Mathematika 64 (2018) 701–729 © 2018 University College London
doi:10.1112/S0025579318000220

TERNARY DIVISOR FUNCTIONS IN ARITHMETIC PROGRESSIONS TO SMOOTH MODULI

PING XI

Abstract. We prove that the exponent of distribution of τ_3 in arithmetic progressions can be as large as $\frac{1}{2}+\frac{1}{34}$, provided that the moduli is squarefree and has only sufficiently small prime factors. The tools involve arithmetic exponent pairs for algebraic trace functions, as well as a double q-analogue of the van der Corput method for smooth bilinear forms.

§1. *Introduction.* Denote by $\tau_k(n)$ the number of positive integral solutions to the equation $x_1x_2\cdots x_k = n$ (and write $\tau_2 = \tau$). A basic problem in analytic number theory is to study the asymptotic behavior of the average

$$\mathscr{D}_k(X;q,a) := \sum_{\substack{n\leqslant X\\ n\equiv a\ (\mathrm{mod}\, q)}} \tau_k(n) \tag{1.1}$$

for $q \geqslant 1$ and $(a,q)=1$.

Define

$$\mathscr{M}_k(X;q,a) := \frac{1}{\varphi(q)} \sum_{\substack{n\leqslant X\\ (n,q)=1}} \tau_k(n),$$

which is expected to be the main term in the asymptotic formula for $\mathscr{D}_k(X;q,a)$. Note that $\mathscr{M}_k(X;q,a) \sim c_{k,q}X(\log X)^{k-1}$ with some positive constant $c_{k,q} \gg (\log\log q)^{-k}$. For each fixed $k \geqslant 1$, we are interested in the admissible value of θ_k being as large as possible such that for any $\varepsilon > 0$, the bound

$$\mathscr{D}_k(X;q,a) - \mathscr{M}_k(X;q,a) \ll \frac{X}{q}(\log X)^{-A} \tag{1.2}$$

holds for all $q < X^{\theta_k-\varepsilon}$ and $A > 0$. It is widely believed that one can take $\theta_k = 1$ for each $k \geqslant 1$; this is, however, only known for the trivial case $k = 1$. Relying on Weil's bound for Kloosterman sums, Selberg [**19**] and Hooley [**12**] obtained independently that $\theta_2 = \frac{2}{3}$ is admissible. For $k \geqslant 3$, the determination of θ_k becomes difficult. Anyway, it was already known to Linnik [**16**] that $\theta_3 = \theta_4 = \frac{1}{2}$ is acceptable. It would be much more difficult to go beyond $\frac{1}{2}$ in general. However, Friedlander and Iwaniec [**9**] were able to prove that one can take

Received 27 October 2017, published online 19 June 2018.
MSC (2010): 11N37, 11N25, 11B25, 11T23, 11L05 (primary).

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

$\theta_3 = \frac{1}{2} + \frac{1}{230}$, and this was later increased to $\frac{1}{2} + \frac{1}{82}$ by Heath-Brown [11]. In the case of prime moduli, Fouvry, Kowalski and Michel [6] can prove that $\theta_3 = \frac{1}{2} + \frac{1}{46}$ is admissible appealing to their bilinear forms of Kloosterman sums with smooth coefficients (see [5, Theorem 1.16]).

The problem becomes much easier if conditions are imposed on the prime factorization of q. For instance, Irving [13] can beat the record of Selberg–Hooley if q has no prime factors exceeding $q^{1/18.1}$. In particular, if $\eta > 0$ is sufficiently small and q has no prime factors exceeding q^η, then the admissible value of θ_2 can be enlarged to $\frac{2}{3} + \frac{1}{246} + O(\eta)$. As a consequence of arithmetic exponent pairs developed in [22], we may obtain a slightly larger value $\frac{2}{3} + \frac{1}{232}$.

In this paper, we focus on the estimation of θ_3 under the assumptions that q is squarefree and has only small prime factors.

THEOREM 1.1. *Let*

$$\theta < \frac{9}{17} = \frac{1}{2} + \frac{1}{34}.$$

There exists some constant $\eta > 0$ such that if $q \asymp X^\theta$ is squarefree and has all its prime factors not exceeding q^η, then (1.2) with $k = 3$ holds uniformly for $(a, q) = 1$ and the implied constant therein may depend on θ and A.

As in [22], we benefit a lot from the q-van der Corput method and arithmetic exponent pairs. In addition, we will develop in Lemma 2.4 a kind of *double* q-van der Corput method to estimate smooth bilinear forms of exponential sums; in particular, Lemma 2.4 is non-trivial if $M > q^\varepsilon$, $N > q^\varepsilon$, and $MN > q^{2/3+\varepsilon}$ (see Remark 2 for detailed arguments). On the other hand, Fouvry, Kowalski and Michel [5, Theorem 1.16] have to assume $MN > q^{3/4+\varepsilon}$ in the case of prime moduli q.

In fact, only two exponent pairs $(\kappa, \lambda) = (\frac{1}{2}, \frac{1}{2}), (\frac{1}{6}, \frac{2}{3})$ will be essentially used, the former of which comes from the completion method and does not require conditions on the factorization of q. In fact, we can restrict q to be a product of two distinct primes of certain sizes; the aim of this paper, however, is to see how far one can go if assuming the factorization of q being as good as possible such that the arithmetic exponent pairs apply freely.

