geq 3/2 - 3\eta \) in which case the condition \( n \neq m \) is void; it is natural to apply the Voronoi summation formula (Lemma 2.3) to the long \( n \)-variable. To this end, we detect the condition \( m \equiv \pm n \mod q \) by additive characters. The trivial character cancels the second term on the right hand side of (7.4), and one obtains the formula
\[
B_{f,g}^\pm(M, N) = \frac{1}{(qMN^*)^{1/2}} \sum_{m,n\geq 1} \lambda_f(m)\lambda_g(n)W_1\left(\frac{m}{M}\right) \frac{1}{N} \hat{W}_{2,N} \left(\frac{n}{q^*}\right) \text{Kl}_2(\pm mn; q)
\]
with \( q^* = q^2/N \), where we use the notation of Lemma 2.3. In particular, by this lemma, the function
\[
y \mapsto \frac{1}{N} \hat{W}_{2,N} \left(\frac{y}{q^2}\right)
\]
decays rapidly for \( y \geq q^* N^* \) and the contribution to \( B_{f,g}^\pm(M, N) \) of those \( n \) that satisfy \( n \geq q^* N^* \) is negligible. By a partition of unity (using Lemma 1.6) we can decompose (7.7) into a sum of \( O(\log q) \) terms of the shape
\[
C^\pm(M, N') = \frac{1}{(qMN^*)^{1/2}} \sum_{m,n\geq 1} \lambda_f(m)\lambda_g(n)W_1\left(\frac{m}{M}\right) W_2\left(\frac{n}{N'}\right) \text{Kl}_2(\pm mn; q)
\]
with \( W_1, W_2 \) satisfying (1.8) and \( N' = q^{\nu'} \leq q^* N^* \).

By Weil’s bound for Kloosterman sums \( |\text{Kl}_2(\pm mn; q)| \leq 2 \), we have the trivial bound
\[
C^\pm(M, N') \ll q^\varepsilon (MN^*/q)^{1/2} = q^{\varepsilon + \mu + \nu^* - 1/2},
\]
which establishes (7.5) unless
\[
1 - 2\eta \leq \mu + \nu^* \leq 1 + 4\eta.
\]
When this condition holds, it follows from (7.6) that
\[
1/2 - 2\eta \leq \mu \leq 1/2 + 2\eta + \varepsilon/2 \quad \text{and} \quad 1/2 - \eta - \varepsilon/2 \leq \nu^* \leq 1/2 + 3\eta.
\]
Now, by Proposition 3.1 (whose conclusion we recall is conditional on Conjecture 6.3 and whose assumption \( U, V \leq q \) is satisfied by the preceding display for \( \eta < 1/6 \), we have
\[
C^\pm(M, N') \ll q^\varepsilon (MN^*/q)^{1/2}((N^*)^{-1/2} + q^{11/64}(MN^*)^{-3/16}) \ll q^{\varepsilon + \mu + \nu^* - 1/2} = q^{\varepsilon - \eta}
\]
for \( \eta = 1/144 \) (the worst bound occurs when \( N' = q^* N^* \)). This concludes the proof of Theorem 1.3. \( \square \)

We remark that the exponent 1/144 is the result of an optimization problem: combining the bounds (4.11), (4.4), (7.8) and Proposition 3.1, 

\[
\min \left( \frac{\mu + \nu}{2} - 1, \max \left( \frac{\nu - \mu - 1}{2}, \frac{\nu - \mu - 1}{4}, \frac{\mu + \nu - 1}{2}, \frac{\mu + \nu^* - 1}{2} \right) + N \right)
\]

employed in the proof of Theorem 1.3.
subject to the constraints

\[ 0 \leq \mu \leq \nu, \quad \mu + \nu \leq 2, \quad \nu^* = 2 - \nu. \]

This is a linear optimization problem that can be solved exactly by computer in a finite search, and a maximum is obtained at \( \mu = 73/144, \ \nu = 71/48. \) The Mathematica code is available after the bibliography; we double-checked the numerical values also with Matlab.

