ON MOMENTS OF TWISTED L-FUNCTIONS

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

Abstract. We study the average of the product of the central values of two L-functions of modular forms f and g twisted by Dirichlet characters to a large prime modulus q. As our principal tools, we use spectral theory to develop bounds on averages of shifted convolution sums with differences ranging over multiples of q, and we use the theory of Deligne and Katz to prove new bounds on bilinear forms in Kloosterman sums with power savings when both variables are near the square root of q. When at least one of the forms f and g is non-cuspidal, we obtain an asymptotic formula for the mixed second moment of twisted L-functions 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 a power saving is proved under a conjectural estimate for certain bilinear forms in Kloosterman sums.

1. Introduction

1.1. **Moments of twisted L-functions.** This paper is motivated by the beautiful work of Matthew Young on the fourth moment of Dirichlet L-functions for prime moduli [29]: for a prime $q > 2$, let

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

where $\varphi^*(q) = q - 2$ is the number of primitive Dirichlet characters modulo q, and

$$L(\chi, s) = \sum_{n \geqslant 1} \frac{\chi(n)}{n^s}, \quad \mathfrak{Re}\, s > 1,$$

is the Dirichlet L-function. Young obtained the asymptotic formula

$$M_4(q) = P_4(\log q) + O(q^{-\frac{1}{80}(1-2\theta)+\varepsilon}) \tag{1.1}$$

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 [23]).

2010 *Mathematics Subject Classification.* 11M06, 11F11, 11L05, 11L40, 11F72, 11T23.

Key words and phrases. L-functions, moments, Eisenstein series, shifted convolution sums, Kloosterman sums, incomplete exponential sums, trace functions of ℓ-adic sheaves, Riemann Hypothesis over finite fields.

V. B. was supported by an ERC starting grant and the Volkswagen Foundation. É. F. thanks ETH Zürich, EPF Lausanne and the Institut Universitaire de France for financial support. Ph. M. was partially supported by the SNF (grant 200021-137488) and the ERC (Advanced Research Grant 228304). V. B., Ph. M. and E. K. were also partially supported by a DFG-SNF lead agency program grant (grant 200021L_153647). D. M. acknowledges partial support by the NSA (Grant H98230-14-1-0139), NSF (Grant DMS-1503629), and ARC (through Grant DP130100674). The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation herein.

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

$$M_{f,g}(q)=\frac{1}{\varphi^*(q)}\sum_{\substack{\chi\,(\mathrm{mod}\,q)\\ \chi\text{ primitive}}}L(f\otimes\chi,1/2)\overline{L(g\otimes\chi,1/2)},\tag{1.2}$$

where q is an integer with $q\not\equiv 2\,(\mathrm{mod}\,4)$ (since otherwise there are no primitive characters modulo q), f and g denote two fixed (holomorphic or non-holomorphic) Hecke eigenforms, not necessarily cuspidal, with respective Hecke eigenvalues $(\lambda_f(n))_{n\geqslant 1}$, $(\lambda_g(n))_{n\geqslant 1}$, and

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

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

$$E(z)=\frac{\partial}{\partial s}\Big|_{s=1/2}E(z,s),\quad\text{with } E(z,s)=\frac{1}{2}\sum_{(c,d)=1}\frac{y^s}{|cz+d|^{2s}}.\tag{1.3}$$

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

$$\lambda_E(n)=d(n)=\sum_{ab=n}1\tag{1.4}$$

(see [18, §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 (1.1).

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

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

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 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*

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

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 cuspidal Hecke eigenforms of level* 1*; if they are either both holomorphic or both Maaß we assume moreover that their root numbers satisfy $\varepsilon(f)\varepsilon(g)=1$. Let q be a prime number. Assume that* Conjecture 5.7 *below holds. Then for any $\varepsilon>0$, we have*

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

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

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

where $P_f(X)$ is an explicit polynomial of degree 1 with coefficients independent of q and leading coefficient $2L(\mathrm{sym}^2 f,1)\zeta(2)^{-1}$.

Remark 1.4. (1) If f and g are either both holomorphic or both Maaß with $\varepsilon(f)\varepsilon(g)=-1$, then $M_{f,g}(q)=0$ for parity reasons (see Remark 2.2).

(2) 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.)

Asymptotic formulas with a power saving for moments in families of L-functions are essential prerequisites for many applications including the techniques of amplification and mollification. Evaluation of moments becomes more difficult as the analytic conductor of the family increases relative to its size (for example, if considering moments involving higher powers of L-functions). In Theorems 1.1–1.3, the family is of size $|\mathcal{F}|\asymp q$ and the analytic conductor is $\asymp_{f,g} q^4\asymp|\mathcal{F}|^4$. As is well-known to experts, this is precisely the critical range of relative sizes at which most current analytic techniques fall just short of producing an asymptotic, and, in the few cases where an asymptotic available in this range, some deep input is typically required.

When $f=g$ and f is cuspidal, the moment (1.2) was studied by Stefanicki [28] and Gao, Khan and Ricotta [15]. In both cases, however, the error term gives only a saving of (at most) a small power of $\log q$ over the main term.

Finally, in relation to our Theorem 1.2 on mixed moments, we note the following asymptotic formula, recently established in [4]:

$$\frac{1}{\varphi^*(q)}\sum_{\substack{\chi\,(\mathrm{mod}\,q)\\ \chi\text{ primitive and even}}}L(f\otimes\chi,1/2)\overline{L(\chi,1/2)}=\frac{1}{2}L(f,1)+O_{f,\varepsilon}(q^{-1/64+\varepsilon}).$$

1.2. **Outline of the proof and bilinear forms in Kloosterman sums.** In this section we outline the proof of Theorems 1.1–1.3. This will also be an occasion to describe the two main ingredients of our approach, which are of independent interest: efficient treatment of shifted convolution sums (with particularly long shift variables) using the full power of spectral theory, and estimates of bilinear forms in Kloosterman sums (in particular when both variables are close to the square root of the conductor), which we treat using algebraic geometry.

Let q be a prime and f,g be Hecke eigenforms (cuspidal or equal to E) of level one. To simplify the forthcoming discussion, we assume in this section that both f and g satisfy the Ramanujan–Petersson conjecture (which is trivial for E and due to Deligne for holomorphic forms [5])

$$|\lambda_f(n)|,|\lambda_g(n)|\leqslant d(n)\tag{1.5}$$

for all $n\geqslant 1$, where $d(n)$ denotes the divisor function.

Using the functional equation of $L(f\otimes\chi,s)L(g\otimes\overline{\chi},s)$ (cf. (2.6)), we represent the central values as a converging series

$$L(f\otimes\chi,1/2)L(g\otimes\overline{\chi},1/2)=\sum\sum_{m,n\geqslant 1}\frac{\lambda_f(m)\lambda_g(n)\chi(m)\overline{\chi}(n)}{(mn)^{1/2}}V\Big(\frac{mn}{q^2}\Big)+\varepsilon(f,g,\chi)\sum\sum_{m,n\geqslant 1}\frac{\lambda_f(m)\lambda_g(n)\overline{\chi}(m)\chi(n)}{(mn)^{1/2}}V\Big(\frac{mn}{q^2}\Big)$$

for some essentially bounded function $V(t)$, which depends on the archimedean local factor $L_\infty(f\otimes\chi,s)L_\infty(g\otimes\overline{\chi},s)$ and which decays rapidly as $t\geqslant q^{\varepsilon}$ (for any fixed $\varepsilon>0$). An important feature is that this archimedean local factor and the root number $\varepsilon(f,g,\chi)=\pm1$ both depend on the character χ only through its *parity*, i.e., through $\chi(-1)=\pm1$. Therefore it is natural to average separately over even or odd characters, and then the root number $\epsilon(f,g,\chi)$ and the cutoff function V will be constant for all χ in the average.

The orthogonality of characters with given parity (given by (6.1) below) shows that theses averages are simple combinations of the quantities

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

The first 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

$$B^{\pm}_{f,g}(M,N)=\frac{1}{(MN)^{1/2}}\sum_{\substack{m\equiv\pm n\,(\mathrm{mod}\,q)\\ m\neq n}}\lambda_f(m)\lambda_g(n)W_1\Big(\frac{m}{M}\Big)W_2\Big(\frac{n}{N}\Big)-\frac{1}{q(MN)^{1/2}}\sum_{m,n}\lambda_f(m)\lambda_g(n)W_1\Big(\frac{m}{M}\Big)W_2\Big(\frac{n}{N}\Big),$$

where $M,N\geqslant 1$, $MN\leqslant q^{2+o(1)}$ and W_1, W_2 are test functions satisfying (1.14) below.

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 [29]). We denote by $\mathrm{ET}^{\pm}_{f,g}(M,N)$ the remaining part of $B^{\pm}_{f,g}(M,N)$ in all cases (thus $\mathrm{ET}^{\pm}_{f,g}(M,N)=B^{\pm}_{f,g}(M,N)$ 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

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

for some absolute constant $\eta>0$.

The trivial bound

$$\mathrm{ET}^{\pm}_{f,g}(M,N)\ll_{\varepsilon}(MN)^{1/2}q^{-1+o(1)},$$

implies that we may assume that MN is close to q^2 when estimating $\mathrm{ET}^{\pm}_{f,g}(M,N)$. For simplicity, we assume that $MN=q^{2+o(1)}$ in this outline.

At this point, the analysis depends on the relative ranges of M and N. There are essentially two cases to consider which are handled by two very different methods.

Balanced configuration and the shifted convolution problem. First, when the sizes of M and N are relatively close to each other, we interpret the congruence condition as an equality over the integers:

$$0\neq m\mp n\equiv 0\,(\mathrm{mod}\,q)\Leftrightarrow\ m\mp n=qr,\ \text{for some } r\neq 0. \tag{1.6}$$

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 [25] for an overview). The most powerful methods known today involve the spectral theory of automorphic forms and usually depend on bounds towards the Ramanujan–Petersson conjecture (not solely for f and g but for all automorphic forms of level 1).

The typical estimate one can obtain is

$$\mathrm{ET}^{\pm}_{f,g}(M,N) \ll \frac{q^{o(1)}}{q^{1/2-\theta}}\left(\frac{M}{N}+\frac{N}{M}\right)^{1/2}. \tag{1.7}$$

(see for instance [29, Theorem 3.3] when $f=g=E$), where we recall that $\theta=7/64$ is the best known approximation to the Ramanujan–Petersson conjecture and depends among other things on the automorphy of the symmetric fourth power of GL(2)-automorphic representations [23]. This bound 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

$$\max(M,N)\geqslant q^{3/2-\theta-\delta} \tag{1.8}$$

for $\delta>0$ fixed as small as we need.

Ideally (under the Ramanujan–Petersson conjecture), it would remain to handle the range

$$\max(M,N)\geqslant q^{3/2-\delta}, \tag{1.9}$$

which we will eventually be able to do (with a small but fixed $\delta>0$) using an alternative set of methods described below and developed in detail in Section 5. Unfortunately and despite the fact that the current value of θ is quite small, these methods do not seem always capable to cover the range (1.8): specifically, to prove Theorem 1.3, with some positive exponent in place of $1/144$, using the bound (1.7) and the results of Section 5, one would need to have $\theta<1/40$. In addition to such a result being unavailable as yet, an argument that does not seriously depend of the numerical value of θ is of interest on its own.

The removal of this dependence on the Ramanujan–Petersson conjecture is precisely one of the main achievements in [3] when f and g are both cuspidal. We adapt this method, which further refines the shifted convolution sum estimates into the very long shift variable range (in particular by exploiting the average over r in (1.6)), to obtain similar uniform estimates also when f or g is the Eisenstein series E (see Section 3, Theorem 3.2). The resulting bound is as follows:

Bound A. *The error term* $\mathrm{ET}^{\pm}_{f,g}(M,N)$ *satisfies*

$$\mathrm{ET}^{\pm}_{f,g}(M,N) \ll q^{o(1)}\left(\frac{1}{q^{1/2}}\Big(\frac{M}{N}+\frac{N}{M}\Big)^{1/2}+\frac{1}{q}\Big(\frac{M}{N}+\frac{N}{M}\Big)\right)^{1/2}+q^{-1/2+\theta+o(1)}. \tag{1.10}$$

The estimate (1.10) is weaker than (1.7) for $M=N$, but crucially it is non-trivial in the full range complementary to (1.9), thereby acting essentially as if $\theta=0$.

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 the remaining range (1.9), which eventually requires us to consider different cases separately.

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

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

where the "dual" length N' satisfies

$$N'\leqslant N^*:=q^2/N$$

and $\widetilde{W}_2$ is another smooth function satisfying (1.14). (Here $\mathrm{Kl}_2(a;q)$ denotes the Kloosterman sum modulo q, normalized so that $|\,\mathrm{Kl}_2(a;q)|\leqslant 2$ by Weil's bound, see (1.15).) 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

$$N'=N^*=q^2/N=q^{o(1)}M,$$

which we now assume.

Such sums are very special cases of bilinear sums in Kloosterman sums

$$B(\mathrm{Kl}_2,\boldsymbol{\alpha}_U,\boldsymbol{\beta}_V)=\sum_{u\leqslant U,}\sum_{v\leqslant V}\alpha_u\beta_v\,\mathrm{Kl}_2(auv;q), \tag{1.11}$$

for $(a,q)=1$, $U,V\leqslant q$ and some complex numbers $(\alpha_u)_{u\leqslant U}$, $(\beta_v)_{v\leqslant V}$. The "trivial" bound (which follows from Weil's bound) is

$$\sum_{u\leqslant U,}\sum_{v\leqslant V}\alpha_u\beta_v\,\mathrm{Kl}_2(auv;q)\ll\|\alpha\|_2\|\beta\|_2(UV)^{1/2},$$

and a natural question is whether one can improve that bound at least for suitable values of the parameters U,V and/or $(\alpha_u)_u$, $(\beta_v)_v$.

Specialized to our current situation (taking $U=M$, $V=N^*=q^{o(1)}M$), the trivial bound yields

$$C^{\pm}(M,N^*)=\frac{1}{(qMN^*)^{1/2}}B(\mathrm{Kl}_2,\boldsymbol{\alpha}_M,\boldsymbol{\beta}_{N^*})\ll q^{o(1)}(MN^*/q)^{1/2}=q^{3/2+o(1)}/N,$$

which is satisfactory as soon as $N\geqslant q^{3/2+o(1)}$. Hence, we are left with a single critical range

$$M=q^{1/2+o(1)},\ \ N^*=q^{1/2+o(1)},$$

which is called the *Pólya–Vinogradov range* (in analogy with, say, character sums modulo q, where sums of length $q^{1/2+o(1)}$ are precisely the longest sums for which an application of the Pólya–Vinogradov inequality does not shorten the sum). It is now sufficient to improve the trivial bound on (1.11) in this most stubborn range.

The cuspidal case. When f and g are both cuspidal, the range of the variables is so short that we don't see a way to exploit the automorphic origin of the sequence $(\lambda_f(m))_n$ and $(\lambda_g(n))_n$. Instead, based on earlier work of Fouvry and Michel [14], we prove in Proposition 5.5 the following bound conditional on a square-root cancellation bound for certain complete 3-dimensional sums of products of Kloosterman sums:

Bound B. *Assume Conjecture 5.7. Then, for* $(a,q)=1$ *and* U,V *satisfying*

$$q^{\frac14}\leqslant UV\leqslant q^{\frac54}\ \text{ and }\ 1\leqslant U\leqslant q^{\frac14}V$$

one has

$$\sum_{u\leqslant U,v\leqslant V}\sum\alpha_u\beta_v\,\mathrm{Kl}_2(auv;q)\leqslant q^{o(1)}\|\alpha\|_2\|\beta\|_2(UV)^{1/2}\big(U^{-\frac12}+q^{\frac{11}{64}}(UV)^{-\frac{3}{16}}\big). \tag{1.12}$$

In the Pólya-Vinogradov range $U \asymp V \asymp q^{1/2}$, the above bound saves a factor $q^{1/64}$ over the trivial bound, leading to Theorem 1.3.

Non-cuspidal cases. If f or g is the Eisenstein series E, we can exploit the decomposition of the Hecke eigenvalues $d(n) = (1 \star 1)(n)$ as a Dirichlet convolution to obtain our unconditional results.

First assuming that $g = E$, the bilinear form (1.11) transforms into trilinear forms with two smooth variables of the type

$$\sum_{m \asymp M} \sum_{n_1 \asymp N_1} \sum_{n_2 \asymp N_2} \lambda_f(m) \operatorname{Kl}_2(\pm m n_1 n_2; q) \quad \text{ with } N_1 N_2 = N^*.$$

We can then group the variables differently to form a new variable (v say) whose length V is *larger* than the Pólya-Vinogradov range $q^{1/2}$ and apply the following bound (see (5.1) of Theorem 5.1):

$$B(\operatorname{Kl}_2, \boldsymbol{\alpha}_U, \boldsymbol{\beta}_V) \ll q^{o(1)} (UV)^{1/2} \|\alpha\|_2 \|\beta\|_2 \big(U^{-1/2} + q^{1/4} V^{-1/2}\big).$$

This bound is non-trivial if $U \geqslant q^{o(1)}$ and $V \geqslant q^{1/2+o(1)}$.

If such a grouping is not possible (because N_1 or N_2 is small), then we are essentially in a situation corresponding to bilinear forms where both variables are in the Pólya-Vinogradov range but one of them is *smooth*. The main new result of this paper regarding bilinear sums of Kloosterman sums (see (5.3) of Theorem 5.1) makes it possible to handle this case:

Bound C. *For $(a, q) = 1$, and U, V satisfying*

$$1 \leqslant U, V \leqslant q, \quad UV \leqslant q^{3/2}, \quad U \leqslant V^2$$

one has

$$\sum_{u \leqslant U,\, v \sim V} \sum \alpha_u \operatorname{Kl}_2(auv; q) \ll q^{o(1)} \|\alpha\|_2 U^{1/2} V \big(q^{\frac{1}{4}} U^{-\frac{1}{6}} V^{-\frac{5}{12}}\big). \tag{1.13}$$

In the critical range $U \asymp V \asymp q^{1/2}$, the above bound saves a factor $q^{1/24}$ over the trivial bound and this combined with arguments from [29] eventually leads to the exponent $1/68$ in the error term of Theorem 1.2.