Notation. As usual, we write $\mathrm{e}(x) = \mathrm{e}^{2\pi \mathrm{i} x}$; μ and φ denote the Möbius and Euler functions, respectively. The multiplicative inverse $\overline{x}$ of $x \pmod q$ is defined by $\overline{x}x \equiv 1 \pmod q$. For a coefficient $\boldsymbol{\alpha} = (\alpha_m)$, denote by $\|\boldsymbol{\alpha}\|$ the ℓ_2-norm $(\sum_m |\alpha_m|^2)^{1/2}$. For $g \in L^1(\mathbf{R})$, define the Fourier transform

$$\widehat{g}(\lambda) = \int_{\mathbf{R}} g(x)\mathrm{e}(-\lambda x)\, dx.$$

We use ε to denote a very small positive number, which might be different at each occurrence; we also write $X^\varepsilon \log X \ll X^\varepsilon$. The notation $n \sim N$ means $N < n \leqslant 2N$. We write $A \ll B$ or $A = O(B)$ to mean $|A| \leqslant cB$ with some positive constant c, which is independent from the involved variables. Also, $A \asymp B$ is defined to be $B \ll A \ll B$.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

§2. *Averages of trace functions.* This section is devoted to several estimates on certain averages of trace functions over finite fields. Since the main concern of this paper is to study the distribution of τ_3, we will not spend too much time on the background of trace functions, which we will leave in a black box. Some necessary terminology on trace functions will be mentioned and complete details can be found, for instance, in [4, 7, 14, 22], etc.

2.1. *Kloosterman sums as trace functions.* For an odd prime p, we may consider the middle-extension sheaves on $\mathbf{A}^1_{\mathbf{F}_p}$. Recall that a sheaf $\mathcal{F}$ on $\mathbf{A}^1_{\mathbf{F}_p}$ is said to be middle-extension, if $\mathcal{F} \simeq j_*(j^*\mathcal{F})$ holds for each $j : U \to \mathbf{A}^1$ as the open immersion of a non-empty open set on which $\mathcal{F}$ is lisse. Denote by $\mathfrak{c}(\mathcal{F})$ the conductor of $\mathcal{F}$. The trace function K associated to $\mathcal{F}$ is given by $K(x) = \iota((\mathrm{tr}\mathcal{F})(\mathbf{F}_p, x))$ for some isomorphism $\iota : \overline{\mathbf{Q}}_\ell \to \mathbf{C}$ with $\ell \neq p$. A trace function K can be associated to different sheaves, and the minimum of the conductors of these sheaves is defined to be the conductor of K, denoted by $\mathfrak{c}(K)$.

In terms of trace functions $(\mathrm{mod}\, p)$, we may construct the *composite* trace functions following the manners in [18] and [22]. Let $q \geqslant 3$ be a squarefree number and $K = K_q(\cdot)$ a *composite* trace function (formally) defined by the product

$$K_q(n) = \prod_{p|q} K_p(n), \tag{2.1}$$

where K_p is the trace function of certain ℓ-adic middle-extension sheaf $\mathcal{F}_p$ on $\mathbf{A}^1_{\mathbf{F}_p}$ for each $p \mid q$. In addition, we may define the conductor of K by

$$\mathfrak{c}(K) = \prod_{p|q} \mathfrak{c}(K_p).$$

Note that $\mathfrak{c}(K) \ll q^\varepsilon$ for any $\varepsilon > 0$ if $\mathfrak{c}(K_p)$ is bounded for each $p \mid q$. Originally, K_p is a function living in $\mathbf{Z}/p\mathbf{Z}$; one can extend its domain to $\mathbf{Z}$ by periodicity, and one can do the same to $K = K_q$.

Some of the subsequent arguments are applicable to general trace functions, however we only focus on Kloosterman sums, which are the main tools in this paper and usually serve as typical examples of trace functions.

There are many ways to define Kloosterman sums. For instance, for a positive integer $\nu \geqslant 1$, we may define the exponential sum

$$S_\nu(\mathbf{n}; q) = q^{(1-\nu)/2} \sum_{\substack{a_1,\dots,a_\nu \,(\mathrm{mod}\, q) \\ a_1\cdots a_\nu \equiv 1 \,(\mathrm{mod}\, q)}} \cdots \sum \mathrm{e}\left(\frac{n_1a_1 + \cdots + n_\nu a_\nu}{q}\right)$$

for each positive integer q and vector $\mathbf{n} = (n_1, \dots, n_\nu) \in \mathbf{Z}^\nu$. One may get slightly more general sums if the congruence condition is replaced by $a_1 \cdots a_\nu \equiv a \,(\mathrm{mod}\, q)$ for some $a \,(\mathrm{mod}\, q)$, which can be expressed in terms of $S_\nu(\mathbf{n}; q)$ by virtue of Möbius inversion.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

For $\nu = 2$, $S_2(\mathbf{n}; q)$ is the classical Kloosterman sum

$$S_2(m, n; q) = \frac{S(m, n; q)}{\sqrt{q}} = \frac{1}{\sqrt{q}} \sideset{}{^*}\sum_{a \,(\mathrm{mod}\, q)} \mathrm{e}\Big(\frac{ma + n\overline{a}}{q}\Big).$$