7.3. The case \( f = g = E. \) Next we prove Theorem 1.1.

Let \( \eta = 1/32. \) We are once more in a symmetric case, so we can assume that \( \mu \leq \nu. \) Moreover, the Ramanujan-Petersson conjecture is trivially true, so we may apply (4.1) to obtain (7.5) if

\[ \mu + \nu \leq 2 - 2\eta, \quad 2\eta \leq \nu - \mu. \]

The grouping and the analysis of the main terms was done in [24], so we will focus on the error term. Applying first (4.4) we obtain

\[ E_{T, E}(M, N) \ll q^{-\eta+\epsilon}, \]

as desired, unless

\[ \mu + \nu^* \leq 1 + 4\eta \]

which we assume from now on.

In this remaining range, the off-diagonal main term \( M_{T, E}^\text{od}(M, N) \ll q^{-7/16+\epsilon} \) is small (cf. the second term in Proposition 4.1), so that we can assume (7.6), and it suffices to prove the estimate (7.9)

\[ B_{T, E}(M, N) \ll q^{-\eta+\epsilon}. \]

We use the letters \( N^*, N' \) etc. as in the preceding subsection. We detect again the congruence by applying the Voronoi summation formula to the \( n \)-variable. This expresses the sum \( B_{T, E}(M, N) \) into a main term and two additional terms. As in (7.18), the main term is \( O(q^{-1+\epsilon}) \), while the error terms decompose into \( O(\log q) \) terms of the shape

\[ \frac{1}{(qMN^*)^{1/2}} \sum_{m,n} d(m)d(n)W_1 \left( \frac{m}{M} \right) W_2 \left( \frac{n}{N'} \right) K_2(\pm mn; q) \]

where \( W_1, W_2 \) satisfy (1.8) (the definition of \( W_2 \) has changed from its preceding appearance).

A trivial estimate shows that (7.9) holds unless

\[ 1 - 2\eta \leq \mu + \nu^* \leq 1 + 4\eta, \]

which we then assume.

We further decompose (7.10) into \( O(\log^4 q) \) terms of the form

\[ \frac{1}{(qMN^*)^{1/2}} \sum_{m_1,m_2,n_1,n_2} W_1 \left( \frac{m_1m_2}{M} \right) W_2 \left( \frac{n_1n_2}{N'} \right) \]

\[ \times W \left( \frac{m_1}{M_1} \right) W \left( \frac{m_2}{M_2} \right) W \left( \frac{n_1}{M_3} \right) W \left( \frac{n_2}{M_4} \right) K_2(\pm m_1n_2n_1n_2; q) \]

with

\[ M_1M_2 = M, \quad M_3M_4 = N' \leq N^*. \]

In (7.12), we separate the variables \( m_1, m_2 \) resp. \( n_1, n_2 \) in \( W_1(m_1m_2/M_1M_2) \) and \( W_2(n_1n_2/M_3M_4) \) by inverse Mellin transform: we write

\[ W_1 \left( \frac{m_1m_2}{M_1M_2} \right) = \frac{1}{2\pi i} \int_0 \widehat{W_1}(s) \frac{M_1^s}{m_1^s} \frac{M_2^s}{m_2^s} ds \]

and exchange the order of summations and integrals. For any \( \epsilon > 0, \) the contribution to the integral of the \( s \) such that \( |s| \geq q^\epsilon \) is negligible, by (1.8) and repeated integration by parts.
Possibly with different $W_i$, $i = 1, 2, 3, 4$, and up to renaming some variables, we are reduced to estimating sums of the shape

$$S^\pm(M_1, M_2, M_3, M_4) = \frac{1}{(qMN^*)^{1/2}} \sum_{m_1, m_2, n_1, n_2} W_1 \left( \frac{m_1}{M_1} \right) W_2 \left( \frac{m_2}{M_2} \right) \times W_3 \left( \frac{m_3}{M_3} \right) W_4 \left( \frac{m_4}{M_4} \right) KL_2(\pm m_1 m_2 m_3 m_4; q),$$

where the $W_i$ satisfy (1.8) and hence (6.4) for $Q = q^r$, and the $M_i$ written in the shape $M_i = q^{\mu_i}$, $i = 1, 2, 3, 4$, satisfy

$$\mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4, \quad \mu_1 + \mu_2 + \mu_3 + \mu_4 = \mu + \nu, \quad \nu' \leq \nu^*.$$

The strategy is the following: if the product of two smooth variables is long (if $\mu_3 + \mu_4$ is large, in particular larger than 3/4) we apply the third part (6.5) of Theorem 6.1 with $MN = M_3 M_4$, and we sum trivially over $m_1$ and $m_2$. If this is not the case, it is possible to factor the product $m_1 m_2 n_1 n_2$ into a product $mn$ in such a way that an application of the general bilinear estimate (6.1) is beneficial.