The double Eisenstein case. 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

$$\sum_{\substack{m_1, m_2, n_1, n_2 \\ m_i \asymp M_i, n_i \asymp N_i}} \cdots \sum \operatorname{Kl}_2(\pm m_1 m_2 n_1 n_2; q),$$

where

$$M_1 M_2 = M \asymp q^{1/2+o(1)}, \;\; N_1 N_2 = N^* \asymp q^{1/2+o(1)}.$$

We now have more possibilities for grouping variables. Especially when two of the variables (say m_2 and n_2) are small, the grouping of m_1, n_1 into a single long variable $n = m_1 n_1$ weighted by a divisor-like function $(1_{M_1} \star 1_{N_1})(n)$ makes it possible to use the general results of [12] which provide quite strong (unconditional) bounds for such types of sums (see (5.5) of Theorem 5.1).

The first step in proving (1.12) and (1.13) in Section 5 is an elaboration of Karatsuba's variant of Burgess's method along the lines of the work of Fouvry and Michel [14]. Using this, bounds for short bilinear sums such as (1.11) (strong in the critical ranges for us) can be obtained if one has upper bounds of the expected square root strength for multivariable *complete* exponential sums. We prove such bounds in the situation of (1.13) by using the Riemann Hypothesis over finite fields of Deligne [5, 6] and a general criterion due to Hooley [17] and Katz [20].

All precise statements concerning bilinear sums of Kloosterman sums above are found in Theorem 5.1 in Section 5.1.

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 §3 with either Young'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) One of the known technical difficulties in the mixed case is that the variables m and n (and their corresponding ranges) do not play the same role. However, applying the Voronoi summation formula *twice* (in different variables) allows us to essentially exchange the roles of m and n in critical ranges (roughly speaking, turning $B^{\pm}_{f,g}(M,N)$ via $C^{\pm}(M,N')$ to a sum of $B^{\pm}_{f,g}(M',N')$, with $M' \leqslant q^2/M$, $N' \leqslant q^2/N$); see Subsection 6.4.2.

(4) When q is suitably *composite*, a bilinear form in Kloosterman sums (1.11) has been estimated in [3] by developing a large sieve-type bound for Kloosterman sums using a variant of q-van der Corput method.

The structure of the paper is as follows. We collect some standard facts in Section 2. As we have seen, the proof depends on two crucial ingredients, the treatment of the shifted convolution sum problem and estimates of the bilinear sums of Kloosterman sums; these are the topics of Sections 3 and 5, respectively, while Section 4 recalls briefly M. Young's method. Finally, Section 6 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. We would also thank the referees for a careful reading and useful suggestions.

1.3. **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

$$W^{(j)}(x) \ll_{j,\varepsilon} q^{j\varepsilon}, \tag{1.14}$$

for any $\varepsilon > 0$ and any $j \geqslant 0$, the implied constant depending on ε and j (but not on q). Sometimes even the stronger bounds $W^{(j)}(x) \ll_j 1$ hold.

From time to time, we will use the ε-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 ε in (1.14) may be different for different functions W).

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

$$S(a,b;c) = \sum_{\substack{d \,(\mathrm{mod}\, c)\\ (d,c)=1}} e_c(ad + b\bar{d})$$

the usual Kloosterman sum, and we also write

$$\mathrm{Kl}_2(a;c) = \frac{1}{\sqrt{c}} S(a,1;c) \tag{1.15}$$

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., [9, Lemme 2]):

Lemma 1.6. *There exists a smooth non-negative function $W(x)$ supported on $[1/2,2]$ and satisfying (1.14) such that*

$$\sum_{k \geqslant 0} W\Big(\frac{x}{2^k}\Big) = 1$$

for any $x \geqslant 1$.

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.

2.1. **Functional equations for Dirichlet L-functions.** Let χ 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 $\mathbf{C}$ and satisfies a functional equation which we now recall (see [19, Theorem 4.15] for instance): let

$$\mathfrak{a}(\chi)=\mathfrak{a}=\frac{1-\chi(-1)}{2}=\begin{cases}0 & \text{if } \chi(-1)=1,\\ 1 & \text{if } \chi(-1)=-1,\end{cases} \tag{2.1}$$

and let

$$\Lambda(\chi,s)=q^{s/2}L_\infty(\chi,s)L(\chi,s),\quad L_\infty(\chi,s)=\pi^{-\frac{s}{2}}\Gamma\Big(\frac{s+\mathfrak{a}}{2}\Big)$$

be the completed L–function. For $s\in\mathbf{C}$ one has

$$\Lambda(\chi,s)=\varepsilon(\chi)\Lambda(\overline{\chi},1-s),$$

where

$$\varepsilon(\chi)=i^{-\mathfrak{a}}\varepsilon_\chi,\ \varepsilon_\chi=\frac{\tau(\chi)}{\sqrt{q}},\quad \tau(\chi)=\sum_{x \bmod q}\chi(x)e(x/q).$$

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

$$\Lambda(E\otimes\chi,s)=\chi(-1)\varepsilon_\chi^2\Lambda(E\otimes\overline{\chi},1-s).$$

2.2. **Cusp forms.** We now describe the functional equation when E is replaced by a cuspidal Hecke eigenform (holomorphic or Maaß) f for the group $\Gamma_0(1)=\mathrm{SL}(2,\mathbf{Z})$. Let $(\lambda_f(n))_{n\geqslant1}$ be the sequence of Hecke eigenvalues of f or equivalently the coefficients of its Hecke L-function:

$$L(f,s):=\sum_{n\geqslant1}\frac{\lambda_f(n)}{n^s}=\prod_p\Big(1-\frac{\lambda_f(p)}{p^s}+\frac{1}{p^{2s}}\Big)^{-1},\quad \mathfrak{Re}\, s>1.$$

The numbers $\lambda_f(n)$ satisfy the multiplicativity relations

$$\lambda_f(m)\lambda_f(n)=\sum_{d\mid(m,n)}\lambda_f\left(\frac{mn}{d^2}\right),\ \lambda_f(mn)=\sum_{d\mid(m,n)}\mu(d)\lambda_f\left(\frac{m}{d}\right)\lambda_f\left(\frac{n}{d}\right).$$

If f is holomorphic, the Ramanujan–Petersson conjecture is known by the work of Deligne [5], and one has

$$|\lambda_f(n)|\leqslant d(n) \tag{2.2}$$

where $d(n)$ is the divisor function. If f is a Maaß form with Laplace eigenvalue $\lambda_f(\infty)=(\frac{1}{2}+it)(\frac{1}{2}-it)$, it follows from the work of Kim-Sarnak [23] that

$$|\lambda_f(n)|\leqslant d(n)n^\theta \text{ for } \theta=7/64$$
(2.3) and
$$\text{either } t\in\mathbf{R} \text{ or } t\in i\mathbf{R} \text{ with } |t|\leqslant\theta.$$

The Ramanujan–Petersson conjecture (that one could take $\theta = 0$ in the above bounds) is at least true on average in the following sense: for any $x \geqslant 1$ and any $\varepsilon > 0$, one has

$$\sum_{n \leqslant x} |\lambda_f(n)|^2 \ll_{\varepsilon,f} x^{1+\varepsilon}. \tag{2.4}$$

Of course, this bound holds also for the divisor function in place of λ_f.

2.3. **Functional equations for twisted L-functions.** For a primitive Dirichlet character χ of prime modulus q, the sequence $(\lambda_f(n)\chi(n))_{n\geqslant 1}$ is the sequence of coefficients of the Hecke L-function of a cusp form $f \otimes \chi$ relative to the group $\Gamma_0(q^2)$ with nebentypus χ^2 (see [19, Propositions 14.19 & 14.20], for instance). The twisted L-function

$$L(f \otimes \chi, s) := \sum_{n \geqslant 1} \frac{\lambda_f(n)\chi(n)}{n^s} = \prod_p \Big(1 - \frac{\lambda_f(p)\chi(p)}{p^s} + \frac{\chi^2(p)}{p^{2s}}\Big)^{-1}, \quad \mathfrak{Re}\, s > 1,$$

has an analytic continuation to $\mathbf{C}$ and satisfies the functional equation (see e.g. [19, Theorem 14.17, Proposition 14.20])

$$\Lambda(f \otimes \chi, s) = \varepsilon(f \otimes \chi)\Lambda(f \otimes \overline{\chi}, 1 - s),$$

where

$$\Lambda(f \otimes \chi, s) = q^s L_\infty(f \otimes \chi, s) L(f \otimes \chi, s),$$

$$L_\infty(f \otimes \chi, s) = \begin{cases} \Gamma_{\mathbf{C}}(\frac{k-1}{2} + s) \text{ if } f \text{ is holomorphic of weight } k, \\ \Gamma_{\mathbf{R}}(s + it + \mathfrak{a})\Gamma_{\mathbf{R}}(s - it + \mathfrak{a}) \text{ if } f \text{ is a Maaß form with eigenvalue } (\frac{1}{2} + it)(\frac{1}{2} - it) \end{cases}$$

with

$$\Gamma_{\mathbf{R}}(s) = \pi^{-s/2}\Gamma(s/2),\ \Gamma_{\mathbf{C}}(s) = (2\pi)^{-s}\Gamma(s),$$

and the root number $\varepsilon(f \otimes \chi)$ is defined by

$$\varepsilon(f \otimes \chi) = \begin{cases} \varepsilon(f)\varepsilon_\chi^2, & \text{if } f \text{ is holomorphic,} \\ \varepsilon(f)\chi(-1)\varepsilon_\chi^2, & \text{if } f \text{ is a Maaß form,} \end{cases} \tag{2.5}$$

where $\varepsilon(f) = \pm 1$ is the root number of $L(f, s)$. Consequently one has the following equations:

Lemma 2.1. *Let f, g be either cuspidal Hecke eigenforms of level 1 or the non-holomorphic Eisenstein series E. Then one has, setting $\varepsilon(E) = 1$,*

$$\Lambda(f \otimes \chi, s)\Lambda(g \otimes \overline{\chi}, s) = \varepsilon(f, g, \chi)\Lambda(f \otimes \overline{\chi}, 1 - s)\Lambda(g \otimes \chi, 1 - s),$$

where

$\varepsilon(f, g, \chi) = \varepsilon(f)\varepsilon(g)$ for f and g both holomorphic or both non-holomorphic,

$\varepsilon(f, g, \chi) = \chi(-1)\varepsilon(f)\varepsilon(g)$ for f holomorphic and g non-holomorphic.

Remark 2.2. Observe that the root number $\varepsilon(f, g, \chi)$ depends on χ at most through its parity $\chi(-1)$ and does not depend on χ at all if f and g are both holomorphic or both non-holomorphic. We will therefore denote it by $\varepsilon(f, g, \pm 1)$ where $\pm 1 = \chi(-1)$.

Next, we state a standard approximate functional equation. We have (similarly as in [19, Theorem 5.3]) the formula

$$L(f \otimes \chi, 1/2)\overline{L(g \otimes \chi, 1/2)} = \sum_{m,n \geqslant 1} \frac{\lambda_f(m)\lambda_g(n)}{(mn)^{1/2}} \chi(m)\bar{\chi}(n) V_{f,g,\pm 1}\left(\frac{mn}{q^2}\right) \\ + \varepsilon(f, g, \pm 1) \sum_{m,n \geqslant 1} \frac{\lambda_f(n)\lambda_g(m)}{(mn)^{1/2}} \chi(m)\bar{\chi}(n) V_{f,g,\pm 1}\left(\frac{mn}{q^2}\right) \tag{2.6}$$

and

$$V_{f,g,\pm 1}(x)=\frac{1}{2\pi i}\int_{(2)}\frac{L_\infty(f\otimes\chi,1/2+s)L_\infty(g\otimes\overline{\chi},1/2+s)}{L_\infty(f\otimes\chi,1/2)L_\infty(g\otimes\overline{\chi},1/2)}x^{-s}\frac{ds}{s}. \tag{2.7}$$

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

2.4. **Voronoi summation and Bessel functions.** The next lemma is a version of the Voronoi formula.

Lemma 2.3. [10, 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 $]0,\infty[$. Let $a_f(n)$ denote Hecke eigenvalues of a Hecke eigenform f of level 1. Then*

$$\begin{aligned}&\sum_{n\geqslant 1}a_f(n)W(n)e\Big(\frac{an}{c}\Big)\\&=\delta_{f=E}\frac{1}{c}\int_0^{+\infty}(\log x+2\gamma-2\log c)W(x)dx+\frac{1}{c}\sum_{\pm}\sum_{n\geqslant 1}a_f(n)\widetilde{W}_\pm\Big(\frac{n}{c^2}\Big)e\Big(\mp\frac{\overline{a}n}{c}\Big),\end{aligned}$$

where γ is Euler's constant and the transforms $\widetilde{W}_\pm:(0,\infty)\to\mathbb{C}$ of W are defined by

$$\widetilde{W}_\pm(y)=\int_0^\infty W(u)\mathcal{J}_\pm(4\pi\sqrt{uy})du$$

with

$$\mathcal{J}_+(x)=2\pi i^k J_{k-1}(x),\quad \mathcal{J}_-(x)=0$$

if f is holomorphic of weight k and

$$\mathcal{J}_+(x)=-\frac{\pi}{\cosh(\pi t)}(Y_{2it}(x)+Y_{-2it}(x))=\frac{\pi i}{\sinh(\pi t)}(J_{2it}(x)-J_{-2it}(x)),$$

$$\mathcal{J}_-(x)=4\cosh(\pi t)K_{2it}(x)$$

if f is non-holomorphic with spectral parameter t (in particular $t=0$ if $f=E$).

For the basic facts concerning the Bessel functions J, Y and K see [18, Appendix B]. In particular, $\mathcal{J}_-$ is rapidly decaying:

$$\mathcal{J}_-(x)\ll x^{-1/2}e^{-x} \tag{2.8}$$

for $x\geqslant 1$ (and fixed $t\in\mathbb{R}$). At one point we shall need the uniform bounds

$$J_{it}(x)\ll e^{|t|/2}(|t|+x)^{-1/2},\quad t\in\mathbf{R},x>0 \tag{2.9}$$

and

$$J_k(x)\ll\min\big(k^{-1/3},|x^2-k^2|^{-1/4}\big),\quad k>0,x>0. \tag{2.10}$$

The first bound follows from the power series expansion [16, 8.402] for $x<t^{1/3}$ (say) and from the uniform expansion [8, 7.13 formula (17)] otherwise. The second bound follows also from the power series expansion for $x<k^{1/3}$ and from Olver's uniform expansion [27, (4.24)].

Integration by parts in combination with [16, 8.472.3] shows the formula

(2.11)

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

for $j \in \mathbb{C}$ 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 [16, 8.451.2]

$$Y_{it}(x) = F_+(x)e^{ix} + F_-(x)e^{-ix} + O(x^{-A}) \tag{2.12}$$

for $x \geqslant 1$, $t \in \mathbb{R}$ with smooth, non-oscillating functions $F_\pm(x)$ (depending on t) satisfying

$$x^j F_\pm^{(j)}(x) \ll_{j,t} x^{-1/2}.$$

Finally, we consider the decay properties of the Bessel transforms $\widetilde{W}, \widetilde{W}_\pm$.

Lemma 2.4. *Let W be a smooth function compactly supported in $[1/2, 2]$ and satisfying (1.14). In the non-holomorphic case set $\vartheta = \mathfrak{Re}\, it$, otherwise set $\vartheta = 0$. For $M \geqslant 1$ let $W_M(x) = W(x/M)$. For any ε, for any $i, j \geqslant 0$ and for all $y > 0$, we have*

$$y^j \widetilde{(W_M)}_\pm^{(j)}(y) \ll_{i,j,\varepsilon} M(1+My)^{j/2}\big(1+(My)^{-2\vartheta-\varepsilon}\big)\big(1+(My)^{1/2}q^{-\varepsilon}\big)^{-i}.$$

In particular, the functions $\widetilde{(W_M)}_\pm(y)$ decay rapidly when $y \gg q^{3\varepsilon}/M$.

Proof. We differentiate j times under the integral sign, followed by i applications of (2.11) (or analogous formulae for K and J) with $z = 0$. Then we estimate trivially, using $\mathcal{B}_\nu'(x) = \frac{1}{2}(\pm\mathcal{B}_{\nu-1} - \mathcal{B}_{\nu+1})$ for $\mathcal{B} \in \{J, Y, K\}$ and the simple bounds

$$J_\nu(x) \ll_\nu \begin{cases} 1, & x \geqslant 1, \\ x^{|\mathfrak{Re}\,\nu|}, & x < 1, \nu \in \mathbb{N}_0, \end{cases} \qquad Y_\nu(x), K_\nu(x) \ll_\nu \begin{cases} 1, & x \geqslant 1, \\ (1+\log|x|)x^{-|\mathfrak{Re}\,\nu|}, & x < 1. \end{cases}$$

□

2.5. **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 $\phi : [0, \infty) \to \mathbf{C}$ satisfying $\phi(0) = \phi'(0) = 0$, $\phi^{(j)}(x) \ll (1+x)^{-3}$ for $0 \leqslant j \leqslant 3$:

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

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

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

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

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

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. [2, Theorem 2]).

Lemma 2.5 (Kuznetsov formula)**.** *Let ϕ be as in the previous paragraph, and let $a, b > 0$ be integers. Then*

$$\sum_{c \geqslant 1} \frac{1}{c} S(a,b;c) \phi\left(\frac{4\pi\sqrt{ab}}{c}\right) = \sum_{\substack{k \geqslant 2 \\ k \text{ even}}} \sum_{f \in \mathcal{B}_k} \dot{\phi}(k) \Gamma(k) \sqrt{ab} \varrho_f(a) \varrho_f(b)$$
$$+ \sum_{f \in \mathcal{B}} \tilde{\phi}(t_f) \frac{\sqrt{ab}}{\cosh(\pi t_f)} \varrho_f(a) \varrho_f(b) + \frac{1}{4\pi} \int_{-\infty}^{\infty} \tilde{\phi}(t) \frac{\sqrt{ab}}{\cosh(\pi t)} \varrho(a,t) \varrho(b,t) dt$$

and

$$\sum_{c \geqslant 1} \frac{1}{c} S(a,-b;c) \phi\left(\frac{4\pi\sqrt{ab}}{c}\right) = \sum_{f \in \mathcal{B}} \check{\phi}(t_f) \frac{\sqrt{ab}}{\cosh(\pi t_f)} \varrho_f(a) \varrho_f(-b)$$
$$+ \frac{1}{4\pi} \int_{-\infty}^{\infty} \check{\phi}(t) \frac{\sqrt{ab}}{\cosh(\pi t)} \varrho(a,t) \varrho(-b,t) dt.$$

Often an application of the Kuznetsov formula is followed directly by an application of the large sieve inequalities of Deshouillers-Iwaniec [7, Theorem 2].