The Weil bound for $S_2(\mathbf{n}; q)$ reads

$$|S_2(m, n; q)| \leqslant \tau(q)(m, n, q)^{1/2}. \tag{2.2}$$

For $\nu \geqslant 3$, one may view $S_\nu(\mathbf{n}; q)$ as the hyper-Kloosterman sum introduced by Mordell (slightly different from that of Deligne in notation). What will be used in this paper is the case $\nu = 3$, for which Smith [20] deduced from Deligne's bound that

$$|S_3(m, n, l; q)| \leqslant \tau_3(q)(m, n, l, q). \tag{2.3}$$

We should remark that $S_\nu(\mathbf{n}; q)$ appears naturally in the study of τ_ν in arithmetic progressions (as well as analytic theory of GL_ν automorphic forms). However, there are more algebraic interpretations with the slightly more general sum

$$\mathrm{Kl}_\nu(m, q) = q^{(1-\nu)/2} \sideset{}{^*}\sum_{\substack{x_1, \dots, x_\nu \,(\mathrm{mod}\, q) \\ x_1 \cdots x_\nu \equiv m \,(\mathrm{mod}\, q)}} \cdots \sideset{}{^*}\sum \mathrm{e}\Big(\frac{x_1 + \cdots + x_\nu}{q}\Big).$$

Note that $S_\nu(\mathbf{n}; q) = \mathrm{Kl}_\nu(n_1 n_2 \cdots n_\nu, q)$ if $(n_1 n_2 \cdots n_\nu, q) = 1$. According to Deligne, there exists an ℓ-adic middle-extension sheaf $\mathcal{K}\ell_\nu$ modulo p, called a Kloosterman sheaf with the corresponding trace function

$$K_{\mathcal{K}\ell_\nu}(x) = \mathrm{Kl}_\nu(x, p), \quad x \in \mathbf{F}_p^\times.$$

Such a sheaf was constructed by Deligne [2]. According to him, $\mathcal{K}\ell_\nu$ is geometrically irreducible, of rank ν, and with the conductor bounded by $\nu + 3$. Therefore, one needs to transform $S_\nu(\mathbf{n}; q)$ to the shape $\mathrm{Kl}_\nu(n, q)$, where n can be defined explicitly by certain products of each coordinate of $\mathbf{n}$. In this way, the average of $S_\nu(\mathbf{n}; q)$ over $\mathbf{n}$ can be interpreted by multilinear sums of $\mathrm{Kl}_\nu(\cdot, q)$, taking shapes of Lemmas 2.3 and 2.4 for instance.

For $\nu = 2$, one may appeal to the classical Kuznetsov–Selberg identity, which reads

$$S_2(m, n; q) = \sum_{d | (m, n, q)} \sqrt{d} S_2(mnd^{-2}, 1; qd^{-1}). \tag{2.4}$$

Such a formula was originally stated by Selberg and the first rigorous proof was given by Kuznetsov [15] appealing to his trace formula. The study of τ_3 requires a similar transformation formula for $\nu = 3$. Smith [21] tried to make an extension of (2.4); as pointed out by Heath-Brown [11], however, the extension is not successful, and there is an alternative approach by virtue of smooth (or friable) numbers. In Appendix A, we will give an elementary proof of (2.4), and the arguments can be easily extended to yield the following formula for $S_3(\mathbf{n}; q)$.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

LEMMA 2.1. *Let q be a squarefree number. For all $m, n, l \in \mathbf{Z}$, we have*

$$S_3(m,n,l;q) = \sum_{d|(m,n,l,q)} \sum_{r|(n/d,l/d,q/d)} \sum_{s|(nl/(dr)^2,m/d,q/dr)} \times d\mu(rs) S_3\left(\frac{\overline{rs}mnl}{d^3r^2s^2}, 1, 1; \frac{q}{drs}\right), \tag{2.5}$$

where $\overline{rs}$ is defined to be the multiplicative inverse of $rs \pmod{q/(drs)}$.

The above formula seems complicated at first sight, but it is much simpler in applications since d, r, s contribute as divisor sums. By virtue of Lemma 2.1, one may transform averages of $S_3(m,n,l;q)$ over m, n, l to bilinear or trilinear forms of $S_3(\gamma mnl, 1, 1; q')$ with some $\gamma \in \mathbf{Z}$ and $q'|q$. Note that $S_3(x, 1, 1; p) = \mathrm{Kl}_3(x, p)$ for all $x \in \mathbf{Z}$, and thus, Lemmas 2.2–2.4 as given below are applicable if the trace function is taken as $x \mapsto \mathrm{Kl}_3(x, q)$.

2.2. *Arithmetic exponent pairs for trace functions.* Let q be a squarefree number and $K = K_q$ defined by (2.1). For a given interval I, we consider the following average of K over I:

$$\mathfrak{S}(K, W) = \sum_{n \in I} K_q(n) W_\delta(n),$$

where W_δ is an arbitrary function defined over $\mathbf{Z}/\delta\mathbf{Z}$ satisfying $\|W_\delta\|_\infty \leqslant 1$. Roughly speaking, the Pólya–Vinogradov bound is non-trivial for $|I| > q^{1/2+\varepsilon}$ (at least for $\delta = 1$). The situation is much easier if conditions are imposed on the prime factorization of q. In [**22**], we developed the method of arithmetic exponent pairs for averages of composite trace functions that may go far beyond the Pólya–Vinogradov bound, as long as the factorization of q is good enough. Such an observation can be at least dated back to Heath-Brown and his proof of Weyl-type subconvex bounds for Dirichlet L-functions to well-factorable moduli.