Explicitly, let $\delta < 1/4$ be some parameter such that

$$1/4 - \delta \leq \frac{1}{6} (\mu + \nu^*).$$

If

$$\mu_1 + \mu_2 \leq \frac{1}{4} - \delta,$$

we apply (6.5) with $MN = M_3 M_4$ and sum trivially over $m_1$ and $m_2$, obtaining the bound

$$q^{-(A+1)} S^\pm(M_1, M_2, M_3, M_4) \ll q^{\frac{1}{2} (\mu + \nu^* - \frac{\delta}{2})} + q^{-\frac{\delta}{2}},$$

for the constant $A$ occurring in (6.3).

On the other hand, if

$$\mu_1 + \mu_2 \geq \frac{1}{4} - \delta,$$

then at least one of $\mu_2$ and $\mu_1 + \mu_2$ is contained in the interval

$$\left[ \frac{1}{4} - \delta, \frac{1}{3} (\mu + \nu^*) \right],$$

since $\mu_2 \leq (\mu + \nu^*)/3$ and $\mu_1 \leq \mu_2$. Let $u$ be one of the numbers $\mu_2$ or $\mu_1 + \mu_2$ satisfying this condition. We then apply (6.1) with

$$(M, N) \leftrightarrow (q^u, MN'q^{-u})$$

there (notice that (7.5) and (7.11) guarantee the assumption $M, N \leq q$ in (6.1)), and we obtain the bound

$$q^{-\varepsilon} S^\pm(M_1, M_2, M_3, M_4) \ll q^{\frac{1}{2} (\mu + \nu^* - 1 - u)} + q^{-\frac{1}{2} (\frac{1}{2} - u)} \ll q^{\frac{1}{2} (\mu + \nu^* - 5/4 + \delta)} + q^{-\frac{1}{2} (\mu + \nu^*) - \frac{1}{4}}.$$

We choose the value of $\delta$ by comparing the second term of the bound (7.14) with the first of the bound (7.3). Precisely, we let take

$$\delta = \frac{1}{2} \left( \frac{5}{4} - (\mu + \nu^*) \right)$$

and therefore we get

$$q^{-\varepsilon} S^\pm(M_1, M_2, M_3, M_4) \ll q^{\frac{1}{2} (\mu + \nu^* - \frac{5}{4})} + q^{-\frac{1}{2} (\mu + \nu^*) - \frac{1}{4}}.$$
under the assumption \( \mu + \nu^* \leq 5/4 \). This is indeed valid, by (7.11), since \( \eta \leq 1/16 \). Therefore, by (7.11), we find that

\[
q^{-(A+1)\varepsilon}S^{\pm}(M_1, M_2, M_3, M_4) \ll q^{-\frac{1}{16}+\eta} + q^{-\frac{1}{16}+\frac{3}{8}\eta} \ll q^{-\frac{1}{32}},
\]
as desired. \( \square \)

Again, the final exponent is the outcome of the following linear program: maximize

\[
\begin{aligned}
\min \left( \frac{\mu+\nu}{2}, -1, \max \left( \frac{\nu-\mu-1}{2}, \frac{\nu-\mu-1}{4} \right) \right), \quad \min \left( \frac{\mu+\nu^*-1}{2}, \max \left( -\frac{1}{8}, \frac{3}{8} - \frac{\mu+\nu}{2} \right) \right), \\
\min \left( \frac{\mu+\nu^*-1}{2}, \max \left( -\frac{\mu+\nu}{2}, \frac{1}{4} - \frac{\mu+\nu+\nu^*}{2} \right), \max \left( -\frac{\mu+\nu^*}{2}, \frac{1}{4} - \frac{\mu+\nu^*}{2} \right) \right)
\end{aligned}
\]

subject to the constraints

\[
0 \leq \mu \leq \nu, \quad \mu + \nu \leq 2, \quad 0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \quad \nu^* = 2 - \nu, \quad \mu_1 + \mu_2 + \mu_3 + \mu_4 = \mu + \nu^*
\]

(again it is clear a priori that the worst bound occurs at \( \nu' = \nu^* \)), and we get the numerical values \( \mu = 17/32, \nu = 45/32, (\mu_1, \mu_2, \mu_3, \mu_4) = (3/32, 3/32, 15/32, 15/32) \).