Lemma 2.6 (Spectral large sieve)**.** *Let $T, M \geqslant 1$, and let (a_m), $M \leqslant m \leqslant 2M$, be a sequence of complex numbers. Then all three quantities*

$$\sum_{\substack{2 \leqslant k \leqslant T \\ k \text{ even}}} \Gamma(k) \sum_{f \in \mathcal{B}_k} \Big| \sum_m a_m \sqrt{m} \varrho_f(m) \Big|^2, \qquad \sum_{\substack{f \in \mathcal{B} \\ |t_f| \leqslant T}} \frac{1}{\cosh(\pi t_f)} \Big| \sum_m a_m \sqrt{m} \varrho_f(\pm m) \Big|^2,$$
$$\int_{-T}^{T} \frac{1}{\cosh(\pi t)} \Big| \sum_m a_m \sqrt{m} \varrho(\pm m, t) \Big|^2 dt$$

are bounded by

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

Finally we quote a special case of [3, Theorem 13] 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^{θ}.

Lemma 2.7. *Let $s \in \mathbf{N}$, $R, T \geqslant 1$, and let $\alpha(r)$, $R \leqslant r \leqslant 2R$, be any sequence of complex numbers with $|\alpha(r)| \leqslant 1$. Then*

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

3. Shifted convolution sums

3.1. **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 (1.14), and for $M,N\geqslant 1$, we denote

$$\mathrm{ET}^{\pm}_{f,g}(M,N)=\frac{1}{(MN)^{1/2}}\sum_{\substack{m\equiv\pm n\,(\mathrm{mod}\,q)\\ m\neq n}}\lambda_f(m)\lambda_g(n)W_1\Big(\frac{m}{M}\Big)W_2\Big(\frac{n}{N}\Big)$$
$$-\frac{1}{q(MN)^{1/2}}\sum_{(mn,q)=1}\lambda_f(m)\lambda_g(n)W_1\Big(\frac{m}{M}\Big)W_2\Big(\frac{n}{N}\Big)-\delta_{f=g=E}\mathrm{MT}^{od,\pm}_{E,E}(M,N),$$

where $\mathrm{MT}^{od,\pm}_{E,E}(M,N)$ is the off-diagonal main term discussed by Young in [29, §6, §8].

We start with the following simple bounds which follow either from the validity of the Ramanujan–Petersson conjecture for the forms in question or the unconditional individual bound (2.3) or the averaged bound (2.4) together with a bound for $\mathrm{MT}^{od,\pm}_{E,E}(M,N)$ given in [29]Lemma 6.1.

Define

$$\theta_g=\begin{cases}0, & \text{if } g=E \text{ or is holomorphic},\\ \theta=7/64, & \text{otherwise},\end{cases}$$

and similarly θ_f.

Proposition 3.1. *Let f,g be either E or cuspidal Hecke eigenforms of level 1. Let q be a prime and assume that W_1,W_2 satisfy (1.14). We have for $1\leqslant M\leqslant N$ the bound*

$$\mathrm{ET}^{\pm}_{f,g}(M,N)\ll\left(N^{\theta_g}\frac{(MN)^{1/2}}{q}+\delta_{f=g=E}\Big(\frac{M}{N}\Big)^{1/2}\right)(qMN)^{\varepsilon}. \tag{3.1}$$

Proof. By [29, Lemma 6.1] we have

$$\mathrm{MT}^{od,\pm}_{E,E}(M,N)\ll(qMN)^{\varepsilon}(M/N)^{1/2}.$$

Using (2.4), the second term in the definition of $\mathrm{ET}^{\pm}_{f,g}(M,N)$ is bounded by $\ll q^{\varepsilon-1}(MN)^{1/2+\varepsilon}$. The first is bounded by $\ll q^{\varepsilon-1}N^{\theta_g}(MN)^{1/2+\varepsilon}$ by using (2.3) for g and (2.4) for f. □

Our main result in this section is the following theorem, which improves on (3.1) in the ranges of critical importance to us.

Theorem 3.2. *Let f,g be either E or cuspidal Hecke eigenforms of level 1; let q be a prime and assume that W_1,W_2 satisfy (1.14). For any $\varepsilon>0$, there exists $\varepsilon'>0$ such that for $N\geqslant M\geqslant 1$ and $MN\leqslant q^{2+\varepsilon'}$, one has*

$$\mathrm{ET}^{\pm}_{f,g}(M,N)\ll q^{\varepsilon}\Big(\frac{N}{qM}\Big)^{1/4}\left(1+\Big(\frac{N}{qM}\Big)^{1/4}\right)+q^{-1/2+\theta+\varepsilon}. \tag{3.2}$$

Remark 3.3. 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 (3.2) improves on [29, Theorem 3.3].

The remaining part of this section is devoted to the proof of Theorem 3.2.

3.2. **Preliminaries.** We start with some general remarks. We denote

$$S^{\pm}_{f,g}(M,N):=\frac{1}{(MN)^{1/2}}\sum_{\substack{m\equiv\pm n\,(\mathrm{mod}\,q)\\ m\neq n}}\lambda_f(m)\lambda_g(n)W_1\Big(\frac{m}{M}\Big)W_2\Big(\frac{n}{N}\Big).$$

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

$$(3.3)\quad \frac{1}{q(MN)^{1/2}}\sum_{m,n}\lambda_f(m)\lambda_g(n)W_1\Big(\frac{m}{M}\Big)W_2\Big(\frac{n}{N}\Big) = \frac{1}{q(MN)^{1/2}}\Big(\operatorname*{res}_{s=1}L(f,s)\widehat{W}_1(s)M^s + O_{f,A}(M^{-A})\Big)\Big(\operatorname*{res}_{s=1}L(g,s)\widehat{W}_2(s)N^s + O_{g,A}(N^{-A})\Big)$$

for any $A \geqslant 0$.

If both f, g are cuspidal, or if f is cuspidal and $M \geqslant q^{\varepsilon}$, this term is very small. In particular, if both f and g are cuspidal, it is enough to obtain the stated bound for the quantity $S^{\pm}_{f,g}(M,N)$ in place of $\mathrm{ET}^{\pm}_{f,g}(M,N)$. In addition, at the cost of an additional error $O(q^{2\theta-1+\varepsilon})$, which is admissible, it suffices to estimate

$$\frac{1}{(MN)^{1/2}}\sum_{\substack{m\equiv\pm n \,(\mathrm{mod}\, q)\\ m\neq n\\ (mn,q)=1}}\lambda_f(m)\lambda_g(n)W_1\Big(\frac{m}{M}\Big)W_2\Big(\frac{n}{N}\Big).$$

Then the required estimate for this last quantity is contained in [3, (3.4), (3.11)] if $N \geqslant 20M$. These estimates hold for non-holomorphic forms as well; see [3, Section 11]. In the complementary case $N \asymp M$ we have

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

(recall that $NM \leqslant q^{2+o(1)}$) by [1, Theorem 1.3] which also holds in the non-holomorphic case. This completes the proof of (3.2) in the case f, g cuspidal.

We prepare similarly for the proof of (3.2) in the case f cuspidal, $g = E$ Eisenstein to which we devote the most work of the section. For convenience we use the Selberg eigenvalue conjecture (known in level 1), which allows to apply Lemma 2.4 with $\vartheta = 0$. 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.

For $M \leqslant q^{1/4}$ the right-hand side of (3.2) is larger than the right-hand side of the simple bound (3.1). We may therefore assume that $M \geqslant q^{1/4}$, in which case it suffices (by (3.3) again) to estimate $S^{\pm}_{f,g}(M,N)$. To begin with, we make no further assumption about the size of M, N, q and write $P := MNq$. For simplicity, we denote $\lambda(m) = \lambda_f(m)$. We open the divisor function, getting

$$(3.4)\qquad (MN)^{1/2}S^{\pm}_{f,E}(M,N) = \sum_{r\neq 0}\sum_{\substack{a,b,m\geqslant 1\\ m\mp ab=rq}}\lambda(m)W_1\left(\frac{ab}{N}\right)W_2\left(\frac{m}{M}\right).$$

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

$$(3.5)\qquad A \leqslant N^{1/2}$$

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

$$(3.6)\qquad S(M,N,q,A) = \sum_{r\neq 0}\sum_{a}\sum_{m\equiv rq\,(\mathrm{mod}\, a)}\lambda(m)W_2\left(\frac{m}{M}\right)W_3\left(\frac{a}{A}\right)W_1\left(\pm\frac{m-rq}{N}\right).$$

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

3.3. **First case.** We first assume that $NP^\varepsilon \geqslant M$. This condition implies $|rq| \leqslant N_0 := 4NP^\varepsilon$. We separate variables by Fourier inversion:

$$S(M,N,q,A) = \int_{-\infty}^{\infty} W_1^\dagger(x) \sum_{1\leqslant |r| \leqslant N_0/q} e\left(\pm \frac{rqx}{N}\right) \sum_a \sum_{m \equiv rq \,(\mathrm{mod}\, a)} \lambda(m) W_2\left(\frac{m}{M}\right) e\left(\mp\frac{mx}{N}\right) W_3\left(\frac{a}{A}\right) dx,$$

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

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

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

$$S_x(M,N,q,A) = \sum_{1\leqslant |r| \leqslant N_0/q} e\left(\pm\frac{rqx}{N}\right) \sum_a W_3\left(\frac{a}{A}\right) \sum_{m\equiv rq \,(\mathrm{mod}\, a)} \lambda(m) V\left(\frac{m}{M}\right). \tag{3.7}$$

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

$$\begin{aligned} S_x(M,N,q,A) &= \sum_{1\leqslant |r| \leqslant N_0/q} V^*\left(\frac{|r|q}{NP^\varepsilon}\right) \sum_a W_3\left(\frac{a}{A}\right) \sum_{m\equiv rq \,(\mathrm{mod}\, a)} \lambda(m) V\left(\frac{m}{M}\right) \\ &= \int_{(\varepsilon)} \widehat{V^*}(u) \sum_{1\leqslant |r| \leqslant N_0/q} \left(\frac{|r|q}{NP^\varepsilon}\right)^{-u} \sum_a W_3\left(\frac{a}{A}\right) \sum_{m\equiv rq \,(\mathrm{mod}\, a)} \lambda(m) V\left(\frac{m}{M}\right) \frac{du}{2\pi i}. \end{aligned} \tag{3.8}$$

We can truncate the u-integral at $|\mathfrak{Im}\, u| \leqslant P^{4\varepsilon}$, and hence it remains to estimate

$$\tilde{S}_u(M,N,q,A) = \sum_{1\leqslant |r| \leqslant N_0/q} |r|^{-u} \sum_a W_3\left(\frac{a}{A}\right) \sum_{m\equiv rq \,(\mathrm{mod}\, a)} \lambda(m) V\left(\frac{m}{M}\right) \tag{3.9}$$

uniformly in $\mathfrak{Re}\, u = \varepsilon$ and $|\mathfrak{Im}\, u| \leqslant P^{4\varepsilon}$. 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

$$\sum_\pm \sum_{d\mid a} \frac{M}{da} \sum_m \lambda(m) S(rq, \pm m; d) \widetilde{V}_\pm\left(\frac{mM}{d^2}\right).$$

By Lemma 2.4, we see that $\widetilde{V}_\pm$ is again a Schwartz class function satisfying

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

for any $k \geqslant 0$. This gives

$$\begin{aligned} \tilde{S}_u(M,N,q,A) &= \sum_\pm \sum_{1\leqslant |r| \leqslant N_0/q} |r|^{-u} \sum_a W_3\left(\frac{a}{A}\right) \sum_{d\mid a} \frac{M}{da} \sum_m \lambda(m) S(rq, \pm m; d) \widetilde{V}_\pm\left(\frac{mM}{d^2}\right) \\ &= \sum_\pm \frac{M}{A} \sum_{0\neq |r| \leqslant N_0/q} |r|^{-u} \sum_b \sum_d \frac{1}{d} \sum_m W_5\left(\frac{db}{A}\right) \lambda(m) S(rq, \pm m; d) \widetilde{V}_\pm\left(\frac{mM}{d^2}\right) \end{aligned}$$

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

$$1 \leqslant R \leqslant \frac{4NP^\varepsilon}{q}, \quad 1 \leqslant M^* \ll \frac{P^{4\varepsilon}A^2}{Mb^2}, \tag{3.10}$$

up to a negligible error. Then we are left with

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

where

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

The support of W_5 restricts the support of Ω to

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

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

$$\begin{aligned}
\Omega(z) &= W_6\left(\frac{zA}{\sqrt{M^*Rqb}}\right)W_5\left(\frac{4\pi\sqrt{|r|mqb}}{zA}\right)\widetilde{V}_\pm\left(\frac{z^2M}{(4\pi)^2|r|q}\right)\\
&= W_6\left(\frac{zA}{\sqrt{M^*Rqb}}\right)\int_{(0)}\int_{(\varepsilon)}\left(\frac{4\pi\sqrt{|r|mqb}}{zA}\right)^{-s}\left(\frac{z^2M}{(4\pi)^2|r|q}\right)^{-t}\widehat{W_5}(s)\widehat{\widetilde{V}}_\pm(t)\frac{dt\,ds}{(2\pi i)^2}\\
&= \int_{(0)}\int_{(\varepsilon)}\left(\frac{4\pi\sqrt{|r|m}}{\sqrt{M^*R}}\right)^{-s}\left(\frac{MM^*Rb^2}{(4\pi A)^2|r|}\right)^{-t}\left(\frac{zA}{\sqrt{M^*Rqb}}\right)^{s-2t}\widehat{W_5}(s)\widehat{\widetilde{V}}_\pm(t)W_6\left(\frac{zA}{\sqrt{M^*Rqb}}\right)\frac{dt\,ds}{(2\pi i)^2}.
\end{aligned}$$

(Assuming $\vartheta = 0$ in lemma 2.4 allows to shift the t-contour to $\mathfrak{Re}\, t = \varepsilon$.) The integrals can be truncated at $|\mathfrak{Im}\, s|, |\mathfrak{Im}\, t| \leqslant P^{4\varepsilon}$ 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

$$S_{u,s,t}(M,N,q,A,R,M^*) = \frac{M}{A}\sum_b\Big|\sum_{R\leqslant|r|\leqslant 2R}|r|^{t-\frac{s}{2}-u}\sum_{M^*\leqslant m\leqslant 2M^*}\lambda(m)m^{-\frac{s}{2}}\Sigma(rq,m)\Big|$$

where

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

and $\mathfrak{Re}\, t = \mathfrak{Re}\, u = \varepsilon$, $\mathfrak{Re}\, s = 0$, $|\mathfrak{Im}\, t|, |\mathfrak{Im}\, u|, |\mathfrak{Im}\, s| \leqslant P^{4\varepsilon}$. This is in a form to apply the Kuznetsov formula (Lemma 2.5). We treat in detail the case $r > 0$, $\pm m > 0$, the other case is essentially identical. We get

$$\begin{aligned}
\Sigma(rq,m) = &\sum_{\substack{k\geqslant 2\\ k \text{ even}}}\sum_{f\in\mathcal{B}_k}\dot{\Theta}(k)\Gamma(k)\sqrt{rqm}\,\varrho_f(rq)\varrho_f(m)\\
&+\sum_{f\in\mathcal{B}}\tilde{\Theta}(t_f)\frac{\sqrt{rqm}}{\cosh(\pi t_f)}\varrho_f(rq)\varrho_f(m) + \frac{1}{4\pi}\int_{-\infty}^{\infty}\tilde{\Theta}(t)\frac{\sqrt{rqm}}{\cosh(\pi t)}\varrho(rq,t)\varrho(m,t)dt.
\end{aligned}$$

By [2, Lemma 2.1] we have

$$\dot{\Theta}(k) \ll_{B,\varepsilon} \frac{P^{12\varepsilon}}{\mathcal{T}}\left(1+\frac{k}{P^{13\varepsilon}\mathcal{T}}\right)^{-B}, \qquad \tilde{\Theta}(t) \ll_{B,\varepsilon} \frac{P^{12\varepsilon}}{\mathcal{T}}\left(1+\frac{|t|}{P^{13\varepsilon}\mathcal{T}}\right)^{-B},$$

where

$$\mathcal{T} = 1 + \frac{\sqrt{M^*Rq}b}{A}$$

(again this uses, for simplicity, the Selberg eigenvalue conjecture, known in the present case of level 1.) From now on, we use ε-convention. By the Cauchy–Schwarz inequality we find that the contribution of the holomorphic spectrum is at most

$$\sum_{b\leqslant P} \frac{P^\varepsilon M}{\sqrt{M^*Rq}b}\Bigg(\sum_{\substack{2\leqslant k\ll P^\varepsilon\mathcal{T}\\ k \text{ even}}} \Gamma(k)\sum_{f\in\mathcal{B}_k}\Big|\sum_{R\leqslant r\leqslant 2R} r^{t-\frac{s}{2}-u}\sqrt{rq}\varrho_f(rq)\Big|^2\Bigg)^{1/2}$$
$$\Bigg(\sum_{\substack{2\leqslant k\leqslant P^\varepsilon\mathcal{T}\\ k \text{ even}}} \Gamma(k)\sum_{f\in\mathcal{B}_k}\Big|\sum_{M^*\leqslant m\leqslant 2M^*} \lambda(m)m^{-\frac{s}{2}}\sqrt{m}\varrho_f(m)\Big|^2\Bigg)^{1/2}.$$

Using the Ramanujan conjecture for $\sqrt{q}\varrho_f(q)$ and the spectral large sieve (Lemma 2.6), this is (recalling (3.5), (3.10))

$$\begin{aligned}
&\sum_{b\leqslant P}\frac{P^\varepsilon M}{\sqrt{M^*Rq}b}\left(\Big(\frac{RM^*qb^2}{A^2}+R\Big)R\right)^{1/2}\left(\Big(\frac{RM^*qb^2}{A^2}+M^*\Big)M^*\right)^{1/2}\\
&\ll P^\varepsilon\sum_{b\leqslant P}\frac{M}{\sqrt{q}b}\left(\frac{N}{M}+\frac{N}{q}\right)^{1/2}\left(\frac{N}{M}+\frac{A^2}{Mb^2}\right)^{1/2} \ll P^\varepsilon\left(\frac{N}{q^{1/2}}+\frac{N\sqrt{M}}{q}\right).
\end{aligned} \tag{3.11}$$

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 $R \geqslant q$) to conclude that the total contribution of the Maaß spectrum is

$$\begin{aligned}
&\sum_{b\leqslant P}\frac{P^\varepsilon M}{\sqrt{M^*Rq}b}\big[((\mathcal{T}^2+R^2)R^2)^{1/4}(\mathcal{T}^2+q)^{1/4}+(\mathcal{T}^2R^{4\theta})^{1/2}\big]((\mathcal{T}^2+M^*)M^*)^{1/2}\\
&\ll P^\varepsilon\sum_{b\leqslant P}\frac{M}{\sqrt{q}b}\left(\frac{N}{M}+\frac{N^2}{q^2}\right)^{1/4}\left(\frac{N}{M}+q\right)^{1/4}\left(\frac{N}{M}+\frac{N}{Mb^2}\right)^{1/2}\\
&\ll P^\varepsilon\left(\frac{N}{q^{1/2}}+\frac{M^{1/4}N^{3/4}}{q^{1/4}}\right)\left(1+\frac{(MN)^{1/4}}{q^{1/2}}\right).
\end{aligned} \tag{3.12}$$

Note that (3.12) is larger than (3.11) when $M \ll NP^\varepsilon$. This completes the analysis of the contribution of $\Sigma(rq, m)$.