To summarize the arithmetic exponent pairs developed in [**22**], we define the *normalized* Fourier transform

$$\widehat{W}_\delta(y) = \frac{1}{\sqrt{\delta}} \sum_{x \,(\mathrm{mod}\, \delta)} W_\delta(x) \mathrm{e}\left(\frac{-yx}{\delta}\right).$$

Recall that a middle-extension sheaf $\mathcal{F}_p$ on $\mathbf{A}^1_{\mathbf{F}_p}$ is said to be d-amiable if it is pointwise pure of weight 0 and no geometrically irreducible component of $\mathcal{F}_p$ is geometrically isomorphic to an Artin–Schreier sheaf of the form $\mathcal{L}_{\psi(P)}$, where $P \in \mathbf{F}_p[X]$ is a polynomial of degree $\leqslant d$. In such a case, we also say the associated trace function K_p is d-amiable. The composite trace function K_q is called d-amiable if K_p is d-amiable for each $p \mid q$. In addition, a sheaf (or its associated trace function) is said to be ∞-amiable if it is amiable for any fixed $d \geqslant 1$.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

We are ready to recall the main results in [22, §§3–6]. Let $\eta > 0$ be a sufficiently small number. Suppose q is a squarefree number with no prime factors exceeding q^η, and $K = K_q(\cdot)$ is an ∞-amiable trace function $(\mathrm{mod}\, q)$. For $|I| < q\delta$, one expects the estimate

$$\mathfrak{S}(K, W) \ll q^\varepsilon \left(\frac{q}{|I|}\right)^\kappa |I|^\lambda \delta^\nu \|\widehat{W}_\delta\|_\infty^\mu \tag{2.6}$$

holds for some tuple $(\kappa, \lambda, \nu, \mu)$, where $\varepsilon > 0$ and the implied constant depends only on ε and the conductor $\mathfrak{c}(K)$ of K. If this is the case, we call $(\kappa, \lambda, \nu, \mu)$ an *exponent pair* for K. In particular, one may take $(\kappa, \lambda, \nu, \mu) = (0, 1, 0, 0)$ trivially.

As shown in [22], other exponent pairs can be produced starting from the trivial one $(\kappa, \lambda, \nu, \mu) = (0, 1, 0, 0)$ by virtue of A- and B-processes in the q-analogue of the van der Corput method. Since we do not input extra conditions on W_δ, it is not a bad choice to utilize the trivial bound $\|\widehat{W}_\delta\|_\infty \leqslant \sqrt{\delta}$. The values of ν, μ are usually not too large and they do not play essential roles in many applications. For instance, we will require the constraints $0 \leqslant \nu + \mu/2 - \kappa \leqslant 1$ in the proof of Lemma 2.3 below, which hold obviously upon many choices of exponent pairs.

LEMMA 2.2. *Keep the notation and convention as above. The estimate (2.6) holds for every $(\kappa, \lambda, \nu, \mu)$ produced by A- and B-processes applied to $(0, 1, 0, 0)$ with*

$$0 \leqslant \kappa \leqslant \tfrac{1}{2} \leqslant \lambda \leqslant 1. \tag{2.7}$$

In particular, one may take

$$(\kappa, \lambda, \nu, \mu) = (\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, 1), \qquad (\tfrac{1}{6}, \tfrac{2}{3}, \tfrac{1}{2}, 0) \quad \textit{and} \quad (\tfrac{2}{7}, \tfrac{4}{7}, \tfrac{2}{7}, 1).$$

2.3. *General bilinear forms of trace functions.* For any coefficients $\boldsymbol{\alpha} = (\alpha_m)$, $\boldsymbol{\beta} = (\beta_n)$ supported in some intervals of length M, N, respectively, we introduce the bilinear form

$$\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K) = \sum_m \sum_n \alpha_m \beta_n K_q(mn).$$

LEMMA 2.3. *With the same notation and assumptions as in Lemma 2.2, we have*

$$\begin{aligned}\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K) \ll (qMN)^\varepsilon (MN)^{1/2} \|\boldsymbol{\alpha}\| \|\boldsymbol{\beta}\| \\ \times (\varepsilon_1 q^{\kappa/2} M^{(\lambda-\kappa-1)/2} + q^{-1/4} + N^{-1/2})\end{aligned}$$

for $(\kappa, \lambda, \nu, \mu)$ with $0 \leqslant \nu + \mu/2 - \kappa \leqslant 1$, where $\varepsilon_1 = 1$ if $M < q$ and vanishes otherwise.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

Remark 1. As one will see later, the proof of Theorem 1.1 only utilizes the choice $(\kappa, \lambda) = (\frac{1}{2}, \frac{1}{2})$, in which case q is not required to have only small prime factors. Lemma 2.3 presents an estimate in which the exponent pair is displayed explicitly, and this might be useful on some other occasions.