**Remark 7.1.** The same strategy, but with (3.1) instead of (4.4), gives a saving of \( q^{-1/24} \) if \( \theta = 0 \) in (3.1).

### 7.4. The mixed case

We consider the mixed case where \( f \) is cuspidal holomorphic and \( g = E \). Let \( \eta = 1/68 \). We will now prove Theorem 1.2 and emphasize again that the proof will be independent of (1.5). In this case, \( M, N \) are not symmetric, and so we will need to distinguish the cases where \( \mu \leq \nu \) and \( \mu > \nu \) on several occasions.

Firstly, applying (4.4), we see that (7.5) holds unless

\[
|\nu - \mu| \geq 1 - 4\eta
\]

which we assume from now on. In particular, the condition \( n \neq m \) is void. In order to avoid pathological cases, we derive first a simple, but useful auxiliary bound by applying the Voronoi formula to the longer of the two variables and estimating trivially. We detect the congruence

\[
B_{f,E}(M, N) = \sum_{a \pmod{q} \atop a \neq 0} \sum_{m,n} \lambda_f(m)d(n)e\left(\frac{a(m+n)}{q}\right)W_1\left(\frac{m}{M}\right)W_2\left(\frac{n}{N}\right).
\]

If, for instance, \( N \geq M \), then applying Lemma 2.3 to the \( n \)-sum yields a “main term”

\[
\frac{1}{q^{2}(MN)^{1/2}} \left( \int_{0}^{+\infty} (\log x + 2\gamma - 2\log q)W_2\left(\frac{x}{N}\right)dx \right) \sum_{m \geq 1} \lambda_f(m)r(m; q)W_1\left(\frac{m}{M}\right)
\]

where \( r(m; q) = q\delta_{q|m} - 1 \) is the Ramanujan sum, and two other terms are of the shape

\[
\frac{1}{q(MN^*)^{1/2}} \sum_{m,n \geq 1} \lambda_f(m)d(n)W_1\left(\frac{m}{M}\right) \frac{1}{N} \left( \overline{W_2,N}_{\sigma} \left(\frac{n}{q^2}\right) \right) S(m, \pm \sigma n; q)
\]

with \( \sigma \in \{\pm 1\} \) and the notation as in Lemma 2.4. A similar strategy (without a “main term”) can be applied if \( M > N \). Using Weil’s bound for Kloosterman sums and estimating trivially, we obtain the bound

\[
B_{f,E}(M, N) \ll q^{\varepsilon} \left( q \min(M, N) \right) \left( \frac{1}{\max(M, N)} \right)^{1/2}.
\]
In particular, (7.3) holds unless
\[(7.20)\quad |\nu - \mu| \leq 1 + 2\eta\]
which we assume from now on. We proceed to derive, by various methods depending on whether \(M > N\) or \(M \leq N\), more elaborate bounds that allow us to treat the range where (7.10) and (7.20) are satisfied.

7.4.1. The case \(M \leq N\). If \(N \leq q\), then by (7.16), we see that \(M \ll q^{4\eta + \epsilon}\), so that (4.3) suffices to prove (7.5). From now on we assume \(N \geq q\). In this case (4.2) is as good as (4.1), but independent of Deligne’s bound, and hence we can as before assume (7.6) without having appealed to (1.5).