3.4. **Second case.** We now assume that $NP^\varepsilon \leqslant M$. We return to (3.6) and begin with some preliminary transformations. We write

$$V(z) = V_{rq}(z) := W_2\left(\frac{Nz+rq}{M}\right) = \int_{-\infty}^{\infty} W_2^\dagger(x)e\left(\frac{Nz+rq}{M}x\right)dx.$$

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

$$\begin{aligned} W_2\left(\frac{m}{M}\right) W_1\left(\pm\frac{m-rq}{N}\right) &= V\left(\frac{m-rq}{N}\right) W_1\left(\pm\frac{m-rq}{N}\right) \\ &= \int_{-\infty}^{\infty} W_2^\dagger(x) e\left(\frac{rqx}{M}\right) W_4\left(\pm\frac{m-rq}{N}\right) dx. \end{aligned}$$

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

$$S_x(M,N,q,A) = \sum_{r \asymp M/q} e\left(\frac{rqx}{M}\right) \sum_a \sum_{m \equiv rq \,(\mathrm{mod}\ a)} \lambda(m) W_3\left(\frac{a}{A}\right) W_4\left(\pm\frac{m-rq}{N}\right),$$

where $r \asymp M/q$ is short for $r \in [c_1 M/q, c_2 M/q]$ for suitable constants c_1, c_2. As in (3.7) – (3.9) we may separate the variables r and q, and need to bound

$$\tilde{S}_u(M,N,q,A) = \sum_{r \asymp M/q} r^{-u} \sum_a \sum_{m \equiv rq \,(\mathrm{mod}\ a)} \lambda(m) W_3\left(\frac{a}{A}\right) W_4\left(\pm\frac{m-rq}{N}\right)$$

with $\mathfrak{Re}\, u = \varepsilon$, $|\mathfrak{Im}\, u| \leqslant P^\varepsilon$. 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

(3.13)

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

where

$$W_4^\pm(z,w) = \int_0^\infty W_4(y) \mathcal{J}_\pm(4\pi\sqrt{z+wy}) dy, \quad 4|w| \leqslant z.$$

Note that by our current size assumption $NP^\varepsilon \leqslant M$, the first argument in $W_4^\epsilon(z,w)$ is substantially larger than the second. We follow the argument of [3, Lemma 17 & Remark after Corollary 18].

As

$$\frac{mrq}{d^2} - 2\frac{mN}{d^2} \gg \frac{M - O(N)}{A^2} \gg \frac{M}{N} \gg P^\varepsilon,$$

the case $\epsilon = -1$ contributes negligibly due to the rapid decay of the Bessel-K-function (this is another reason why we separate the two cases in the somewhat artificial way $NP^\varepsilon \geqslant M$ and $NP^\varepsilon \leqslant M$), cf. (2.8). Hence it suffices to consider only the case $\epsilon = 1$. For later purposes (see (3.15) below) it is convenient to insert into (3.13) a smooth, redundant weight function $W_0(mrq/d^2, \pm mN/d^2)$ such that $W_0(z,w) = 0$ for $z \leqslant 1$ or $3|w| \geqslant z$, and $W_0(z,w) = 1$ for $z \geqslant 2$ and $4|w| \leqslant z$. We write $W_5(z,w) = W_0(z,w) W_4^+(z,w)$, so that

$$\tilde{S}_u(M,N,q,A) = \sum_{r \asymp M/q} r^{-u} \sum_a W_3\left(\frac{a}{A}\right) \sum_{d \mid a} \frac{N}{da} \sum_m \lambda(m) S(rq, m; d) W_5\left(\frac{mrq}{d^2}, \pm\frac{mN}{d^2}\right)$$

up to a negligible error coming from $\epsilon = -1$. An integral transform similar to $W_5(z,w)$ was analyzed in [3, Lemma 17]. Our general assumption in the forthcoming analysis is

$$z \asymp z + wy \gg P^\varepsilon.$$

Repeated application of the formula (2.11) yields the preliminary bound

$$W_5(z,w) \ll_k \left(\frac{\sqrt{z}}{w}\right)^k$$

for any $k \geqslant 0$. In particular, up to a negligible error of $O(P^{-k})$, we can assume that

$$\sqrt{z} \geqslant wP^{-\varepsilon}. \tag{3.14}$$

In this range we use the asymptotic formula (2.12), so that

$$W_5(z,w) = W_+(z,w)e(2\sqrt{z}) + W_-(z,w)e(-2\sqrt{z}) + \mathrm{O}(P^{-k}),$$

where

$$z^i|w|^j \frac{\partial^i}{\partial z^i}\frac{\partial^j}{\partial w^j} W_\pm(z,w) \ll P^{\varepsilon(i+j)} z^{-1/4}.$$

It is now easy to see (cf. [3, Corollary 18]) that its double Mellin transform

$$\widehat{W}_{\pm,\pm}(s,t) = \int_0^\infty \int_0^\infty W_\pm(z,\pm w) z^{s-1} w^{t-1} dz\, dw \tag{3.15}$$

is rapidly decaying on vertical lines (i.e. is $\ll_{k,\ell,\varepsilon} P^\varepsilon |s|^{-k}|t|^{-\ell}$ for $|s|,\ |t| \geqslant 1$) and absolutely convergent in $\mathfrak{Re}\, t > 0$, $\mathfrak{Re}\, s + \mathfrak{Re}\, t/2 < 1/4$.

We can restrict m to a dyadic range $M^* \leqslant m \leqslant 2M^*$, and (3.14) implies

$$M^* \ll \frac{P^{2\varepsilon} M A^2}{(bN)^2}.$$

This leaves us with bounding

$$\begin{aligned}
&\tilde{S}_u(M,N,q,A,M^*) \\
&= \sum_{r \asymp M/q} r^{-u} \sum_{b \leqslant P} \frac{1}{b} \sum_d W_3\left(\frac{bd}{A}\right)\frac{N}{d^2} \sum_{M^* \leqslant m \leqslant 2M^*} \lambda(m) S(rq,m;d) e\left(\pm\frac{2\sqrt{mrq}}{d}\right) W_\pm\left(\frac{mrq}{d^2}, \frac{mN}{d^2}\right) \\
&= \sum_{b \leqslant P} \frac{N}{b} \sum_{r \asymp M/q} r^{-u} \sum_{M^* \leqslant m \leqslant 2M^*} \lambda(m) \sum_d \frac{1}{d} S(rq,m;d)\Omega\left(\frac{4\pi\sqrt{mrq}}{d}\right),
\end{aligned}$$

where

$$\Omega(z) = W_3\left(\frac{4\pi b\sqrt{mrq}}{Az}\right)\frac{z}{4\pi\sqrt{mrq}} W_\pm\left(\frac{z^2}{(4\pi)^2}, \frac{z^2}{(4\pi)^2}\frac{N}{rq}\right)\exp(\pm iz)$$

has support contained in

$$z \asymp Z := \frac{b\sqrt{M^* M}}{A} \gg 1.$$

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:

$$\begin{aligned}
\Omega(z) &= W_6\left(\frac{z}{Z}\right)\exp(\pm iz) \\
&\quad \times \int_{(0)}\int_{(\varepsilon)}\int_{(1/4-\varepsilon)} \widehat{W_3}(v)\widehat{W}_\pm(s,t)\left(\frac{4\pi b\sqrt{mrq}}{Az}\right)^{-v} \frac{z}{4\pi\sqrt{mrq}}\left(\frac{z}{4\pi}\right)^{-2s-2t}\left(\frac{N}{rq}\right)^{-t}\frac{ds}{2\pi i}\frac{dt}{2\pi i}\frac{dv}{2\pi i} \\
&= \int_{(0)}\int_{(\varepsilon)}\int_{(1/4-\varepsilon)} \frac{\widehat{W_3}(v)\widehat{W}_\pm(s,t)}{(4\pi)^{1+v-2s-2t}} W_6\left(\frac{z}{Z}\right)\exp(\pm iz)\left(\frac{b}{A}\right)^{-v} \\
&\quad \times (M^*)^{-\frac{v+1}{2}} M^{t-\frac{v+1}{2}} Z^{1-2s-2t} N^{-t}\left(\frac{m}{M^*}\right)^{-\frac{v+1}{2}}\left(\frac{rq}{M}\right)^{t-\frac{v+1}{2}}\left(\frac{z}{Z}\right)^{1-2s-2t}\frac{ds}{2\pi i}\frac{dt}{2\pi i}\frac{dv}{2\pi i}.
\end{aligned}$$

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

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

where $\alpha = \frac{v+1}{2} - t + u$ 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 16], the spectral sum can be truncated (with a negligible error) at spectral parameter $P^{3\varepsilon}Z^{1/2}$, and we obtain

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

(up to a negligible error), where

$$\mathcal{H} = \sum_{\substack{2 \leqslant k \ll P^{3\varepsilon}Z^{1/2} \\ k \text{ even}}} \sum_{f \in \mathcal{B}_k} 4i^k \Gamma(k) \int_0^\infty J_{k-1}(z) \Theta(z) \frac{dz}{z} \sqrt{mrq}\, \varrho_f(m) \varrho_f(rq),$$

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

$$\mathcal{E} = \int_{|t| \ll P^{3\varepsilon}Z^{1/2}} \frac{i}{2} \int_0^\infty \frac{J_{2it}(z) - J_{-2it}(z)}{\sinh(\pi t)} \Theta(z) \frac{dz}{z} \frac{\sqrt{mrq}\, \varrho(m,t) \varrho(rq,t)}{\cosh(\pi t)} dt$$

denote the respective contributions of the holomorphic cusp forms, non-holomorphic (Maaß) cusp forms and of the Eisenstein series. It follows from (2.9) and (2.10) that

$$J_{k-1}(z), \frac{J_{2it}(z) - J_{-2it}(z)}{\sinh(\pi t)} \ll P^{\varepsilon} Z^{-1/2}, \quad z \asymp Z, \quad t, k \ll P^{3\varepsilon} Z^{1/2}.$$

(Indeed, if $k \asymp Z$, then $Z \ll P^{6\varepsilon}$ and $k^{-1/3} \asymp Z^{1/6} Z^{-1/2} \ll P^{\varepsilon} Z^{-1/2}$.) We estimate the z-integral trivially. From now on we use ε-convention. The Maaß contribution is at most

$$\sum_{b \leqslant P} \frac{P^{\varepsilon} N}{b\sqrt{M^*M}} \sum_{\substack{f \in \mathcal{B} \\ t_f \ll P^{3\varepsilon}Z^{1/2}}} \Big| \sum_{r \asymp M/q} r^{-\alpha} \sum_{M^* \leqslant m \leqslant 2M^*} \lambda(m) \left(\frac{m}{M^*}\right)^{-\frac{v+1}{2}} \frac{\sqrt{mrq}\, \varrho_f(m) \varrho_f(rq)}{\cosh(\pi t_f)} \Big|,$$

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

$$\sum_b \frac{P^{\varepsilon} N}{\sqrt{bM^*M}} \Bigg(\sum_{\substack{f \in \mathcal{B} \\ t_f \ll P^{\varepsilon}Z^{1/2}}} \Big| \sum_{r \asymp M/q} r^{-\alpha} \frac{\sqrt{rq}\, \varrho_f(rq)}{\cosh(\pi t_f)} \Big|^2 \Bigg)^{1/2} \times \Bigg(\sum_{\substack{f \in \mathcal{B} \\ t_f \ll P^{\varepsilon}Z^{1/2}}} \sum_{M^* \leqslant m \leqslant 2M^*} \lambda(m) \left(\frac{m}{M^*}\right)^{-\frac{v+1}{2}} \frac{\sqrt{m}\, \varrho_f(m)}{\cosh(\pi t_f)} \Big|^2 \Bigg)^{1/2}.$$

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

$$\ll \sum_{b\leqslant P} \frac{P^\varepsilon N}{b\sqrt{M^*M}} \left[(Z+q)^{1/4} \left(\left(Z + \frac{M^2}{q^2}\right)\frac{M^2}{q^2}\right)^{1/4} + \left(\left(Z + \frac{M}{q}\right)\frac{M}{q}\right)^{1/2} \right] ((Z+M^*)M^*)^{1/2}$$
$$\ll \sum_{b\leqslant P} \frac{P^\varepsilon N}{b\sqrt{M}} \left[\left(\frac{M}{N}+q\right)^{1/4} \left(\left(\frac{M}{N} + \frac{M^2}{q^2}\right)\frac{M^2}{q^2}\right)^{1/4} + \left(\left(\frac{M}{N} + \frac{M}{q}\right)\frac{M}{q}\right)^{1/2} \right] \left(\frac{M}{N} + \frac{MA^2}{(bN)^2}\right)^{1/2}.$$

By (3.5) this

$$\begin{aligned} &\ll P^\varepsilon N^{1/2}\left[\frac{M}{N^{1/2}q^{1/2}} + \frac{M^{5/4}}{qN^{1/4}} + \frac{M^{3/4}}{N^{1/4}q^{1/4}} + \frac{M}{q^{3/4}}\right] \\ &\ll P^\varepsilon \left(\frac{M}{q^{1/2}} + \frac{M^{3/4}N^{1/4}}{q^{1/4}}\right)\left(1 + \frac{(MN)^{1/4}}{q^{1/2}}\right). \end{aligned} \tag{3.16}$$

Combining (3.12) and (3.16), and recalling the extra factor $(MN)^{1/2}$ in (3.4), we complete the proof of (3.2) in the case f cuspidal, $g = E$ Eisenstein.

3.5. **The case $f = g = E$.** Here we merely indicate the points where the proof of [29, Thm. 3.3] needs some modifications. We will freely borrow the notations of that paper, where the error term $\mathrm{ET}_{E,E}(M,N)$ is denoted by $E_{M,N}$.

In [29, §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 E_K where K is a parameter around which the Laplace eigenvalues of the Maaß forms are localized.

In [29, (9.12)], we apply Hölder's inequality with exponents $(1/4, 1/4, 1/2)$, and obtain

$$\Big| \sum_{K\leqslant\kappa_j<2K} \frac{|\varrho_j(1)|^2}{\cosh(\pi\kappa_j)}\lambda_j(q)L_j(1/2+s_1)L_j(1/2+s_2)^2\Big| \leqslant \Big|\sum_{K\leqslant\kappa_j<2K}\frac{|\varrho_j(1)|^2}{\cosh(\pi\kappa_j)}|\lambda_j(q)|^4\Big|^{1/4}$$
$$\cdot \Big|\sum_{K\leqslant\kappa_j<2K}\frac{|\varrho_j(1)|^2}{\cosh(\pi\kappa_j)}|L_j(1/2+s_1)|^4\Big|^{1/4} \cdot \Big|\sum_{K\leqslant\kappa_j<2K}\frac{|\varrho_j(1)|^2}{\cosh(\pi\kappa_j)}|L_j(1/2+s_1)|^4\Big|^{1/2}.$$

Exactly as in [29, (9.13)], we bound the last two factors by $K^{3/2+\varepsilon}$. For the first factor we write

$$|\lambda_j(q)|^4 \leqslant 2(|\lambda_j(1)| + |\lambda_j(q^2)|)^2$$

and use [26, Lemma 2.4] to estimate this factor by $(qK)^\varepsilon(K^2+q)^{1/4}$. In this way, [29, (9.15)] becomes

$$E_{m\pm} \ll q^{-1/2+\varepsilon}\left(\frac{N}{M}\right)^{1/2} + q^{-1/4+\varepsilon}\left(\frac{N}{M}\right)^{1/4}. \tag{3.17}$$

Then [29, Proposition 9.3] and the last bound of [29, 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 (3.17), at least for $\theta < 1/4$ and under the general assumption $N \geqslant M$. Hence we get an improved version of [29, Theorem 3.4]:

$$\mathrm{ET}_{E,E}(M,N) \ll q^{-1/2+\varepsilon}\left(\frac{N}{M}\right)^{1/2} + q^{-1/4+\varepsilon}\left(\frac{N}{M}\right)^{1/4} \tag{3.18}$$

under the assumption $MN \ll q^{2+\varepsilon}$ and $N \geqslant M$. □

Remark 3.4. Inserting (3.18) into the subsequent analysis of the piecewise linear function at the end of Section 3 in [29], we obtain the exponent $-1/82$ in the error term of [29, Theorem 1] (with no θ-dependence), the maximum being taken at $a = 21/41$ and $b = 60/41$.

4. Young's method

In this section we prove the following small variation of [29, Lemma 4.1, 4.2].

Proposition 4.1. *Let $\lambda(m)$ denote either the (normalized) Fourier coefficients of a cuspidal Hecke eigenform f of level 1, or the divisor function. Let $N, N_1, N_2, M \geqslant 1$ be parameters with $N_1N_2 = N$, $N_1 \leqslant N_2$, and let $q \in \mathbb{N}$. Let W_1, W_2, W_3 be smooth compactly supported weight functions satisfying* (1.14)*. Then*

$$\frac{1}{\sqrt{MN}} \sum_{n_1n_2 \equiv \pm m \,(\mathrm{mod}\ q)} \lambda(m) W_1(n_1/N_1) W_2(n_2/N_2) W_3(m/M)$$

is bounded by the following two quantities:

$$(MNq)^{\varepsilon} \cdot \begin{cases} \frac{\sqrt{MN}}{q^{2-\theta}} + \min\left(\frac{(Mq)^{1/2}}{N^{1/2}} + \frac{N_1M^{1/2}}{qN^{1/2}}, \frac{q^{1/4}}{N_1^{1/2}} + \frac{(MN_1)^{1/2}}{N^{1/2}} + \frac{N_1M^{1/2}}{qN^{1/2}}, \frac{M^{1/2}N_1}{N^{1/2}}\right), \\ \frac{\sqrt{MN}}{q^{2-\theta}} + \min\left(\frac{N_1^2}{(MN)^{1/2}}, \frac{N^{1/6}N_1q^{1/2}}{N_2M^{2/3}}\right) + \frac{M^{1/2}}{N^{1/2}} + \frac{M^{1/2}N_1}{qN^{1/2}} + \frac{M^{3/2}}{N_2N^{1/2}}. \end{cases}$$

Remark 4.2. 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 1.2, which dualizes the variables n_1, n_2 simultaneously in the form of Voronoi summation.