Proof. By Cauchy's inequality, we have

$$|\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K)|^2 \leqslant \|\boldsymbol{\alpha}\|^2 \sum_m \Big| \sum_n \beta_n K_q(mn) \Big|^2.$$

Squaring out and switching summations, we get

$$|\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K)|^2 \leqslant \|\boldsymbol{\alpha}\|^2 \sum_{n_1} \sum_{n_2} |\beta_{n_1}||\beta_{n_2}||\Sigma(n_1, n_2; q)|, \tag{2.8}$$

where

$$\Sigma(n_1, n_2; q) = \sum_m K_q(mn_1)\overline{K_q(mn_2)}.$$

For $n_1 = n_2$, we utilize the trivial bound $\Sigma(n_1, n_2; q) \ll Mq^{\varepsilon}$, giving a contribution to the right-hand side of (2.8) at most $O(q^{\varepsilon} M \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2)$. We now assume $n_1 \neq n_2$ and put $d = (n_1 - n_2, q)$, which is also a squarefree number and $(d, q/d) = 1$. From the Chinese remainder theorem, it follows that

$$K_q(mn_1)\overline{K_q(mn_2)} = |K_d(mn_1)|^2 K_{q/d}(mn_1)\overline{K_{q/d}(mn_2)}.$$

One may check that $m \mapsto K_{q/d}(mn_1)\overline{K_{q/d}(mn_2)}$ is an ∞-amiable trace function $(\mathrm{mod}\, q/d)$ as long as q/d is large enough. From Lemma 2.2 it follows that

$$\Sigma(n_1, n_2; q) \ll (qM)^{\varepsilon} \left(\frac{q}{dM}\right)^{\kappa} M^{\lambda} d^{\nu+\mu/2}$$

if $q > M$; otherwise we have

$$\Sigma(n_1, n_2; q) \ll (qM)^{\varepsilon} \left[\frac{M}{q}\right] (q/d)^{1/2} d \ll (qM)^{\varepsilon} M (d/q)^{1/2}.$$

Hence, we conclude that

$$|\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K)|^2 \ll q^{\varepsilon} M \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2 + \varepsilon_1 \mathscr{B}_1 + (1 - \varepsilon_1)\mathscr{B}_2,$$

where

$$\mathscr{B}_1 = (qM)^{\varepsilon} q^{\kappa} M^{\lambda-\kappa} \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2 \sum_{d|q} d^{\nu+\mu/2-\kappa} \sum_{\substack{n_1 \neq n_2 \\ n_1 \equiv n_2 \,(\mathrm{mod}\, d)}} \sum |\beta_{n_1}||\beta_{n_2}|,$$

$$\mathscr{B}_2 = (qM)^{\varepsilon} q^{-1/2} M \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2 \sum_{d|q} d^{1/2} \sum_{\substack{n_1 \neq n_2 \\ n_1 \equiv n_2 \,(\mathrm{mod}\, d)}} \sum |\beta_{n_1}||\beta_{n_2}|.$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

Note that $0 \leqslant \nu + \mu/2 - \kappa \leqslant 1$. We then have

$$\begin{aligned}\mathscr{B}_1 &\ll (qM)^{\varepsilon} q^{\kappa} M^{\lambda-\kappa} \|\boldsymbol{\alpha}\|^2 \sum_{n_1 \neq n_2} \sum |\beta_{n_1}||\beta_{n_2}|(n_1 - n_2, q)^{\nu+\mu/2-\kappa} \\ &\ll (qM)^{\varepsilon} q^{\kappa} M^{\lambda-\kappa} \|\boldsymbol{\alpha}\|^2 \sum_{n_2} |\beta_{n_2}|^2 \sum_{n \ll N} (n, q)^{\nu+\mu/2-\kappa} \\ &\ll (qMN)^{\varepsilon} q^{\kappa} M^{\lambda-\kappa} N \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\mathscr{B}_2 &\ll (qM)^{\varepsilon} q^{-1/2} M \|\boldsymbol{\alpha}\|^2 \sum_{n_1 \neq n_2} \sum |\beta_{n_1}||\beta_{n_2}|(n_1 - n_2, q)^{1/2} \\ &\ll (qMN)^{\varepsilon} q^{-1/2} MN \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2.\end{aligned}$$

This establishes Lemma 2.3 by collecting the above estimates of $\mathscr{B}_1$ and $\mathscr{B}_2$. □

2.4. *Double q-van der Corput method.* The classical van der Corput method can also be applied to the two-dimensional exponential sums of the shape

$$\sum_{(m,n)\in\Omega} \sum e(f(m, n)),$$

where Ω is a rectangle, or some set bounded by simple algebraic curves. As an analogue of q-van der Corput, one can also develop a double q-van der Corput method to study the double sums of trace functions

$$\sum_{(m,n)\in\Omega} \sum \mathscr{K}(m, n),$$

where $\mathscr{K}$ is the trace function of an ℓ-adic middle-extension sheaf on $\mathbf{A}^1_{\mathbf{F}_p} \times \mathbf{A}^1_{\mathbf{F}_p}$. In applications to Theorem 1.1, we need the simpler case $\mathscr{K}(m, n) = K(mn)$ with K being the amiable trace function given as above. This leads to the study of the bilinear form $\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K)$, and we will prove the following alternative estimate for $\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K)$ with smooth coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, provided that K is the trace function of certain Kloosterman sheaves.