First, we observe that (7.20) and (7.6) imply \(\mu \geq 1/2 - 2\eta > 2/5\), so that the second term in (7.4) is negligible. In the first term, we open the divisor function, apply smooth proportions of unity and are left with bounding the triple sum
\[(7.21)\quad C(M, N, N_1, N_2) := \frac{1}{\sqrt{MN}} \sum_{n_1, n_2 = \pm m(q)} \lambda_f(m)W_1(n_1/N_1)W_2(n_2/N_2)W_3(m/M)
\]
where
\[(7.22)\quad N_1N_2 = N, \quad N_1 \leq N_2\]
and \(W_1, W_2, W_3\) are (new) smooth, compactly supported weight functions satisfying (1.8). We can now apply Proposition 5.1 getting
\[(7.23)\quad C(M, N, N_1, N_2) \ll q^\epsilon \left(\frac{\sqrt{MN}}{q^{7/4}} + \min\left(\frac{\sqrt{MN}}{q^{7/4}}, \frac{N_1}{N_1^{1/2}}, \frac{N_2}{N_2^{1/2}}, \frac{N_1^{1/6}N_2^{1/2}}{(MN)^{1/2}}, \frac{M^{1/2}N_1^{1/2}}{N_2^{3/2}}, \frac{M^{1/2}N_2^{1/2}}{N_1^{3/2}}\right)\right),
\]
The term \(\sqrt{MN}/q^{7/4} \ll q^{-3/4 + \epsilon}\) is acceptable and can be dropped.

Alternatively, we can apply Poisson summation to both \(n_1, n_2\) (mimicking Voronoi summation on the original \(n\)-sum). We conclude from (7.6) and (7.16) that \(\mu \leq 1/2 + 2\eta < 3/5\), so that in particular \((m, q) = (n_1n_2, q) = 1\) in (7.21). We obtain
\[C(M, N, N_1, N_2) = \frac{1}{\sqrt{MN}} N \sum_{m, h_1, h_2} \lambda_f(m)W_3(m/M)W_1^\dagger(h_1N_1/q)W_2^\dagger(h_2N_2/q)S(\pm mh_1, h_2; q),
\]
where \(W_1^\dagger\) and \(W_2^\dagger\) denote the Fourier transforms of \(W_1\) and \(W_2\). Since \((q, m) = 1\) and the \(m\)-sum is sufficiently long, the contribution of the terms \(q | h_1h_2\) is negligible. After a smooth partition of unity, we are left with \(O(\log^2 q)\) terms of the form
\[C'(M, N, N_1, N_2) := \frac{1}{\sqrt{qMN_1N_2}} \sum_{m, h_1, h_2} \lambda_f(m)W_3(m/M)W_1(h_1N_1/q)W_2(h_2N_2/q)Kl_2(\pm mh_1h_2; q),
\]
where
\[(7.24)\quad N_1' \leq N_1^0, \quad N_2' \leq N_2^0, \quad N_1^0 = q/N_1, \quad N_2^0 = q/N_2,
\]
and \(W_1, W_2, W_3\) are (new) smooth, compactly supported weight functions satisfying (1.8). Notice that \(N_1' \geq N_1^0\). We can now use our results on multi-linear forms in Kloosterman sums as developed in Section 6. In particular, we can apply the bound (6.1) with \((M, N) \leftrightarrow (N_2', MN_1')\) in the notation of Theorem 6.1 or the bound (6.3) with \((M, N) \leftrightarrow (MN_2', N_1')\). This gives (using (1.6) several times)
\[(7.25)\quad C'(M, N, N_1, N_2) \ll q^\epsilon \frac{MN_1'N_2'}{\sqrt{MN_1N_2}}((N_2')^{-1/2} + q^{1/4}(MN_1')^{-1/2}), \quad \text{if } MN_1' \leq q,
\]
and

\[(7.26) \quad C'(M, N, N_1, N_2) \ll q^\varepsilon \frac{M N_1' N_2'}{\sqrt{q M N_1' N_2'}} (q^{1/4}(M N_2')^{-1/6}(N_1')^{-5/12}), \quad \text{if } MN_2' \leq (N_1')^2, \]

since the condition \(MN_1'N_2' \leq q^{3/2}\) and \(N_2', MN_2', N_1' \leq q\) are automatic by (7.6), (7.16), (7.22) and (7.26).

Combining all estimates we have derived so far, that is (4.1), (4.4), (7.19), (7.23), (7.25) and (7.26), we need to find the maximum of the piecewise linear function

\[
\min \left( \frac{\mu+\nu}{2} - 1, \max \left( \frac{\mu-\nu}{2}, \frac{\nu-\mu}{2} \right) \right), \quad 1+\mu-\nu,
\]

\[
\max \left( \frac{\mu+\nu}{2}, \frac{2\nu+\mu-2-\nu}{2}, \frac{\mu-\nu}{2}, \frac{2\nu+\mu-2-\nu}{2} + \frac{\mu+\nu}{2} \right), \quad \frac{\mu+2\nu+2\nu'-1-\nu_2'-\nu_2'-\mu}{2} + \frac{\mu+2\nu+2\nu'-1-\nu_2'-\nu_2'-\mu}{2} - \frac{\mu+\nu}{2}
\]