Proof. We follow closely the argument in [29, Lemma 4.1, 4.2] and keep track of the following two differences. We drop the assumption $MN \leqslant q^{2+\varepsilon}$ and we allow that λ can be the divisor function or the sequence of Hecke eigenvalues (and we make sure to use only bounds of the type (2.4) and (2.3)).

The first bound is the analogue of [29, Lemma 4.1]. We can exclude the terms $n_1n_2 \equiv 0 \pmod q$ at the cost of an error $(MNq)^{\varepsilon}(MN)^{1/2}q^{-2+\theta}$. 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 [29, (4.6)] by

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

where

$$S(K,L,q) = \sum_{\substack{l \leqslant L \\ (l,q)=1}} |(\lambda * 1)(l)| \cdot \Big| \sum_{(k,q)=1} e\left(\frac{l\bar{k}}{q}\right) W\left(\frac{k}{K}\right) \Big|$$

is analogous to [29, (4.7)]. The proof of [29, Proposition 4.3] provides bounds for $S(K, L, q)$ in the situation where λ 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 [29, Proposition 4.3] uses only ∞-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

$$S(K,L,q) \ll (KLq)^{\varepsilon} \min\Big(Lq^{1/2} + KL/q, (Lq^{3/2} + L^2K + L^2K^2/q)^{1/2}, LK\Big),$$

where the first bound is the analogue of [29, (4.11)], the second bound is the analogue of the last display in [29, 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 in Proposition 4.1 is the analogue of [29, Lemma 4.2], and again we only indicate the changes in Young's proof. The error term in the second display of [29, 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 [29, (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 [29, p. 22] and arrive at a quantity analogous to [29, (4.15)]. Finally, the pointwise bound on the quantity $V(h, k)$ defined under [29, (4.17)] and proved above [29, (4.18)] allows us to reach the analogue of [29, (4.16)–(4.17)], which yields the second bound of our proposition (and we notice that the assumption $N \gg q^{1+\varepsilon}$ in [29, 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 4.3. *Let $\lambda(m)$ denote either the (normalized) Fourier coefficients of a cuspidal Hecke eigenform f of level 1, or the divisor function. Let*

$$1 \leqslant N' \leqslant N^*, \quad N_1N_2 = N', \quad N_2 \geqslant N_1 \geqslant 1, \quad 1 \leqslant M' \leqslant M^*$$

be parameters and let $q \in \mathbb{N}$. Let W_1, W_2, W_3 be smooth compactly supported weight functions satisfying (1.14). Then

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

is bounded by the following two quantities:

$$(M^*N^*q)^{\varepsilon}\cdot\begin{cases}\frac{\sqrt{M'N'}}{q^{2-\theta}} + \min\left(\frac{M'q^{1/2}}{\sqrt{M^*N^*}} + \frac{N_1M'}{q\sqrt{M^*N^*}}, \frac{\sqrt{M'N_2}q^{1/4}}{\sqrt{M^*N^*}} + \frac{M'N_1^{1/2}}{\sqrt{M^*N^*}} + \frac{N_1M'}{q\sqrt{M^*N^*}}, \frac{M'N_1}{\sqrt{M^*N^*}}\right),\\ \frac{\sqrt{M'N'}}{q^{2-\theta}} + \min\left(\frac{N_1^2}{\sqrt{M^*N^*}}, \frac{(M')^{-1/6}(N')^{2/3}N_1q^{1/2}}{N_2\sqrt{M^*N^*}}\right) + \frac{M'}{\sqrt{M^*N^*}} + \frac{M'N_1}{q\sqrt{M^*N^*}} + \frac{(M')^2}{N_2\sqrt{M^*N^*}}.\end{cases}$$

5. Bilinear forms in Kloosterman sums

5.1. **Statements of results.** We begin by stating the precise results we obtain concerning the bilinear forms of type (1.11), i.e.

$$B(\mathrm{Kl}_2, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_m \sum_n \alpha_m\beta_n \,\mathrm{Kl}_2(amn; q),$$

where the normalized Kloosterman sum $\mathrm{Kl}_2(a; q)$ is defined in (1.15).

We recall from Section 1.2 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 5.5, being of similar (or better) quality for special coefficients (α_m) and (β_n). Proposition 5.5 itself will be proved in Subsection 5.5, depending on a conjecture on certain complete sums over finite fields.

Theorem 5.1. *Let q be a prime number, let a be an integer coprime with q, $M, N \geqslant 1$, $\mathcal{N}$ an interval of length N, and $(\alpha_m)_m$, $(\beta_n)_n$ two sequences supported respectively on $[1, M]$ and $\mathcal{N}$.*

(1) *If $M, N \leqslant q$, we have*

$$\sum\sum_{m\leqslant M,\, n\in\mathcal{N}} \alpha_m\beta_n \operatorname{Kl}_2(amn; q) \ll (qMN)^{\varepsilon}(MN)^{1/2}\|\alpha\|_2\|\beta\|_2\big(M^{-1/2} + q^{1/4}N^{-1/2}\big) \tag{5.1}$$

for any $\varepsilon > 0$.

(2) *If the conditions*

$$M, N \leqslant q, \quad MN \leqslant q^{3/2}, \quad M \leqslant N^2, \tag{5.2}$$

are satisfied, then we have

$$\sum\sum_{m\leqslant M,\, n\in\mathcal{N}} \alpha_m \operatorname{Kl}_2(amn; q) \ll (qMN)^{\varepsilon}(\|\alpha\|_1\|\alpha\|_2)^{1/2}M^{1/4}N\big(q^{\frac{1}{4}}M^{-\frac{1}{6}}N^{-\frac{5}{12}}\big). \tag{5.3}$$

(3) *Let W_i, for $1 \leqslant i \leqslant 2$, be smooth, compactly supported functions satisfying*

$$W_i^{(j)}(x) \ll_j Q^j, \ i = 1, 2, \tag{5.4}$$

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

$$\sum_m\sum_n W_1\Big(\frac{m}{M}\Big)W_2\Big(\frac{n}{N}\Big)\operatorname{Kl}_2(amn; q) \ll_{\varepsilon} q^{\varepsilon}Q^A MN\Big(\frac{1}{q^{1/8}} + \frac{q^{3/8}}{(MN)^{1/2}}\Big). \tag{5.5}$$

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

$$\|\alpha\|_1 = \sum_m |\alpha_m|, \quad \|\alpha\|_2 = \Big(\sum_m |\alpha_m|^2\Big)^{1/2}.$$

The first part (5.1) and the third (5.5) have been proven by Fouvry, Kowalski and Michel in Theorems 1.17 and 1.16 (respectively) of [12] (building on [11]), in considerably greater generality. Thus it remains to prove the second part.

5.2. **General setup.** Some of our arguments are valid for bilinear forms involving a more general kernel $K(mn)$ modulo q than the Kloosterman sums $\operatorname{Kl}_2(amn; q)$. It is therefore useful to consider first the general problem of bounding a general "type I" sum

$$B(K, \boldsymbol{\alpha}_M, 1_{\mathcal{N}}) := \sum_{m\leqslant M} \alpha_m \sum_{n\in\mathcal{N}} K(mn), \tag{5.6}$$

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

The proof is a slight generalization of the method of [14]: given $A, B \geqslant 1$ such that

$$AB \leqslant N, \ 2AM < q, \tag{5.7}$$

we have

$$\begin{aligned} B(K, \boldsymbol{\alpha}_M, 1_{\mathcal{N}}) &= \frac{1}{AB}\sum\sum_{\substack{A<a\leqslant 2A\\ B<b\leqslant 2B}}\sum_{m\leqslant M}\alpha_m\sum_{n+ab\in\mathcal{N}} K(m(n+ab)) \\ &= \frac{1}{AB}\sum\sum_{\substack{A<a\leqslant 2A\\ B<b\leqslant 2B}}\sum_{m\leqslant M}\alpha_m\sum_{n+ab\in\mathcal{N}} K(am(\bar{a}n+b)), \end{aligned}$$

where, as usual, $a\overline{a}\equiv 1 \,(\mathrm{mod}\, q)$. By the method of [14, p. 116], we get

$$B(K,\boldsymbol{\alpha}_M,1_{\mathcal{N}}) \ll \frac{\log q}{AB} \sum\sum_{r\,(\mathrm{mod}\,q),\, s\leqslant 2AM} \nu(r,s)\Big|\sum_{B<b\leqslant 2B}\eta_b K(s(r+b))\Big|$$

for

$$\nu(r,s) = \sum\sum\sum_{\substack{A<a\leqslant 2A,\ m\leqslant M,\ n\in\mathcal{N}',\\ am=s,\ \overline{a}n\equiv r\,(\mathrm{mod}\,q)}} |\alpha_m|,$$

where $\mathcal{N}'\supset\mathcal{N}$ is some interval of length $2N$ and $(\eta_b)_{B<b\leqslant 2B}$ are some complex numbers such that $|\eta_b|\leqslant 1$. We have the bounds

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

and

$$\sum_{r,s}\nu(r,s)^2 \;=\; \sum_{\substack{a,m,n,a',m',n'\\ am=a'm',\\ a'n=an'\,(\mathrm{mod}\,q)}}\!\!\cdots\sum |\alpha_m||\alpha_{m'}| \ll \sum_{a,m}|\alpha_m|^2 \sum_{\substack{n,a',m',n'\\ am=a'm',\\ a'n=an'\,(\mathrm{mod}\,q)}}\!\!\cdots\sum 1 \ll q^{\varepsilon}AN\sum_m|\alpha_m|^2.$$

Here, we have used the inequality $|\alpha_m||\alpha_{m'}|\leqslant|\alpha_m|^2+|\alpha_{m'}|^2$ 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 $n\in\mathcal{N}'$, the congruence $a'n=an'\,(\mathrm{mod}\,q)$ has at most two solutions in the interval $\mathcal{N}'$ since it has length $\leqslant 2q$ (cf. [14, p. 116]).

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

$$AB\cdot B(K,\boldsymbol{\alpha}_M,1_{\mathcal{N}}) \ll q^{\varepsilon}(AN)^{3/4}(\|\alpha\|_1\|\alpha\|_2)^{1/2}\Big(\sum\sum_{r\,(\mathrm{mod}\,q),1\leqslant s\leqslant AM}\Big|\sum_{B<b\leqslant 2B}\eta_b K(s(r+b))\Big|^4\Big)^{1/4}. \tag{5.8}$$

Expanding the fourth power, the inner term of the second factor can be written as

$$\sum_{\boldsymbol{b}\in\mathcal{B}}\eta(\boldsymbol{b})\Sigma(K,\boldsymbol{b}),$$

where $\mathcal{B}$ denotes the set of quadruples $\boldsymbol{b}=(b_1,b_2,b_1',b_2')$ of integers satisfying $B<b_i,b_i'\leqslant 2B$ $(i=1,2)$, the coefficients $\eta(\boldsymbol{b})$ satisfy $|\eta(\boldsymbol{b})|=1$ for all $\boldsymbol{b}\in\mathcal{B}$, and we denote

$$\Sigma(K,\boldsymbol{b}) = \sum\sum_{\substack{r\,(\mathrm{mod}\,q)\\ 1\leqslant s\leqslant AM}} K(s(r+b_1))K(s(r+b_2))\overline{K(s(r+b_1'))K(s(r+b_2'))}. \tag{5.9}$$

Let $\mathcal{B}^{\Delta}$ be the subset of $\boldsymbol{b}$ 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 $\boldsymbol{b}$, we use the trivial bound for $\Sigma(K,\boldsymbol{b})$, getting

$$\sum_{\boldsymbol{b}\in\mathcal{B}^{\Delta}}|\Sigma(K,\boldsymbol{b})| \ll AB^2Mq, \tag{5.10}$$

where the implied constant depends only on H.

To bound the contribution of the $\boldsymbol{b}\notin\mathcal{B}^{\Delta}$, we complete the s-sum using additive characters and obtain

$$\Sigma(K,\boldsymbol{b}) \ll (\log q)\max_{h\,(\mathrm{mod}\,q)}|\Sigma(K,\boldsymbol{b},h;q)|,$$

where

$$\Sigma(K,\boldsymbol{b},h) := \sum\sum_{r,s\,(\mathrm{mod}\,q)} K(s(r+b_1))K(s(r+b_2))\overline{K(s(r+b_1'))K(s(r+b_2'))}e_q(hs). \tag{5.11}$$

The procedure we have described gives a general scheme for estimating special bilinear forms (5.6) with a general uniformly bounded kernel K: the sum $B(K, \boldsymbol{\alpha}_M, 1_{\mathcal{N}})$ is estimated as in (5.8), where the contribution of the diagonal quadruples $\boldsymbol{b} \in \mathcal{B}^{\Delta}$ is bounded in (5.10), and the contributions of off-diagonal quadruples $\boldsymbol{b} \notin \mathcal{B}^{\Delta}$ are estimated in terms of the *complete* sums $\Sigma(K, \boldsymbol{b}, h)$ given by (5.11).

If we now insert the trivial bound $\Sigma(K, \boldsymbol{b}, h) \ll q^2$, we obtain a bound that is never better than the trivial estimate $B(K, \boldsymbol{\alpha}_M, 1_{\mathcal{N}}) \ll MN$. We must improve on this by exhibiting cancellation in the complete sum $\Sigma(K, \boldsymbol{b}, h)$ by exploiting the structure of K.

In Section 5.3, we will further show how, for kernels K that are themselves given by a complete exponential sum of a specific shape, the estimation of $\Sigma(K, \boldsymbol{b}, h)$ 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 5.2. *Let q be a prime and define $K(a) = \mathrm{Kl}_2(a; q)$. With notation as above, for all $\boldsymbol{b} \in \mathcal{B} - \mathcal{B}^{\Delta}$ and all $h \in \mathbf{F}_q$, we have*

$$|\Sigma(K, \boldsymbol{b}, h)| \ll q, \tag{5.12}$$

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 (5.8), (5.10), and (5.12) that

$$B(K, \boldsymbol{\alpha}_M, 1_{\mathcal{N}}) \ll q^{\varepsilon}(AB)^{-1}(AN)^{3/4}(\|\alpha\|_1\|\alpha\|_2)^{1/2}(AB^2Mq + B^4q)^{1/4}.$$

We then finish the proof of the second part of Theorem 5.1 by choosing

$$A = \frac{1}{2}M^{-\frac{1}{3}}N^{\frac{2}{3}},\ B = (MN)^{\frac{1}{3}};$$

note that the conditions (5.7) as well as $A, B \geqslant 1$ are satisfied by (5.2). ☐

Remark 5.3. The estimate (5.12) achieves square-root cancellation in the two-variable complete sum $\Sigma(K, \boldsymbol{b}, h)$. A weaker bound $\Sigma(K, \boldsymbol{b}, h) \ll q^{3/2}$ can be proved easily and quite generally directly (by fixing one of the variables) from, say, [13], but this yields a power saving over the trivial bound for $B(K, \boldsymbol{\alpha}_M, 1_{\mathcal{N}})$ (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 $M, N \asymp q^{1/2}$.

5.3. **Reduction to two-variable character sums.** We will now study the sums $\Sigma(K, \boldsymbol{b})$ 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

$$K(x) = q^{-1/2} \sideset{}{^*}\sum_{u \,(\mathrm{mod}\, q)} e_q(f(u) + xu),$$

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 (5.9) by this formula and performing the averaging over s, we obtain

$$\Sigma(K, \boldsymbol{b}, h) = q^{-1} \sum_{r \,(\mathrm{mod}\, q)} \sideset{}{^*}\sum_{(u,v,u',v') \in V_r(\mathbf{F}_q)} e_q(f(u) + f(v) - f(u') - f(v')),$$

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

$$r(u + v - u' - v') + b_1u + b_2v - (b_1'u' + b_2'v') + h = 0.$$

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 Σ_1 of these terms to $\Sigma(K, \boldsymbol{b}, h; q)$ of these terms is equal to

$$\Sigma_1 = q^{-1} \sideset{}{^*}\sum_{\substack{u,v,u',v' \,(\mathrm{mod}\, q)\\ u+v-u'-v'\neq 0}} e_q\big(f(u) + f(v) - f(u') - f(v')\big) = q^{-1}\Big(q^2|K(0)|^4 - q \sum_{r \,(\mathrm{mod}\, q)} |K(r)|^4\Big) \ll q.$$

We are left with the following 2-dimensional exponential sum over $\mathbf{F}_q$

$$S(f, h, \boldsymbol{b}) := \sum\sum_{(u,v,u',v')\in W(\mathbf{F}_q)} e_q(f(u) + f(v) - f(u') - f(v')),$$

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

$$\begin{cases} u + v = u' + v', \\ b_1 u + b_2 v = b_1' u' + b_2' v' - h. \end{cases}$$

We will prove the following estimate for these sums:

Theorem 5.4. *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 $h \in \mathbf{F}_q$, we have

$$\sum_{u,v} e_q\big(f(u) + f(v) - f(l_3(u, v) + h) - f(l_4(u, v) - h)\big) \ll q,$$

where the implied constant is absolute.

We will prove this in the next section. Assuming the result, we conclude the proof of Proposition 5.2 (and hence of Theorem 5.1) as follows: if $b_1' \neq b_2'$, we can write

$$S(f, h, \boldsymbol{b}) = \sum_{u,v} e_q\big(f(u) + f(v) - f(l_3(u, v) + h') - f(l_4(u, v) - h')\big),$$

where

$$l_3(u, v) = \frac{b_1 - b_2'}{b_1' - b_2'} u + \frac{b_2 - b_2'}{b_1' - b_2'} v, \qquad l_4(u, v) = \frac{b_1' - b_1}{b_1' - b_2'} u + \frac{b_1' - b_2}{b_1' - b_2'} v,$$

$$h' = \frac{h}{b_1' - b_2'}.$$

Simple checks show that the sets $\{l_1, l_2\}$ and $\{l_3, l_4\}$ thus defined coincide only if $\boldsymbol{b} \in \mathcal{B}^{\Delta}$. Hence we then get

$$S(f, h, \boldsymbol{b}) \ll q$$

for all $\boldsymbol{b} \in \mathcal{B} - \mathcal{B}^{\Delta}$ and all $h \in \mathbf{F}_q$.