LEMMA 2.4. *Suppose $U, V \in \mathscr{C}^{\infty}(\mathbf{R}^+)$ are two bounded smooth functions with compact support and let*

$$\alpha_m = U\left(\frac{m}{M}\right), \qquad \beta_n = V\left(\frac{n}{N}\right).$$

Let a be a fixed integer and $q \geqslant 1$ a squarefree number which has a factor $q_1 \in (q^{1/3-\varepsilon}, q^{1/3+\varepsilon})$. For $K = \mathrm{Kl}_3(a\cdot, q)$, we have

$$\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K) \ll (qMN)^{\varepsilon} MN\{q^{-1/3} + M^{-1/2} + N^{-1/2} + q^{1/3}(MN)^{-1/2}\}.$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

Remark 2. If $\max\{M, N\} \ll q^{2/3}$, the upper bound in Lemma 2.4 can be reduced to

$$(qMN)^{\varepsilon} MN \cdot (MN)^{-1/2} q^{1/3} = (qMN)^{\varepsilon} \left(\frac{q}{M}\right)^{\kappa} \left(\frac{q}{N}\right)^{\kappa} M^{\lambda} N^{\lambda}$$

with $(\kappa, \lambda) = (\frac{1}{6}, \frac{2}{3})$. This coincides with the two-dimensional van der Corput method in the classical case; see Conjecture Q in [**10**, p. 73] for details.

One can see that Lemma 2.4 is non-trivial for $M > q^{\varepsilon}, N > q^{\varepsilon}$ and $MN > q^{2/3+\varepsilon}$. This can be compared with the estimate of Fouvry, Kowalski and Michel [**5**, Theorem 1.16] in the case of prime moduli p, where they have to assume $MN > p^{3/4+\varepsilon}$.

We believe Lemma 2.4 holds for general trace functions K. The situation is much simplier if K is a non-trivial Dirichlet character mod q, in which case the bilinear form becomes the product of two linear forms, so that arithmetic exponent pairs developed in [**22**] apply.

Proof. The lemma becomes trivial if $\max\{M, N\} < 2q^{1/3}$. We first assume $M < 2q^{1/3} \leqslant N < q$, in which case one may apply the arithmetic exponent pairs to the sum over n with $(\kappa, \lambda) = (\frac{1}{6}, \frac{2}{3})$, getting

$$\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K) \ll (qMN)^{\varepsilon} M \left(\frac{q}{N}\right)^{1/6} N^{2/3} \ll (qMN)^{\varepsilon} q^{1/6} MN^{1/2}.$$

This is at most $(qMN)^{\varepsilon} q^{1/3} (MN)^{1/2}$ since $M < 2q^{1/3}$. It remains to consider the case $\min\{M, N\} \geqslant 2q^{1/3}$, which we henceforth assume.

Suppose $q = rs$. For any $l_1, l_2 \in \mathbf{Z}$, we have

$$\begin{aligned}
&\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K) \\
&\quad = \sum_m \sum_n U\left(\frac{m + l_1 r}{M}\right) V\left(\frac{n + l_2 r}{N}\right) K_{rs}((m + l_1 r)(n + l_2 r)) \\
&\quad = \sum_m \sum_n U\left(\frac{m + l_1 r}{M}\right) V\left(\frac{n + l_2 r}{N}\right) K_r(mn) K_s((m + l_1 r)(n + l_2 r)),
\end{aligned}$$

where K_s, K_r are the corresponding trace functions $(\mathrm{mod}\, s)$ and $(\mathrm{mod}\, r)$, respectively, both of which are also amiable as long as r, s are both large enough.

Summing over $l_1 \leqslant L_1$ and $l_2 \leqslant L_2$ with $L_1 \leqslant 1 + [M/r]$, $L_2 \leqslant 1 + [N/r]$, we have

$$\begin{aligned}
\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K) &\ll \frac{r^{\varepsilon}}{L_1 L_2} \\
&\times \sum_{m \ll M} \sum_{n \ll N} \left| \sum_{l_1 \leqslant L_1} \sum_{l_2 \leqslant L_2} U\left(\frac{m + l_1 r}{M}\right) V\left(\frac{n + l_2 r}{N}\right) K_s((m + l_1 r)(n + l_2 r)) \right|.
\end{aligned}$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

By Fourier inversion, we have

$$U\left(\frac{m+l_1r}{M}\right)=\int_{\mathbf{R}}\widehat{U}(y)\mathrm{e}\left(\frac{(m+l_1r)y}{M}\right)dy,$$
$$V\left(\frac{n+l_2r}{N}\right)=\int_{\mathbf{R}}\widehat{V}(y)\mathrm{e}\left(\frac{(n+l_2r)y}{N}\right)dy,$$

from which we may derive that

$$\mathscr{B}(\boldsymbol{\alpha},\boldsymbol{\beta};K)\ll\frac{r^{\varepsilon}}{L_1L_2}\times\sum_{m\ll M}\sum_{n\ll N}\Bigg|\sum_{l_1\leqslant L_1}\sum_{l_2\leqslant L_2}\mathrm{e}(l_1y_1+l_2y_2)K_s((m+l_1r)(n+l_2r))\Bigg|$$