subject to the constraints

\[
0 \leq \mu \leq \nu, \quad \mu+\nu \leq 2, \quad \nu_1+\nu_2 = \nu, \quad 0 \leq \nu_1 \leq \nu_2, \quad 0 \leq \nu_1' \leq \nu_1 = 1-\nu_1, \quad 0 \leq \nu_2' \leq \nu_2 = 1-\nu_2.
\]

(Of course this expression can be simplified quite a bit.) A computer search shows that the maximum \(-1/68\) is attained at \(\mu = 161/306, \nu = 449/306, (\nu_1, \nu_2) = (9/17, 287/306)\) and (unsurprisingly) \(\nu_1' = \nu_1, \nu_2' = \nu_2\).

7.4.2. The case \(M \geq N\). We now assume \(\mu \geq \nu\) and observe that (7.16) and (7.20) are still in force, but we will not use (7.6) which is based on (4.1) (and hence on Deligne’s bound), but (4.3) instead.

In the present case it turns out to be most efficient to apply Voronoi summation in (7.17) in both variables. In the critical range this has essentially the effect of switching \(N\) and \(M\). The “main term” of the \(n\)-sum is given by (7.18) and trivially bounded by \(q^{-1+\varepsilon}\), which is acceptable. Applying Lemma 2.4 the usual partition of unity to the remaining terms in the Voronoi formula, we are left with bounding

\[
\tilde{B}_{f,E}^\pm(M, N) := \frac{1}{\sqrt{M^*N^*}} \sum_{m \equiv \pm n \pmod{q}} \lambda_f(m) d(n) W_1 \left( \frac{m}{M'} \right) W_2 \left( \frac{n}{N'} \right) - \frac{1}{q \sqrt{M^*N^*}} \sum_{m,n} \lambda_f(m) d(n) W_1 \left( \frac{m}{M'} \right) W_2 \left( \frac{n}{N'} \right)
\]

where

\[
M^* = \frac{q^2}{M}, \quad N^* = \frac{q^2}{N}, \quad M' \ll M^* q^\varepsilon, \quad N' \ll N^* q^\varepsilon
\]

and \(W_1, W_2\) are new weight functions satisfying (1.8). The second term is negligible unless \(M' \leq q^\varepsilon\), in which case it is trivially bounded by \(O(q^{\varepsilon-1}(M/N)^{1/2})\). By (7.20), this is \(O(q^{\varepsilon-1/2+\eta})\), which is acceptable. For the first term, we can apply Corollary 5.4. This leads to the linear program to
maximize
\[ \min \left( \frac{\mu + \nu}{2} - 1 + \frac{4}{3}, \max \left( \frac{\mu - \nu - 1}{2}, \frac{\mu - \nu - 1}{4} \right) \right), \frac{1 + \nu - \mu}{2}, \]
\[ \max \left( \frac{2\mu' - \mu^* - \nu^* + 1}{2}, \frac{2\nu + 2\mu' - 2 - \mu^* - \nu^*}{2} \right), \]
\[ \max \left( \frac{\mu' + \nu_2 + \frac{1}{2} - \nu^* - \mu^*}{2}, \frac{2\mu' + \nu_1 - \mu^* - \nu^*}{2}, \frac{2\nu + 2\mu' - 2 - \mu^* - \nu^*}{2} \right), \mu' + \nu_1 - \frac{\mu^* + \nu^*}{2}, \]
\[ \max \left( \min \left( \frac{2\nu - \frac{\mu + \nu^*}{2}, \frac{2}{3} \nu' + \nu_1 + \frac{1}{2} - \frac{1}{6} \mu' - \nu_2 - \frac{\mu^* + \nu^*}{2} \right), \right), \]
\[ \frac{2\mu' - \mu^* - \nu^* - \nu^*}{2}, \frac{2\mu' + \nu_1 - \mu^* - \nu^*}{2}, \mu' - \nu_2 - \frac{\mu^* + \nu^*}{2} \right) \]
subject to the constraints
\[ 0 \leq \nu \leq \mu, \quad \mu + \nu \leq 2, \quad \nu^* = 2 - \nu, \quad \mu^* = 2 - \mu, \]
\[ 0 \leq \nu' \leq \nu^*, \quad 0 \leq \mu' \leq \mu^*, \quad \nu_1 + \nu_2 = \nu', \quad 0 \leq \nu_1 \leq \nu_2. \]