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 $\boldsymbol{b}$ are also in $\mathcal{B}^{\Delta}$. ∎

5.4. **Estimate of two-variable character sums.** We prove Theorem 5.4 in this section. By a general criterion (due to Hooley [17, Theorem 5] and Katz [20, Cor. 4], see also [14, Prop. 2.1]), the desired estimate follows from the Riemann Hypothesis over finite fields of Deligne for any $h \in \mathbf{F}_q$ such that the rational function

$$F(U,V) = f(U) + f(V) - f(l_3(U,V)+h) - f(l_4(U,V)-h) \in \mathbf{F}_q(U,V)$$

is *not composed*, which means that it is not of the shape

$$F = Q \circ P,$$

where $P \in \bar{\mathbf{F}}_q(U,V)$ and where $Q \in \bar{\mathbf{F}}_q(T)$ 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 $\bar{\mathbf{F}}_q$, under the assumption that $\{l_1, l_2\} \neq \{l_3, l_4\}$.

In the following, we denote by $C \in \bar{\mathbf{F}}_q$ a non–zero constant, the value of which may change from one line to another. We follow closely the method of [14, Proposition 2.3] but first we make the birational change of variables

$$X = U/V,\ Y = V,$$

so that

$$\begin{aligned} F(U,V) &= f(XY) + f(Y) - f(Yl_3(X,1)+h) - f(Yl_4(X,1)-h) \\ &= \frac{1}{XY} + \frac{1}{Y} - \frac{1}{Yl_3(X,1)+h} - \frac{1}{Yl_4(X,1)-h}. \end{aligned}$$

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 $P_1(X,Y), P_2(X,Y) \in \bar{\mathbf{F}}_q[X,Y]$ are coprime polynomials in two variables and

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

are coprime polynomials in one variable (here $m(\lambda)$ and $m(\mu)$ denote the multiplicity of the zeros λ and μ). 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

$$q_1 > q_2, \tag{5.13}$$

which means that ∞ is a pole of Q. Our objective is then to show that

$$q_1 = 1. \tag{5.14}$$

We have the identity

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

In this latter expression, the numerator and denominator, $\mathrm{NUM}(X,Y)$ and $\mathrm{DEN}(X,Y)$, are coprime. We also have

$$F(XY,Y) = \frac{(Yl_3(X,1)+h)(Yl_4(X,1)-h)(1+X) - XY^2(l_3(X,1)+l_4(X,1))}{XY(Yl_3(X,1)+h)(Yl_4(X,1)-h)}. \tag{5.15}$$

By the assumption (5.13), 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 (5.15) are coprime and are equal to $C \cdot \mathrm{NUM}(X, Y)$ and $C \cdot \mathrm{DEN}(X, Y)$ 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

$$q_1 - q_2 = 1, \tag{5.16}$$

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

In particular if $q_2 = 0$, we obtain (5.14) and we are done.

Suppose now that $q_2 \geqslant 1$. 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 λ 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

$$\mathrm{NUM}(X, 0) = CP_1(X, 0)^{q_1} \text{ or } CP_2(X, 0)^{q_1},$$

but $\mathrm{NUM}(X, 0) = -h^2(1 + X)$ and therefore $q_1 = 1$. This contradicts the equality (5.16) and the assumption $q_2 \geqslant 1$.

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

$$F(XY, Y) = \frac{l_3(X, 1)l_4(X, 1)(1 + X) - X(l_3(X, 1) + l_4(X, 1))}{XYl_3(X, 1)l_4(X, 1)}. \tag{5.17}$$

Let us assume that $F(XY, Y) \neq 0$. The polynomials $\mathrm{NUM}(X, Y)$ and $\mathrm{DEN}(X, Y)$ divide the numerator and denominator of the right-hand side of (5.17), in particular $\mathrm{NUM}(X, Y)$ does not depend on Y. Suppose that $q_2 \geqslant 1$; as above (possibly up to a change of variable $T \mapsto T + \mu$), we may assume that $0 \notin \{\lambda \mid Q_1(\lambda) = 0\}$ 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

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

since the left-hand side only depends on X and the λ are $\neq 0$. 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 $\{l_3, l_4\} = \{l_1, l_2\}$. □

5.5. **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 establish the following proposition concerning bilinear sums of Kloosterman sums conditionally on a square root cancellation bound for a certain complete sum of products of Kloosterman sums in three variables, which we state as Conjecture 5.7.

Proposition 5.5 (Bilinear forms of Kloosterman sums). *Let q be a prime, $(a, q) = 1$, $\mathcal{N} \subset \mathbf{R}$ an interval of length N and $(\alpha_m)_m$, $(\beta_n)_n$ be two sequences of complex numbers supported respectively on $[1, M]$ and $\mathcal{N}$ and with ℓ_2-norms given by*

$$\|\alpha\|_2^2 = \sum_{m \leqslant M} |\alpha_m|^2, \ \|\beta\|_2^2 = \sum_{n \in \mathcal{N}} |\beta_n|^2.$$

Assuming that

$$1 \leqslant M, N \leqslant q, \ q^{\frac{1}{4}} \leqslant MN \leqslant q^{\frac{5}{4}} \text{ and } M \leqslant q^{\frac{1}{4}} N, \tag{5.18}$$

and that Conjecture 5.7 *holds, one has*

$$\sum_{m\leqslant M, n\in\mathcal{N}}\sum \alpha_m\beta_n \operatorname{Kl}_2(amn;q) \ll q^{\varepsilon}\|\alpha\|_2\|\beta\|_2(MN)^{1/2}\big(M^{-\frac{1}{2}} + q^{\frac{11}{64}}(MN)^{-\frac{3}{16}}\big)$$

for any $\varepsilon > 0$*, uniformly in* a*.*

Remark 5.6. The main reason to believe that the above bound should hold unconditionally is the unconditional bound obtained by Fouvry and Michel in [14, §VII], for bilinear forms of the shape

$$B(K, \boldsymbol{\alpha}_M, \boldsymbol{\beta}_N) = \sum_{m\leqslant M, n\leqslant N}\sum \alpha_m\beta_n K(mn) \tag{5.19}$$

for kernels $K(x)$ of the shape

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

for some fixed integer $k \neq 0, 1, 2$. They obtained the bound

$$\sum_{\substack{M<m\leqslant 2M\\ N<n\leqslant 2N}}\sum \alpha_m\beta_n K(mn) \ll_{\varepsilon,k} q^{\varepsilon}\|\alpha\|_2\|\beta\|_2(MN)^{1/2}(M^{-1/2} + q^{\frac{11}{64}}(MN)^{-\frac{3}{16}})$$

for any $\varepsilon > 0$ (actually 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 [14] yields the slightly stronger bound presented here.) This bound was ultimately a consequence of bounds for families of multivariable complete algebraic exponential sums which where obtained using the work of Deligne and Katz.

We establish Proposition 5.5 by repeating the argument of Fouvry and Michel. With $K(x) = \operatorname{Kl}_2(ax;q)$, the Cauchy–Schwarz inequality gives

$$\Big|\sum_{m\leqslant M,\ n\in\mathcal{N}}\sum \alpha_m\beta_n K(mn)\Big|^2 \leqslant \|\beta\|_2^2 \sum_{m_1,m_2\leqslant M}\sum \alpha_{m_1}\overline{\alpha}_{m_2}\sum_{n\in\mathcal{N}} K(m_1n)\overline{K}(m_2n) =: \|\beta\|_2^2(\Sigma^{=} + \Sigma^{\neq}),$$

where $\Sigma^{=}$ is the contribution of the diagonal terms $m_1 \equiv m_2 \pmod q$ and $\Sigma^{\neq}$ is the remaining off-diagonal contribution. The diagonal term is bounded by $O(\|\alpha\|_2^2 N)$. For the remaining terms we apply again Vinogradov's "shift by ab" trick: given $A, B \geqslant 1$ satisfying the conditions (5.7) (these will be satisfied by (5.18) after a suitable choice of A, B), the off-diagonal term $\Sigma^{\neq}$ is bounded by

$$\frac{1}{AB}\sum_{\substack{A<a\leqslant 2A\\ B<b\leqslant 2B}}\sum\ \sum_{\substack{m_1,m_2\leqslant M\\ m_1\not\equiv m_2 \,(\mathrm{mod}\, q)}}\sum \alpha_{m_1}\overline{\alpha}_{m_2}\sum_{n+ab\in\mathcal{N}} K(am_1(\overline{a}n+b))\overline{K}(am_2(\overline{a}n+b)),$$

which is itself bounded by

$$\ll \frac{q^{\varepsilon}}{AB}\sum_{\substack{r \,(\mathrm{mod}\, q),\ 1\leqslant s_1,s_2\leqslant AM\\ s_1\not\equiv s_2 \,(\mathrm{mod}\, q)}}\sum \nu(r,s_1,s_2)\Big|\sum_{B<b\leqslant 2B}\eta_b K(s_1(r+b))\overline{K}(s_2(r+b))\Big|,$$

where $|\eta_b| \leqslant 1$ and

$$\nu(r,s_1,s_2) = \sum_{\substack{A<a\leqslant 2A,\ m_1,m_2\leqslant M,\ n\in\mathcal{N}',\\ am_1=s_1,\ am_2=s_2,\ \overline{a}n\equiv r \,(\mathrm{mod}\, q)}}\cdots\sum |\alpha_{m_1}||\alpha_{m_2}|$$

with $\mathcal{N}' \supset \mathcal{N}$ as before.

By the same reasoning as above we have

$$\sum_{r,s_1,s_2}\sum\sum \nu(r,s_1,s_2) \ll AN\|\alpha\|_1^2 \leqslant AMN\|\alpha\|_2^2$$

and

$$\sum\sum\sum_{r,s_1,s_2} \nu(r,s_1,s_2)^2 \ll q^\varepsilon AN\|\alpha\|_2^4.$$

From these bounds and Hölder's inequality and (5.7), we obtain as in [14, Lemma 7.1] that $\Sigma^{\neq}$ is bounded by

$$\frac{q^\varepsilon}{AB}(AN)^{3/4}M^{1/2}\|\alpha\|_2^2$$
$$\Bigl(\sum_{\boldsymbol{b}}\Bigl|\sum_{r\,(\mathrm{mod}\,q)}\ \mathop{\sum\sum}_{\substack{1\leqslant s_1,s_2\leqslant AM\\ s_1\not\equiv s_2\,(\mathrm{mod}\,q)}}\prod_{i=1}^{2}K(s_1(r+b_i))\overline{K}(s_2(r+b_i))\overline{K(s_1(r+b_{i+2}))\overline{K}(s_2(r+b_{i+2}))}\Bigr|\Bigr)^{1/4}$$

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

$$\mathcal{B}^\Delta\subset\mathcal{B}$$

be the "diagonal" set of elements $\boldsymbol{b}$ for which some pair (b_i,b_j) $(i,j\leqslant 4)$ 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 $\boldsymbol{b}\in\mathcal{B}^\Delta$ we use the trivial bound to obtain

$$\sum_{\boldsymbol{b}\in\mathcal{B}^\Delta}\Bigl|\sum_{r\,(\mathrm{mod}\,q)}\ \mathop{\sum\sum}_{s_1,s_2}\cdots\Bigr|\ll qA^2B^2M^2.$$

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

$$\delta(s_1\not\equiv s_2\,(\mathrm{mod}\,q))=1-\frac{1}{q}\sum_{\lambda\,(\mathrm{mod}\,q)}e_q(\lambda(s_1-s_2)).$$

We then complete the s_1, s_2 sums, also using additive characters: for $\lambda,\mu_1,\mu_2\in\mathbf{Z}/q\mathbf{Z}$ let

$$\mathcal{S}(r,\lambda;q)=\sum_{s\,(\mathrm{mod}\,q)}K(s(r+b_1))K(s(r+b_2))\overline{K(s(r+b_3))K(s(r+b_4))}e_q(\lambda s),$$

$$\mathcal{R}(\mu_1,\mu_2;q)=\sum_{r\,(\mathrm{mod}\,q)}\mathcal{S}(r,\mu_1;q)\overline{\mathcal{S}(r,\mu_2;q)},$$

and

$$\Sigma(\boldsymbol{b},\mu_1,\mu_2;q)=\mathcal{R}(\mu_1,\mu_2;q)-\frac{1}{q}\sum_{\lambda\,(\mathrm{mod}\,q)}\mathcal{R}(\mu_1+\lambda,\mu_2+\lambda;q).$$

5.6. **Correlation sums.** We now formulate a conjectural bound on the sum $\Sigma(\boldsymbol{b},\mu_1,\mu_2;q)$. To motivate this conjecture, let us briefly examine the structure and the significance of the sums $\mathcal{S},\mathcal{R}$ and Σ.

Given r,λ and $\boldsymbol{b}$, the sum $\mathcal{S}(r,\lambda;q)$ is a one-variable sum of a product of the four Kloosterman sums $s\mapsto K(s(r+b_i))$, $i=1,\ldots,4$, and the additive phase $e_q(\lambda s)$. It is well known that as s varies, the Kloosterman sums oscillate rather wildly and moreover, for distinct values of the b_i's, these oscillations are independent of each other, so that typically square-root cancellation occurs:

$$\mathcal{S}(r,\lambda;q)=O(q^{1/2}).$$

For this and more general sums of that type, we refer to the article [13] which builds crucially on the works of Deligne and Katz [6, 21, 22]. The sum $\mathcal{R}(\mu_1,\mu_2;q)$ deals with the variation of the sums $\mathcal{S}(r,\lambda;q)$; more precisely it measures to which extent the functions $r\mapsto\mathcal{S}(r,\mu_1;q)$ and

$r \mapsto \mathcal{S}(r, \mu_2; q)$ correlate. If there is no correlation, it is then natural to expect from Deligne's formalism of weights that square-root cancellation occurs again, and so

$$\mathcal{R}(\mu_1, \mu_2; q) = O(q^{3/2}).$$

On the other hand, when the sums *do* correlate, one expect $\mathcal{R}(\mu_1, \mu_2; q)$ to be the sum of a main term of size q^2 and of an error term, more precisely

$$\mathcal{R}(\mu_1, \mu_2; q) = q^2 + O(q^{3/2}).$$

This is essentially the content of the conjecture below which also incorporates the correlation and non-correlation cases; see [24] for further discussions on this conjecture.

Conjecture 5.7. *There exists a constant C such that for any prime q, every integer a coprime with q, every $\mu_1, \mu_2 \in \mathbf{F}_q$ and every $\boldsymbol{b} \in \mathcal{B}^{gen} := \mathcal{B} \setminus \mathcal{B}^{\Delta}$ we have*

$$|\Sigma(\boldsymbol{b}, \mu_1, \mu_2; q)| \leqslant C q^{3/2};$$

here the sum Σ is the sum relative to the function $K(x) = \mathrm{Kl}_2(ax; q)$.

If we assume Conjecture 5.7, we obtain that

$$\sum_{\boldsymbol{b} \in \mathcal{B}^{gen}} \Big| \sum_{r \,(\mathrm{mod}\, q)} \sum_{s_1, s_2} \sum \cdots \Big| \ll q^{\varepsilon} B^4 q^{3/2}.$$

Hence, under this assumption, we have

$$\Sigma^{\neq} \ll \frac{q^{\varepsilon}}{AB} (AN)^{3/4} M^{1/2} \|\alpha\|_2^2 (A^2 B^2 M^2 q + B^4 q^{3/2})^{1/4}.$$

We may choose (see [14, p.128])

$$A = q^{1/8} M^{-1/2} N^{1/2}, \ B = q^{-1/8} (MN)^{1/2},$$

for which (5.7) as well as $A, B \geqslant 1$ are satisfied by (5.18). Combining this bound with that for $\Sigma^{=}$, we conclude that Proposition 5.5 follows from Conjecture 5.7. ☐

6. Evaluation of moments of L-functions

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

6.1. **First steps.** Let f, g be either 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

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

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 χ 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 (2.6), we have

$$L(f\otimes\chi,1/2)\overline{L(g\otimes\chi,1/2)}=2\sum_{m,n\geqslant 1}\frac{\lambda_f(m)\lambda_g(n)}{(mn)^{1/2}}\chi(m)\bar{\chi}(n)V_{f,g,\sigma}\left(\frac{mn}{q^2}\right),$$

where the function $V_{f,g,\sigma}$ is given by (2.7).

We now average over χ of parity σ. The orthogonality relation for these characters is

$$\frac{2}{q-1}\sum_{\substack{\chi\,(\mathrm{mod}\,q)\\ \chi(-1)=\sigma}}\chi(m)\overline{\chi(n)}=\delta_{m\equiv n\,(\mathrm{mod}\,q)}+\sigma\delta_{m\equiv -n\,(\mathrm{mod}\,q)} \tag{6.1}$$

for q prime and any integers m and n such that $(mn,q)=1$. Inserted in the above formula, it yields

$$M_{f,g,+1}(q)=B^+_{f,g,+1}(q)+B^-_{f,g,+1}(q),\qquad M_{f,g,-1}(q)=B^+_{f,g,-1}(q)-B^-_{f,g,-1}(q),$$

where

$$B^{\pm}_{f,g,\sigma}(q)=\sum_{\substack{m\equiv\pm n\,(\mathrm{mod}\,q)\\ (mn,q)=1}}\frac{\lambda_f(m)\lambda_g(n)}{(mn)^{1/2}}V_{f,g,\sigma}\left(\frac{mn}{q^2}\right)-\frac{1}{\varphi^*(q)}\sum_{(mn,q)=1}\frac{\lambda_f(m)\lambda_g(n)}{(mn)^{1/2}}V_{f,g,\sigma}\left(\frac{mn}{q^2}\right) \tag{6.2}$$

(indeed, the second term in (6.2) 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 $\mathrm{MT}^d_{f,g,\sigma}(q)$ is given by the contribution of $n=m$ in $B^+_{f,g,\sigma}(q)$ (note that $n=m$, $m\equiv -n\neq 0\,(\mathrm{mod}\ q)$ is impossible for q odd). By Mellin inversion and a contour shift, we can compute explicitly:

$$\begin{aligned}\mathrm{MT}^d_{f,g,\sigma}(q)&=\sum_{\substack{m\geqslant 1\\ (m,q)=1}}\frac{\lambda_f(m)\lambda_g(m)}{m}V_{f,g,\sigma}\left(\frac{m^2}{q^2}\right)\\ &=\operatorname*{res}_{s=0}\frac{L_\infty(f\otimes\chi,1/2+s)L_\infty(g\otimes\overline{\chi},1/2+s)}{L_\infty(f\otimes\chi,1/2)L_\infty(g\otimes\overline{\chi},1/2)}\frac{L^{(q)}(f\otimes g,1+2s)}{\zeta^{(q)}(2+4s)}\frac{q^{2s}}{s}+O(q^{-1/2+\varepsilon}),\end{aligned}$$

for any $\varepsilon>0$, where χ 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

$$\mathrm{MT}^d_{f,g,\sigma}(q)=\mathrm{MT}^0_{f,g,\sigma}(q)+O(q^{\varepsilon-1/2}),$$

where

$$\mathrm{MT}^0_{f,g,\sigma}(q)=\begin{cases}P_{1,f,\sigma}(\log q) & \text{for } P_{1,f,\sigma}(X) \text{ a degree 1 polynomial if } f=g \text{ is cuspidal,}\\ \frac{L(f\otimes g,1)}{\zeta(2)} & \text{if } f\neq g \text{ are both cuspidal,}\\ \frac{L(f,1)^2}{\zeta(2)} & \text{if } f \text{ is cuspidal and } g=E,\\ P_{4,\sigma}(\log q) & \text{for } P_{4,\sigma}(X) \text{ a polynomial of degree 4 if } f=g=E.\end{cases}$$

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.