for some $y_1, y_2\in\mathbf{R}$. By Cauchy's inequality, we find

$$\begin{aligned}\mathscr{B}(\boldsymbol{\alpha},\boldsymbol{\beta};K)^2&\ll\frac{r^{\varepsilon}MN}{(L_1L_2)^2}\times\sum_{m\ll M}\sum_{n\ll N}\Bigg|\sum_{l_1\leqslant L_1}\sum_{l_2\leqslant L_2}\\&\quad\times\mathrm{e}(l_1y_1+l_2y_2)K_s((m+l_1r)(n+l_2r))\Bigg|^2\\&\ll\frac{r^{\varepsilon}MN}{(L_1L_2)^2}\sum_{l_1,l_1'\leqslant L_1}\sum_{l_2,l_2'\leqslant L_2}\Bigg|\sum_{m\ll M}\sum_{n\ll N}\\&\quad\times K_s((m+l_1r)(n+l_2r))\overline{K_s((m+l_1'r)(n+l_2'r))}\Bigg|\\&\ll\frac{MNr^{\varepsilon}}{L_1L_2}\sum_{0\leqslant|l_1|\leqslant L_1}\sum_{0\leqslant|l_2|\leqslant L_2}|\Pi(l_1,l_2)|\end{aligned}$$

with

$$\Pi(l_1,l_2)=\sum_{m\in\mathscr{M}(l_1)}\sum_{n\in\mathscr{N}(l_2)}K_s(mn)\overline{K_s((m+l_1r)(n+l_2r))},$$

where $\mathscr{M}(l_1), \mathscr{N}(l_2)$ are some intervals with $|\mathscr{M}(l_1)|\ll M$ and $|\mathscr{N}(l_2)|\ll N$. Here we have used the estimates

$$|\widehat{U}(y)|,\qquad|\widehat{V}(y)|\ll(1+|y|)^{-A},\quad A\geqslant 0$$

from integration by parts.

We claim, for all $l_1, l_2\in\mathbf{Z}$, that

$$|\Pi(l_1,l_2)|\ll(MNs)^{\varepsilon}\{s^{-1}MN(l_1,l_2,s)+M+N+s\}.\tag{2.9}$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

Hence,

$$\begin{aligned}\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K)^2 &\ll \frac{MNq^{\varepsilon}}{L_1L_2} \sum_{0\leqslant|l_1|\leqslant L_1} \sum_{0\leqslant|l_2|\leqslant L_2} \{s^{-1}MN(l_1, l_2, s) + M + N + s\} \\ &\ll \frac{MNq^{\varepsilon}}{L_1L_2}\{MN + L_1L_2(s^{-1}MN + M + N + s)\}.\end{aligned}$$

By taking $L_1 = 1 + [M/r]$, $L_2 = 1 + [N/r]$, we get

$$\mathscr{B}(\boldsymbol{\alpha}, \boldsymbol{\beta}; K) \ll (MN)^{1/2+\varepsilon}\{r + (s^{-1}MN + M + N + s)^{1/2}\},$$

provided that $\min\{M, N\} \gg r$. The lemma follows by choosing r, s such that $q^{1/3-\varepsilon} \ll r \ll q^{1/3+\varepsilon}$.

It remains to prove (2.9). By the completion method, we have

$$\begin{aligned}&|\Pi(l_1, l_2)| \\ &\ll \frac{1}{s^2} \sum_{|h_1|<s/2} \sum_{|h_2|<s/2} \min\left\{M, \frac{s}{|h_1|}\right\}\min\left\{N, \frac{s}{|h_2|}\right\}|\Sigma(h_1, h_2; l_1, l_2; s)|,\end{aligned}$$

where

$$\Sigma(h_1, h_2; l_1, l_2; s) = \sum_{x,y \,(\mathrm{mod}\, s)} \sum K_s(xy)\overline{K_s((x + l_1)(y + l_2))}\mathrm{e}\left(\frac{h_1x + h_2y}{s}\right).$$

Then (2.9) follows from an upper bound for $\Sigma(h_1, h_2; l_1, l_2; s)$. From the Chinese remainder theorem, it suffices to consider the case of prime moduli, which has been studied in Theorem B.1 in Appendix B. More precisely, put $s_1 = (s, a)$ and $s_2 = s/s_1$, we then have $(as_1, s_2) = 1$ since s is squarefree. From the Chinese remainder theorem, we may decompose

$$\begin{aligned}&\Sigma(h_1, h_2; l_1, l_2; s) \\ &\quad = \Sigma(\overline{s_1}h_1, \overline{s_1}h_2; l_1, l_2; s_2)|K_{s_1}(0)|^2 \sum_{x_1,y_1 \,(\mathrm{mod}\, s_1)} \sum \mathrm{e}\left(\frac{h_1x_1 + h_2y_1}{s_1}\right).\end{aligned}$$

Note that $|K_{s_1}(0)|^2 = 1/s_1^2$. Thus,

$$|\Sigma(h_1, h_2; l_1, l_2; s)| \leqslant |\Sigma(\overline{s_1}h_1, \overline{s_1}h_2; l_1, l_2; s_2)|.$$