A computer search shows that the maximum in this case is in fact a bit smaller, namely \(-1/64\), attained at \(\mu = 47/32, \nu = 17/32, (\nu_1, \nu_2) = (17/32, 15/16)\) and \(\nu' = \nu^*, \mu' = \mu^*\). This completes the proof of Theorem \([1, 2]\). \(\square\)

**Remark 7.2.** The reader may wonder why we use the “switching trick” at the beginning of Subsection \([7.4.2]\) and why the exponents in Subsection \([7.4.1]\) and \([7.4.2]\) are different. Young’s technique in the version of Proposition \([5.1]\) is only efficient if the divisor function is attached to the longer variable, which explains why we need to switch \(N\) and \(M\) at the beginning of the last subsection. Under this transformation of two applications of the Voronoi summation formula (one for each sum), the range \(MN \leq q^2\) becomes \(M^*N^* \geq q^2\). Of course, we are mostly interested in the case \(MN = q^2\) in which case the size conditions are essentially self-dual, but when it comes to optimizing exponents, the “worst case” of Subsection \([7.4.1]\) satisfies \(MN = q^{2-\delta}\) for \(\delta = 1/34\). For the dual problem, however, \(M^*N^* = q^{2-\delta}\) is forbidden, because we have the general assumption \(MN \ll q^{2+o(1)}\), therefore the exponent in Subsection \([7.4.2]\) becomes a little bit better.

**References**

[1] V. Blomer, *Shifted convolution sums and subconvexity bounds for automorphic L-functions*, Int. Math. Res. Not. (2004), 3905-3926.

[2] V. Blomer, G. Harcos, and Ph. Michel, *A Burgess-like subconvex bound for twisted L-functions. Appendix 2 by Z. Mao*, Forum Math. **19** (2007), no. 1, 61–105.

[3] V. Blomer and D. Miličević, *The second moment of twisted modular L-functions*, preprint arXiv:1404.7845 (2014).

[4] P. Deligne, *La conjecture de Weil, I*, Inst. Hautes Études Sci. Publ. Math. **43** (1974), 273–307.

[5] J.-M. Deshouillers and H. Iwaniec, *Kloosterman sums and Fourier coefficients of cusp forms*, Invent. Math. **70** (1982/83), no. 2, 219–288.

[6] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. Tricomi, *Higher transcendental functions II*, McGraw-Hill, 1953.

[7] É. Fouvry, *Sur le problème des diviseurs de Titchmarsh*, J. Reine Angew. Math. **357** (1985), 51–76.

[8] É. Fouvry, S. Ganguly, E. Kowalski, and Ph. Michel, *Gaussian distribution for the divisor function and Hecke eigenvalues in arithmetic progressions*, Comm. Math. Helv., to appear.

[9] É. Fouvry, E. Kowalski, and Ph. Michel, *Algebraic twists of modular forms and Hecke orbits*, GAFA, to appear.

[arXiv:1207.0617]

[10] **Algebraic trace functions over the primes**, Duke Math. J. **163** (2014), no. 9, 1683–1736.

[11] É. Fouvry, E. Kowalski, and P. Michel, *A study in sums of products*, preprint arXiv:1405.2293 (2014).

[12] É. Fouvry and Ph. Michel, *Sur certaines sommes d'expomientielles sur les nombres premiers*, Ann. Sci. École Norm. Sup. (4) **31** (1998), no. 1, 93–130.

[13] P. Gao, R. Khan, and G. Ricotta, *The second moment of Dirichlet twists of Hecke L-functions*, Acta Arith. **140** (2009), no. 1, 57–65.
8. Appendix: Mathematica Code

Section 7.2.

In[1] := Maximize[{Min[(m + n)/2 - 1, Max[(n - m - 1)/2, (n - m - 1)/4],
                     (n + nstar - 1)/2, (m + nstar - 1)/2 + Max[-nstar/2, 11/64
                     - 3 (m + nstar)/16]], 0 <= m, m <= n, m + n <= 2, nstar == 2 - n},
                    {m, n, nstar}]
Out[1] := {1
Section 7.3.