We apply a partition of unity to the m, n variables and are led to evaluate the dyadic sums

$$\sum_{M,N\geqslant 1}^{\mathrm{dy}} \sum_{\substack{m\equiv \pm n \,(\mathrm{mod}\, q)\\ m\neq n}} \frac{\lambda_f(m)\lambda_g(n)}{(mn)^{1/2}} V_{f,g,\pm 1}\Big(\frac{mn}{q^2}\Big) W_1\Big(\frac{m}{M}\Big) W_2\Big(\frac{n}{N}\Big)$$
$$- \sum_{M,N\geqslant 1}^{\mathrm{dy}} \frac{1}{q} \sum_{m,n} \frac{\lambda_f(m)\lambda_g(n)}{(mn)^{1/2}} V_{f,g,\pm 1}\left(\frac{mn}{q^2}\right) W_1\Big(\frac{m}{M}\Big) W_2\Big(\frac{n}{N}\Big),$$

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

$$1 \leqslant MN \leqslant q^{2+\varepsilon}. \tag{6.3}$$

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 [3, 29]): using the definition (2.7) of $V_{f,g,\pm 1}(x)$ as a Mellin transform, we shift the line of integration to $\mathfrak{Re}\, s = \varepsilon$ and approximate

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

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

$$B_{f,g}^{\pm}(M, N) = \frac{1}{(MN)^{1/2}} \sum_{\substack{m\equiv \pm n \,(\mathrm{mod}\, q)\\ m\neq n}} \lambda_f(m)\lambda_g(n) W_1\Big(\frac{m}{M}\Big) W_2\Big(\frac{n}{N}\Big) - \frac{1}{q(MN)^{1/2}} \sum_{m,n} \lambda_f(m)\lambda_g(n) W_1\Big(\frac{m}{M}\Big) W_2\Big(\frac{n}{N}\Big), \tag{6.4}$$

with new test functions W_1, W_2 (which depend on s) satisfying (1.14), since $s = \varepsilon + it$ and $|t| < \log^2 q$.

As explained in Section 1.2, our objective is then to show that (assuming Conjecture 5.7 if both f and g are cuspidal)

$$B_{f,g}^{\pm}(M, N) = \delta_{f=g=E} \mathrm{MT}_{E,E}^{od,\pm}(M, N) + O(q^{-\eta+\varepsilon}) \tag{6.5}$$

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 $\mu, \nu, \mu^*, \nu^*, \varrho$ by

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

By (6.3) we have

$$0 \leqslant \mu + \nu \leqslant 2 + \varepsilon.$$

We consider the three main results in turn.

6.2. **The case f and g cuspidal.** Let $\eta = 1/144$. We prove (6.5) subject to Conjecture 5.7.

By symmetry, we may assume that $0 \leqslant \mu \leqslant \nu \leqslant 2+\varepsilon$ (up to exchanging the roles of f and g). We review the various bounds that are available and the ranges of the parameters μ, ν for which (6.5) holds.

The trivial bound. By (3.1), we obtain (6.5) immediately if $\mu+\nu \leqslant 2-2\eta-2\theta\nu$. We can therefore assume that

$$2-2\eta-2\theta\nu \leqslant \mu+\nu \leqslant 2+\varepsilon \tag{6.6}$$

and therefore

$$-2\eta-2\theta\nu \leqslant \mu-\nu^* \leqslant \varepsilon. \tag{6.7}$$

The shifted convolution bound. From (3.2), we obtain that (6.5) holds unless

$$1-4\eta \leqslant \nu-\mu \text{ or equivalently } \mu+\nu^* \leqslant 1+4\eta. \tag{6.8}$$

The trivial Voronoi summation bound. By (6.6) and (6.8), we then have

$$\nu \geqslant 3/2-3\eta-2\theta \geqslant 1+\frac{1}{1000},$$

in which case the condition $n \neq m$ is void (since $\mu \leqslant 1+\varepsilon/2$); it is then natural to apply the Voronoi summation formula (Lemma 2.3) to the (long) n-variable. To this end, we detect the condition $m \equiv \pm n \,(\mathrm{mod}\, q)$ by additive characters. The trivial character cancels the second term on the right hand side of (6.4), and one obtains the formula

$$B^{\pm}_{f,g}(M,N) = \frac{1}{(qMN^*)^{1/2}} \sum_{m,n\geqslant 1} \lambda_f(m)\lambda_g(n) W_1\left(\frac{m}{M}\right) \frac{1}{N}\widetilde{W_{2,N}}\left(\frac{n}{q^2}\right) \mathrm{Kl}_2(\pm mn;q) \tag{6.9}$$

with $N^* = q^2/N$, where we use the notation of Lemma 2.4. In particular, by this lemma, the function

$$y \mapsto \frac{1}{N}\widetilde{W_{2,N}}\Big(\frac{y}{q^2}\Big)$$

decays rapidly for $y \geqslant q^{\varepsilon}N^*$ and the contribution to $B^{\pm}_{f,g}(M,N)$ of those n that satisfy $n \geqslant q^{\varepsilon}N^*$ is negligible. By a partition of unity (using Lemma 1.6), we can decompose (6.9) into a sum of $O(\log q)$ terms of the shape

$$C^{\pm}(M,N') = \frac{(1+N^*/N')^{2\theta+\varepsilon}}{(qMN^*)^{1/2}} \sum_{m,n\geqslant 1} \lambda_f(m)\lambda_g(n) W_1\left(\frac{m}{M}\right) W_2\left(\frac{n}{N'}\right) \mathrm{Kl}_2(\pm mn;q),$$

with W_1, W_2 satisfying (1.14) and $N' = q^{\nu'} \leqslant q^{\varepsilon}N^*$.

By Weil's bound for Kloosterman sums $|\,\mathrm{Kl}_2(\pm mn;q)| \leqslant 2$ and (2.4) we have the trivial bound

$$C^{\pm}(M,N') \ll q^{\varepsilon}(MN^*/q)^{1/2} = q^{\varepsilon+\frac{\mu+\nu^*-1}{2}}, \tag{6.10}$$

which establishes (6.5) unless (cf. (6.8) for the upper bound)

$$1-2\eta \leqslant \mu+\nu^* \leqslant 1+4\eta. \tag{6.11}$$

This together with (6.7) implies that

$$\mu \geqslant \frac{1}{2}-2\eta-\theta\nu.$$

The bilinear sum bound. Now, applying Proposition 5.5 (whose conclusion we recall is conditional on Conjecture 5.7) with $\mathrm{M}=2M,\ \mathrm{N}=2N^*$,

$$(\alpha_m)_{m\leqslant 2M}=\lambda_f(m)W_1\left(\frac{m}{M}\right),\ (\beta_n)_{n\leqslant 2N^*}=\lambda_g(n)W_2\left(\frac{n}{N'}\right),$$

and whose assumptions (5.18) are satisfied by the preceding display, we have (using (2.4))

$$C^{\pm}(M,N')\ll q^{\varepsilon}\Big(\frac{MN^*}{q}\Big)^{1/2}(M^{-1/2}+q^{11/64}(MN^*)^{-3/16})\ll q^{\varepsilon}(q^{3\eta+\frac{1}{2}\theta\nu-\frac{1}{4}}+q^{-\frac{1}{64}+\frac{5}{4}\eta})\ll q^{\varepsilon-\eta}$$

for $\eta=1/144$ and $\nu\leqslant 2+\varepsilon$. This concludes the proof of Theorem 1.3. ☐

6.3. **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 $\mu\leqslant\nu$. Moreover, the Ramanujan–Petersson conjecture is trivially true, so we may apply (3.1) to obtain (6.5) if

$$\mu+\nu\leqslant 2-2\eta,\ 2\eta\leqslant\nu-\mu.$$

The grouping and the analysis of the main terms was done in [29], so we will focus on the error term. Applying first (3.2) we obtain

$$\mathrm{ET}^{\pm}_{E,E}(M,N)\ll q^{-\eta+\varepsilon},$$

as desired, unless

$$\mu+\nu^*\leqslant 1+4\eta,$$

which we assume from now on.

In this remaining range, the off-diagonal main term $\mathrm{MT}^{od,\pm}_{E,E}(M,N)\ll q^{-7/16+\varepsilon}$ is small (cf. the second term in Proposition 3.1), so that we can assume (6.7), and it suffices to prove the estimate

$$B^{\pm}_{E,E}(M,N)\ll q^{-\eta+\varepsilon}. \tag{6.12}$$

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 (note that here we have $\vartheta=0$ in the notation of Lemma 2.4). This expresses the sum $B^{\pm}_{E,E}(M,N)$ into a main term and two additional terms. As in (6.21), the main term is $O(q^{-1+\varepsilon})$, while the error terms decompose into $O(\log q)$ terms of the shape

$$\frac{1}{(qMN^*)^{1/2}}\sum_{m,n}\sum d(m)d(n)W_1\left(\frac{m}{M}\right)W_2\left(\frac{n}{N'}\right)\mathrm{Kl}_2(\pm mn;q), \tag{6.13}$$

where W_1, W_2 satisfy (1.14) (the definition of W_2 has changed from its preceding appearance).

A trivial estimate shows that (6.12) holds unless

$$1-2\eta\leqslant\mu+\nu^*\leqslant 1+4\eta, \tag{6.14}$$

which we then assume.

We further decompose (6.13) into $O(\log^4 q)$ terms of the form

$$\frac{1}{(qMN^*)^{1/2}}\sum_{m_1,m_2,n_1,n_2}\sum W_1\left(\frac{m_1m_2}{M}\right)W_2\left(\frac{n_1n_2}{N'}\right) \times W\left(\frac{m_1}{M_1}\right)W\left(\frac{m_2}{M_2}\right)W\left(\frac{n_1}{M_3}\right)W\left(\frac{n_2}{M_4}\right)\mathrm{Kl}_2(\pm m_1m_2n_1n_2;q) \tag{6.15}$$

with

$$M_1M_2=M,\quad M_3M_4=N'\leqslant N^*.$$

In (6.15), 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 $\varepsilon > 0$, the contribution to the integral of the s such that $|s| \geqslant q^{\varepsilon}$ is negligible, by (1.14) 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

$$(6.16)\quad S^{\pm}(M_1, M_2, M_3, M_4) = \frac{1}{(qMN^*)^{1/2}} \sum\sum_{m_1,m_2,n_1,n_2} W_1\left(\frac{m_1}{M_1}\right) W_2\left(\frac{m_2}{M_2}\right) \times W_3\left(\frac{m_3}{M_3}\right) W_4\left(\frac{m_4}{M_4}\right) \mathrm{Kl}_2(\pm m_1m_2m_3m_4; q),$$

where the W_i satisfy (1.14) and hence (5.4) for $Q = q^{\varepsilon}$, and the M_i written in the shape $M_i = q^{\mu_i}$, $i = 1, 2, 3, 4$, satisfy

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

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 (5.5) of Theorem 5.1, with $MN = M_3M_4$, and we sum trivially over m_1 and m_2. If this is not the case, it is possible to factor the product $m_1m_2n_1n_2$ into a product mn in such a way that an application of the general bilinear estimate (5.1) is beneficial.

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

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

If

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

we apply (5.5) with $MN = M_3M_4$ and sum trivially over m_1 and m_2, obtaining the bound

$$(6.17)\qquad q^{-(A+1)\varepsilon} S^{\pm}(M_1, M_2, M_3, M_4) \ll q^{\frac{1}{2}(\mu+\nu^*-\frac{5}{4})} + q^{-\frac{\delta}{2}}$$

for the constant A occurring in (5.5).

On the other hand, if

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

then at least one of μ_2 and $\mu_1 + \mu_2$ is contained in the interval

$$(6.18)\qquad \left[\frac{1}{4} - \delta, \frac{1}{3}(\mu + \nu^*)\right],$$

since $\mu_2 \leqslant (\mu + \nu^*)/3$ and $\mu_1 \leqslant \mu_2$. Let u be one of the numbers μ_2 or $\mu_1 + \mu_2$ satisfying this condition. We then apply (5.1) with

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

there (notice that (6.18) and (6.14) guarantee the assumption $M, N \leqslant q$ in (5.1)), and we obtain the bound

$$q^{-\varepsilon} S^{\pm}(M_1, M_2, M_3, M_4) \ll q^{\frac{1}{2}(\mu+\nu^*-1-u)} + q^{-\frac{1}{2}(\frac{1}{2}-u)} \ll q^{\frac{1}{2}(\mu+\nu^*-5/4+\delta)} + q^{\frac{1}{6}(\mu+\nu^*)-\frac{1}{4}}.$$

We choose the value of δ by comparing the second term of the bound (6.17) with the first of the bound (6.3). Precisely, we take

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

and therefore we get

$$q^{-\varepsilon} S^{\pm}(M_1, M_2, M_3, M_4) \ll q^{\frac{1}{4}(\mu+\nu^*-\frac{5}{4})} + q^{\frac{1}{6}(\mu+\nu^*)-\frac{1}{4}}$$

under the assumption $\mu + \nu^* \leqslant 5/4$. This is indeed valid, by (6.14), since $\eta \leqslant 1/16$. Therefore, by (6.14), we find that

$$q^{-(A+1)\varepsilon} S^{\pm}(M_1, M_2, M_3, M_4) \ll q^{-\frac{1}{16}+\eta} + q^{-\frac{1}{12}+\frac{2}{3}\eta} \ll q^{-\frac{1}{32}},$$

as desired. □

Remark 6.1. The same strategy, but with (1.7) instead of (3.2), gives a saving of $q^{-1/24}$ if $\theta = 0$ in (1.7).

6.4. **The mixed case.** We will now prove Theorem 1.2 and consider the mixed case where f is cuspidal and $g = E$. Let $\eta = 1/68$. In the present case, M and N are not symmetric, and so we will need to distinguish the cases where $\mu \leqslant \nu$ and $\mu > \nu$ on several occasions.

Firstly, applying (3.2), we see that (6.5) holds unless

$$|\nu - \mu| \geqslant 1 - 4\eta, \tag{6.19}$$

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 condition in (6.4) with additive characters and cancel the contribution of the trivial character with the second term. This gives

$$B^{\pm}_{f,E}(M, N) = \frac{1}{q(MN)^{1/2}} \sum_{\substack{a \ (\mathrm{mod}\ q) \\ a \not\equiv 0}} \sum_{m,n} \lambda_f(m) d(n) e\left(\frac{a(m \mp n)}{q}\right) W_1\Big(\frac{m}{M}\Big) W_2\Big(\frac{n}{N}\Big). \tag{6.20}$$

If, for instance, $N \geqslant M$, then applying Lemma 2.3 to the n-sum yields a "main term"

$$\frac{1}{q^2(MN)^{1/2}} \Big(\int_0^{+\infty} (\log x + 2\gamma - 2\log q) W_2\Big(\frac{x}{N}\Big) dx \Big) \sum_{m \geqslant 1} \lambda_f(m) r(m; q) W_1\Big(\frac{m}{M}\Big) \tag{6.21}$$

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

$$\frac{1}{q(MN^*)^{1/2}} \sum_{m,n \geqslant 1} \lambda_f(m) d(n) W_1\left(\frac{m}{M}\right) \frac{1}{N} (\widetilde{W_{2,N}})_{\sigma} \left(\frac{n}{q^2}\right) S(m, \pm\sigma n; q)$$

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 (using (2.4)), we obtain the bound

$$B^{\pm}_{f,E}(M, N) \ll q^{\varepsilon} \left(\frac{q \min(M, N)}{\max(M, N)}\right)^{1/2}. \tag{6.22}$$

In particular, (6.5) holds unless

$$|\nu - \mu| \leqslant 1 + 2\eta, \tag{6.23}$$

which we assume from now on. We proceed to derive, by various methods depending on whether $M > N$ or $M \leqslant N$, more elaborate bounds that allow us to treat the range where (6.19) and (6.23) are satisfied.