In[2] := Maximize[{Min[(m + n)/2 - 1, Max[(n - m - 1)/2, (n - m - 1)/4],
                     (m + nstar - 1)/2, (m + nstar - 1)/2 + Max[-1/8, 1/4 - (m1 + m3 + m4)/2],
                     (m + nstar - 1)/2 + Max[-m2/2, 1/4 - (m1 + m3 + m4)/2], (m + nstar - 1)/2
                     + Max[-(m1 + m2)/2, 1/4 - (m3 + m4)/2]], 0 <= m, m <= n, m + n <= 2,
                     0 <= m1, m1 <= m2, m2 <= m3, m3 <= m4, nstar == 2 - n,
                     m1 + m2 + m3 + m4 == m + nstar}, {m, n, nstar, m1, m2, m3, m4}]
Out[2] := {1
Section 7.4.1.

41
In[3] := Maximize[

{Min[(m + n)/2 - 1, Max[(n - m - 1)/2, (m + n - 1)/4],
(1 + m - n)/2, Max[(m + n - 1)/2, (2 n1 + m - 2 - n)/2],
Max[(m + n - 1)/2, (m + n - 1)/4, (m + n - 1)/2, (2 n1 + m - 2 - n)/2],
Max[1/4 - n1/2, (m - n)/2, (2 n1 + m - 2 - n)/2, (m + n1 - n)/2, (m + n1 - n)/2, (m + n1 - n)/2, (m + n1 - n)/2, (m + n1 - n)/2, (m + n1 - n)/2],
If[m + n1prime <= 1,
(2 m + 2 n1prime + 2 n2prime - 1 - n1circ - n2circ - m)/2
+ Max[-n2prime/2, 1/4 - (m + n1prime)/2], 10],
If[m + n2prime <= 2 n1prime, (2 m + 2 n1prime + 2 n2prime - 1
- n1circ - n2circ - m)/2 + 1/4 - 5 n1prime/12 - (m + n2prime)/6, 10]],
\[m \geq 0, n \geq m, m + n \leq 2, n1 + n2 == n, n1 \leq n2, n1 \geq 0,
n1prime \geq 0, n1prime \leq n1, n2prime \geq 0, n2prime \leq n2circ, n2circ == 1 - n2\],
\{m, n, n1, n2, n1prime, n2prime, n1circ, n2circ\}]

Out[3] := \{-1/68, \{m \rightarrow 161/306, n \rightarrow 449/306, n1 \rightarrow 9/17, n2 \rightarrow 287/306, n1prime \rightarrow 8/17, n2prime \rightarrow 19/306, n1circ \rightarrow 8/17, n2circ \rightarrow 19/306\}\}

Section 7.4.2.

In[4] := Maximize[

{Min[(m + n)/2 - 1, Max[(m + n - 1)/2, (m + n - 1)/4],
(1 + m - n)/2, Max[(2 mprime - mstar - nstar + 1)/ 2,
(2 n1 + 2 mprime - 2 - mstar - nstar)/2],
Max[(mprime + n2 + 1/2 - nstar - mstar)/ 2, (2 mprime + n1 - (mstar + nstar)/2),
2/3 nprime + n1 + 1/2 - 1/6 mprime - n2 - (mstar + nstar)/2],
(2 mprime - mstar - nstar)/ 2, (2 mprime + 2 n1 - 2 - mstar - nstar)/2,
2 mprime - n2 - (mstar + nstar)/2],
\[0 \leq n, n \leq m, m + n \leq 2, nstar == 2 - n, mstar == 2 - m, 0 \leq nprime, nprime \leq nstar, 0 \leq mprime, mprime \leq mstar, n1 + n2 == nprime, 0 \leq n1, n1 \leq n2\],
\{m, n, n1, n2, mprime, nprime, nstar, mstar\}]

Out[4] := \{-1/64, \{m \rightarrow 47/92, n \rightarrow 17/92, n1 \rightarrow 17/92, n2 \rightarrow 15/92, nprime \rightarrow 47/92, nstar \rightarrow 47/92, mprime \rightarrow 17/92, mstar \rightarrow 17/92\}\}