6.4.1. *The case $M \leqslant N$.* If $N \leqslant q$, then by (6.19), we see that $M \ll q^{4\eta+\varepsilon}$, so that (3.1) suffices to prove (6.5). From now on we assume $N \geqslant q$. As the Ramanujan–Petersson conjecture is trivially true for the divisor function, (3.1) holds with $\theta_g = 0$, and hence (6.7) holds in the stronger form

$$-2\eta \leqslant \mu - \nu^* \leqslant \varepsilon. \tag{6.24}$$

First, we observe that (6.23) and (6.24) imply $\mu \geqslant 1/2 - 2\eta > 2/5$, so that the second term in (6.4) is negligible. In the first term, we open the divisor function, apply smooth partitions of unity and are left with bounding the triple sum

$$C(M,N,N_1,N_2) := \frac{1}{\sqrt{MN}} \sum_{n_1n_2 \equiv \pm m \,(\mathrm{mod}\, q)} \lambda_f(m) W_1(n_1/N_1) W_2(n_2/N_2) W_3(m/M), \tag{6.25}$$

where

$$N_1N_2 = N, \quad N_1 \leqslant N_2 \tag{6.26}$$

and W_1, W_2, W_3 are (new) smooth, compactly supported weight functions satisfying (1.14). We can now apply Proposition 4.1, getting

$$C(M,N,N_1,N_2) \ll q^{\varepsilon} \cdot \begin{cases} \frac{\sqrt{MN}}{q^{2-\theta}} + \min\left(\frac{(Mq)^{1/2}}{N^{1/2}} + \frac{N_1M^{1/2}}{qN^{1/2}}, \frac{q^{1/4}}{N_1^{1/2}} + \frac{M^{1/2}}{N_2^{1/2}} + \frac{N_1M^{1/2}}{qN^{1/2}}, \frac{M^{1/2}N_1}{N^{1/2}}\right), \\ \frac{\sqrt{MN}}{q^{2-\theta}} + \min\left(\frac{N_1^2}{(MN)^{1/2}}, \frac{N^{1/6}N_1q^{1/2}}{M^{2/3}N_2}\right) + \frac{M^{1/2}}{N^{1/2}} + \frac{M^{1/2}N_1}{qN^{1/2}} + \frac{M^{3/2}}{N_2N^{1/2}}. \end{cases} \tag{6.27}$$

The term $\sqrt{MN}/q^{2-\theta} \ll q^{-3/4+\varepsilon}$ 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 (6.24) and (6.19) that $\mu \leqslant 1/2 + 2\eta < 3/5$, so that in particular $(m,q) = (n_1n_2, q) = 1$ in (6.25). We obtain

$$C(M,N,N_1,N_2) = \frac{1}{\sqrt{MN}} \frac{N}{q^2} \sum_{m,h_1,h_2} \lambda_f(m) W_3(m/M) W_1^{\dagger}(h_1N_1/q) W_2^{\dagger}(h_2N_2/q) S(\pm m h_1, h_2; q),$$

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 $q \mid h_1h_2$ is negligible. After a smooth partition of unity, we are left with $O(\log^2 q)$ terms of the form

$$C'(M,N,N_1,N_2) := \frac{1}{\sqrt{qMN_1^{\circ}N_2^{\circ}}} \sum_{m,h_1,h_2} \lambda_f(m) W_3(m/M) W_1(h_1/N_1') W_2(h_2/N_2') \,\mathrm{Kl}_2(\pm m h_1 h_2; q),$$

where

$$N_1' \leqslant N_1^{\circ}, \quad N_2' \leqslant N_2^{\circ}, \quad N_1^{\circ} = q/N_1, \quad N_2^{\circ} = q/N_2, \tag{6.28}$$

and W_1, W_2, W_3 are (new) smooth, compactly supported weight functions satisfying (1.14). Notice that $N_1^{\circ} \geqslant N_2^{\circ}$. We can now use our results on multi-linear forms in Kloosterman sums as developed in Section 5. In particular, we can apply the bound (5.1) with $(M,N) \leftarrow (N_2', MN_1')$ in the notation of Theorem 5.1, or the bound (5.3) with $(M,N) \leftarrow (MN_2', N_1')$. This gives (using (2.4) several times)

$$C'(M,N,N_1,N_2) \ll q^{\varepsilon} \frac{MN_1'N_2'}{\sqrt{qMN_1^{\circ}N_2^{\circ}}} ((N_2')^{-1/2} + q^{1/4}(MN_1')^{-1/2}), \quad \text{if } MN_1' \leqslant q, \tag{6.29}$$

and

$$C'(M,N,N_1,N_2) \ll q^{\varepsilon} \frac{MN_1'N_2'}{\sqrt{qMN_1^{\circ}N_2^{\circ}}} (q^{1/4}(MN_2')^{-1/6}(N_1')^{-5/12}), \quad \text{if } MN_2' \leqslant (N_1')^2, \tag{6.30}$$

since the condition $MN_1'N_2' \leqslant q^{3/2}$ and $N_2', MN_2', N_1' \leqslant q$ are automatic by (6.24), (6.19), (6.26) and (6.28).

Combining all estimates we have derived so far, that is (3.1) with $\theta_g = 0$, (3.2), (6.22), (6.27), (6.29) and (6.30), we need to find the maximum of the piecewise linear function

$$\begin{aligned}
\min\Big(&\tfrac{\mu+\nu}{2}-1, \max\Big(\tfrac{\nu-\mu-1}{2}, \tfrac{\nu-\mu-1}{4}\Big), \tfrac{1+\mu-\nu}{2},\\
&\max\Big(\tfrac{\mu+1-\nu}{2}, \tfrac{2\nu_1+\mu-2-\nu}{2}\Big), \max\Big(\tfrac{1}{4}-\tfrac{\nu_1}{2}, \tfrac{\mu-\nu_2}{2}, \tfrac{2\nu_1+\mu-\nu-2}{2}\Big), \tfrac{\mu+2\nu_1-\nu}{2},\\
&\max\Big(\min\Big(2\nu_1-\tfrac{\mu+\nu}{2}, \tfrac{\nu}{6}+\nu_1+\tfrac{1}{2}-\nu_2-\tfrac{2\mu}{3}\Big), \tfrac{\mu-\nu}{2}, \tfrac{\mu+2\nu_1-2-\nu}{2}, \tfrac{3\mu}{2}-\nu_2-\tfrac{\nu}{2}\Big),\\
&\Big(\tfrac{2\mu+2\nu_1'+2\nu_2'-1-\nu_1^\circ-\nu_2^\circ-\mu}{2}+\max\Big(-\tfrac{\nu_2'}{2}, \tfrac{1}{4}-\tfrac{\mu+\nu_1'}{2}\Big)\Big)\delta_{\mu+\nu_1'\leqslant 1},\\
&\Big(\tfrac{2\mu+2\nu_1'+2\nu_2'-1-\nu_1^\circ-\nu_2^\circ-\mu}{2}+\tfrac{1}{4}-\tfrac{\mu+\nu_2'}{6}-\tfrac{5\nu_1'}{12}\Big)\delta_{\mu+\nu_2'\leqslant 2\nu_1'}\Big)
\end{aligned}$$

subject to the constraints

$$0\leqslant \mu\leqslant \nu, \quad \mu+\nu\leqslant 2, \quad \nu_1+\nu_2=\nu, \quad 0\leqslant \nu_1\leqslant \nu_2, \quad 0\leqslant \nu_1'\leqslant \nu_1^\circ = 1-\nu_1, \quad 0\leqslant \nu_2'\leqslant \nu_2^\circ = 1-\nu_2.$$

(Of course this expression can be simplified quite a bit.) This is a linear optimization problem that can be solved exactly by computer in a finite search. One obtains that the maximum $-1/68$ is attained at $\mu = 161/306$, $\nu = 449/306$, $(\nu_1, \nu_2) = (9/17, 287/306)$ and (unsurprisingly) $\nu_1' = \nu_1^\circ$, $\nu_2' = \nu_2^\circ$. The Mathematica code is available after the bibliography.

6.4.2. *The case $M \geqslant N$.* We now assume $\mu \geqslant \nu$ and observe that (6.19) and (6.23) are still in force. (However, we cannot use (6.24).)

In the present case it turns out to be most efficient to apply Voronoi summation in (6.20) 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 (6.21) and trivially bounded by $O(q^{-1+\varepsilon})$, which is acceptable. Applying Lemma 2.4 and the usual partition of unity to the remaining terms in the Voronoi formula, we are left with bounding

$$\begin{aligned}
\tilde{B}^{\pm}_{f,E}(M,N) :=& \frac{1+(M^*/M')^{2\theta}}{\sqrt{M^*N^*}}\Big|\sum_{m\equiv \pm n \,(\mathrm{mod}\, q)} \lambda_f(m)d(n)W_1\Big(\frac{m}{M'}\Big)W_2\Big(\frac{n}{N'}\Big)\Big|\\
&+\frac{1+(M^*/M')^{2\theta}}{q\sqrt{M^*N^*}}\Big|\sum_{m,n}\lambda_f(m)d(n)W_1\Big(\frac{m}{M'}\Big)W_2\Big(\frac{n}{N'}\Big)\Big|,
\end{aligned}$$

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 (1.14). The second term is negligible unless $M' \leqslant q^{\varepsilon}$, in which case it is trivially bounded by $O(q^{\varepsilon-1}(M/N)^{1/2})$. By (6.23), this is $O(q^{\varepsilon-1/2+\eta})$, which is acceptable. For the first term, we can apply Corollary 4.3 in addition to the other bounds (3.1), (3.2) and (6.22). This leads to the linear program to maximize

$$\min\Big(\tfrac{\mu+\nu}{2}-1+\theta\mu, \max\Big(\tfrac{\mu-\nu-1}{2}, \tfrac{\mu-\nu-1}{4}\Big), \tfrac{1+\nu-\mu}{2},$$
$$\theta(\mu^*-\mu')+\max\Big(\tfrac{2\mu'-\mu^*-\nu^*+1}{2}, \tfrac{2\nu_1+2\mu'-2-\mu^*-\nu^*}{2}\Big),$$
$$\theta(\mu^*-\mu')+\max\Big(\tfrac{\mu'+\nu_2+\frac{1}{2}-\nu^*-\mu^*}{2}, \tfrac{2\mu'+\nu_1-\mu^*-\nu^*}{2}, \tfrac{2\nu_1+2\mu'-2-\mu^*-\nu^*}{2}\Big),$$
$$\theta(\mu^*-\mu')+\theta(\mu^*-\mu')+\mu'+\nu_1-\tfrac{\mu^*+\nu^*}{2},$$
$$\theta(\mu^*-\mu')\max\Big(\min\Big(2\nu_1-\tfrac{\mu^*+\nu^*}{2}, \tfrac{2}{3}\nu'+\nu_1+\tfrac{1}{2}-\tfrac{1}{6}\mu'-\nu_2-\tfrac{\mu^*+\nu^*}{2}\Big),$$
$$\tfrac{2\mu'-\mu^*-\nu^*}{2}, \tfrac{2\mu'+2\nu_1-2-\mu^*-\nu^*}{2}, 2\mu'-\nu_2-\tfrac{\mu^*+\nu^*}{2}\Big)\Big)$$

subject to the constraints

$$0\leqslant\nu\leqslant\mu,\quad \mu+\nu\leqslant 2,\quad \nu^*=2-\nu,\quad \mu^*=2-\mu,$$
$$0\leqslant\nu'\leqslant\nu^*,\quad 0\leqslant\mu'\leqslant\mu^*,\quad \nu_1+\nu_2=\nu',\quad 0\leqslant\nu_1\leqslant\nu_2.$$

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. □

Remark 6.2. The reader may wonder why we use the "switching trick" at the beginning of Subsection 6.4.2 and why the exponents in Subsection 6.4.1 and 6.4.2 are different. Young's technique in the version of Proposition 4.1 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 $MN\leqslant q^2$ becomes $M^*N^*\geqslant q^2$. 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 6.4.1 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 6.4.2 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. Milićević, *The second moment of twisted modular L-functions*, Geom. Funct. Anal. **25** (2015), 453-516.
[4] S. Das and R. Khan, *Simultaneous nonvanishing of Dirichlet L-functions and twists of Hecke-Maass L-functions*, J. Ramanujan Math. Soc. **30** (2015), no. no. 3, 237–250.
[5] P. Deligne, *La conjecture de Weil, I*, Inst. Hautes Études Sci. Publ. Math. **43** (1974), 273–307.
[6] ———, *La conjecture de Weil, II*, Inst. Hautes Études Sci. Publ. Math. **52** (1980), 137–252.
[7] J.-M. Deshouillers and H. Iwaniec, *Kloosterman sums and Fourier coefficients of cusp forms*, Invent. math. **70** (1982/83), no. 2, 219–288.
[8] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. Tricomi, *Higher transcendental functions II*, McGraw-Hill, 1953.
[9] É. Fouvry, *Sur le problème des diviseurs de Titchmarsh*, J. reine angew. Math. **357** (1985), 51–76.
[10] É. Fouvry, S. Ganguly, E. Kowalski, and Ph. Michel, *Gaussian distribution for the divisor function and Hecke eigenvalues in arithmetic progressions*, Comment. Math. Helv. **89** (2014), no. 4, 979–1014.
[11] É. Fouvry, E. Kowalski, and Ph. Michel, *Algebraic twists of modular forms and Hecke orbits*, Geom. Funct. Anal. **25** (2015), no. 2, 580–657.
[12] ———, *Algebraic trace functions over the primes*, Duke Math. J. **163** (2014), no. 9, 1683–1736.

[13] ———, *A study in sums of products*, Philos. Trans. A **373** (2015), no. 2040, 20140309, 26pp.
[14] É. Fouvry and Ph. Michel, *Sur certaines sommes d'exponentielles sur les nombres premiers*, Ann. Sci. École Norm. Sup. (4) **31** (1998), no. 1, 93–130.
[15] 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.
[16] I.S. Gradshteyn and I.M. Ryzhik, *Table of integrals, series, and products*, Seventh ed., Elsevier/Academic Press, Amsterdam, 2007.
[17] C. Hooley, *On exponential sums and certain of their applications*, Number theory days (Exeter, 1980), London Math. Soc. Lecture Note Series, vol. 56, Cambridge Univ. Press, 1982, pp. 92–122.
[18] H. Iwaniec, *Spectral methods of automorphic forms. Second edition*, Graduate Studies in Mathematics, vol. 53, American Mathematical Society; Revista Matemática Iberoamericana, Madrid, Providence, RI, 2002.
[19] H. Iwaniec and E. Kowalski, *Analytic number theory*, Colloquium Publications, vol. 53, American Mathematical Society, Providence, RI, 2004.
[20] N.M. Katz, *Sommes exponentielles*, Astérisque, vol. 79, Société Mathématique de France, Paris, 1980.
[21] ———, *Gauss sums, Kloosterman sums, and monodromy groups*, Annals of Mathematics Studies, vol. 116, Princeton University Press, Princeton, NJ, 1988.
[22] ———, *Exponential sums and differential equations*, Annals of Mathematics Studies, vol. 124, Princeton University Press, Princeton, NJ, 1990.
[23] H.H. Kim, *Functoriality for the exterior square of* GL_4 *and the symmetric fourth of* GL_2*. With appendix 1 by Dinakar Ramakrishnan and appendix 2 by Kim and Peter Sarnak.*, J. Amer. Math. Soc. **16** (2003), no. 1, 139–183.
[24] E. Kowalski, Ph. Michel, and W. Sawin, *Bilinear forms with Kloosterman sums and applications*. `arXiv:1511.01636`.
[25] Ph. Michel, *Analytic number theory and families of automorphic L-functions in Automorphic forms and Applications*, Vol. 12, IAS/Park City Math. Ser., Amer. Math. Soc., Providence, RI, 2007.
[26] Y. Motohashi, *Spectral theory of the Riemann zeta-function*, Cambridge Tracts in Mathematics, vol. 127, Cambridge University Press, Cambridge, 1997.
[27] F.W.J. Olver, *The asymptotic expansion of Bessel functions of large order*, Phil. Trans. R. Soc. Lond. A **247** (1954), 328-368.
[28] T. Stefanicki, *Non-vanishing of L-functions attached to automorphic representations of* GL(2) *over* **Q**, J. reine angew. Math. **474** (1996), 1–24.
[29] M.P. Young, *The fourth moment of Dirichlet L-functions*, Ann. of Math. (2) **173** (2011), no. 1, 1–50.

7. Appendix: Mathematica code

Section 7.4.1.

```
In[1] :=    Maximize[{Min[(m + n)/2 - 1, Max[(n - m - 1)/2, (n - m - 1)/4],
            (1 + m - n)/2, Max[(m + 1 - n)/2, (2 n1 + m - 2 - n)/2],
            Max[1/4 - n1/2, (m - n2)/2, (2 n1 + m - n - 2)/2], (m + 2 n1 - n)/ 2,
            Max[Min[2 n1 - (m + n)/2, n/6 + n1 + 1/2 - n2 - 2 m/3], (m - n)/2,
            (m + 2 n1 - 2 - n)/2, 3 m/2 - n2 - 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 >= 0, n >= m, m + n <= 2, n1 + n2 == n, n1 <= n2, n1 >= 0,
            n1prime >= 0, n1prime <= n1circ, n1circ == 1 - n1, n2prime >= 0,
            n2prime <= n2circ, n2circ == 1 - n2}, {m, n, n1, n2, n1prime,
            n2prime, n1circ, n2circ}]
```

Out[1] := $\left\{-\frac{1}{68}, \left\{\mathtt{m} \to \frac{161}{306}, \mathtt{n} \to \frac{449}{306}, \mathtt{n1} \to \frac{9}{17}, \mathtt{n2} \to \frac{287}{306}, \mathtt{n1prime} \to \frac{8}{17}, \mathtt{n1prime} \to \frac{19}{306}, \mathtt{n1circ} \to \frac{8}{17}, \mathtt{n1circ} \to \frac{19}{306}\right\}\right\}$

Section 7.4.2.

```
In[2] :=    Maximize[{Min[(m + n)/2 - 1 + 7m/64, Max[(m - n - 1)/2, (m - n - 1)/4],
            (1 + n - m)/2, 7/64(mstar-mprime) + Max[(2 mprime - mstar - nstar + 1)/ 2,
            (2 n1 + 2 mprime - 2 - mstar - nstar)/2],
            7/64(mstar-mprime) + Max[(mprime + n2 + 1/2 - nstar - mstar)/ 2,
            (2 mprime + n1 - mstar - nstar)/ 2, (2 n1 + 2 mprime - 2 - mstar - nstar)/2],
            7/64(mstar-mprime) + mprime + n1 - (mstar + nstar)/2,
            7/64(mstar-mprime) + Max[Min[2 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 <= n, n <= m, m + n <= 2,
            nstar == 2 - n, mstar == 2 - m, 0 <= nprime, nprime <= nstar,
            0 <= mprime, mprime <= mstar, n1 + n2 == nprime, 0 <= n1, n1 <= n2},
            {m, n, n1, n2, nprime, nstar, mprime, mstar}]
```

Out[2] := $\left\{-\frac{1}{64}, \left\{\mathtt{m} \to \frac{47}{32}, \mathtt{n} \to \frac{17}{32}, \mathtt{n1} \to \frac{17}{32}, \mathtt{n2} \to \frac{15}{32}, \mathtt{nprime} \to \frac{47}{32}, \mathtt{nstar} \to \frac{47}{32}, \mathtt{mprime} \to \frac{17}{32}, \mathtt{mstar} \to \frac{17}{32}\right\}\right\}$

Mathematisches Institut, Universität Göttingen, Bunsenstr. 3-5, 37073 Göttingen, Germany
E-mail address: vblomer@math.uni-goettingen.de

Laboratoire de Mathématiques d'Orsay, Université Paris–Saclay, 91405 Orsay, France
E-mail address: etienne.fouvry@math.u-psud.fr

ETH Zürich – D-MATH, Rämistrasse 101, CH-8092 Zürich, Switzerland
E-mail address: kowalski@math.ethz.ch

EPF Lausanne, Chaire TAN, Station 8, CH-1015 Lausanne, Switzerland
E-mail address: philippe.michel@epfl.ch

Department of Mathematics, Bryn Mawr College, Park Science Building, 101 North Merion Avenue, Bryn Mawr, PA 19010-2899, U.S.A.
E-mail address: dmilicevic@brynmawr.edu