By Theorem B.1 we then have

$$|\Sigma(h_1, h_2; l_1, l_2; s)| \ll s_2^{1+\varepsilon}(h_1, h_2, l_1, l_2, s_2) \leqslant s^{1+\varepsilon}(h_1, h_2, l_1, l_2, s),$$

from which we derive that

$$\begin{aligned}&|\Pi(l_1, l_2)| \\ &\ll \frac{1}{s^2} \sum_{|h_1|\leqslant s/2} \sum_{|h_2|\leqslant s/2} \min\left\{M, \frac{s}{|h_1|}\right\}\min\left\{N, \frac{s}{|h_2|}\right\}s^{1+\varepsilon}(h_1, h_2, l_1, l_2, s) \\ &\ll (MNs)^{\varepsilon}\{s^{-1}MN(l_1, l_2, s) + M + N + s\}\end{aligned}$$

as claimed in (2.9), and thus completing the proof of Lemma 2.4. □

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

§3. *Reduction of the problem.* In what follows, we assume $X \geqslant 3$ is large. Recall that

$$\mathscr{D}_3(X; q, a) = \sum_{\substack{n_1 n_2 n_3 \leqslant X \\ n_1 n_2 n_3 \equiv a \,(\mathrm{mod}\, q)}} 1. \tag{3.1}$$

A fundamental tool in the transformation of $\mathscr{D}_3(X; q, a)$ is the Poisson summation formula. Hence, we would like to remove the condition $n_1 n_2 n_3 \leqslant X$, so that the three variables n_1, n_2, n_3 are independent and smooth. To this end, we follow the arguments in [**6**] and appeal to the following lemma as a smooth partition of unity, which works well in our situation (see [**3**, Lemme 2] for instance).

LEMMA 3.1. *For every $\Delta > 1$, there exists a sequence $\{b_{l,\Delta}\}_{l \geqslant 0}$ of smooth functions with support included in $[\Delta^{l-1}, \Delta^{l+1}]$, such that*

$$\sum_{l \geqslant 0} b_{l,\Delta}(x) = 1 \quad \textit{for all } x \geqslant 1,$$

and

$$x^j b_{l,\Delta}^{(j)}(x) \ll (1 - \Delta^{-1})^{-j} \quad \textit{for all } x \geqslant 1 \textit{ and } j, l \geqslant 0. \tag{3.2}$$

In our applications, we take

$$\Delta = 1 + (\log X)^{-B}$$

for some parameter $B \geqslant 1$. From now on, we denote by N_j, $1 \leqslant j \leqslant 3$, some parameters of the form

$$N_j = \Delta^l = (1 + (\log X)^{-B})^l \tag{3.3}$$

for some $l \geqslant 1$. For such a variable $N_j = \Delta^l$, we define

$$V_j = b_{l,\Delta} \tag{3.4}$$

as given in Lemma 3.1. Thus, the derivatives of V_j satisfy

$$x^l V_j^{(l)}(x) \ll (\log X)^{Bl} \tag{3.5}$$

for all $l \geqslant 0$, where the implied constant may depend on l. Moreover, integration by parts gives

$$\widehat{V}_j(\xi) \ll N_j (\log X)^{Bl} (1 + |\xi| N_j)^{-l} \tag{3.6}$$

for any fixed $l \geqslant 0$, and the implied constant also depends on l.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

For $\boldsymbol{N} = (N_1, N_2, N_3)$, we now consider the smoothed sum

$$\mathscr{D}_3(\boldsymbol{N}; q, a) = \sum_{\substack{n_1,n_2,n_3\in\mathbf{Z}\\ n_1n_2n_3\equiv a \,(\mathrm{mod}\, q)}}\sum\sum V_1(n_1)V_2(n_2)V_3(n_3). \tag{3.7}$$

Define

$$\mathscr{D}_3(\boldsymbol{N}; q) = \sum_{\substack{n_1,n_2,n_3\in\mathbf{Z}\\ (n_1n_2n_3,q)=1}}\sum\sum V_1(n_1)V_2(n_2)V_3(n_3).$$

We appeal to the following lemma for the separation of variables; the proof can be found in [6, Lemma 4.2].

LEMMA 3.2. *For any $A > 0$, one can choose $B \geqslant 1$ such that*

$$\mathscr{D}_3(X; q, a) - \mathscr{M}_3(X; q, a) = \sum_{\boldsymbol{N}}\left\{\mathscr{D}_3(\boldsymbol{N}; q, a) - \frac{1}{\varphi(q)}\mathscr{D}_3(\boldsymbol{N}; q)\right\} + O\left(\frac{X}{q}(\log X)^{-A}\right),$$

where $\boldsymbol{N} = (N_1, N_2, N_3)$ runs over triples of $\boldsymbol{N}$ as above such that

$$X(\log X)^{-B} \leqslant N_1N_2N_3 \leqslant X. \tag{3.8}$$

Due to the symmetry of the problem, we assume

$$N_1 \leqslant N_2 \leqslant N_3. \tag{3.9}$$

Since the number of triples $\boldsymbol{N}$ satisfying (3.8) with N_i of the form (3.3) is $O((\log X)^{3B+3})$, Lemma 3.2 shows that Theorem 1.1 will follow if we can show that for any $\theta < \frac{9}{17}$ and $A > 0$, we have

$$\mathscr{D}_3(\boldsymbol{N}; q, a) - \frac{1}{\varphi(q)}\mathscr{D}_3(\boldsymbol{N}; q) \ll \frac{X}{q}(\log X)^{-A} \tag{3.10}$$

uniformly for all triples $\boldsymbol{N} = (N_1, N_2, N_3)$ satisfying (3.8), (3.9), and (3.3) and for all integers a coprime to q satisfying $q \asymp X^{\theta}$.

To evaluate $\mathscr{D}_3(\boldsymbol{N}; q, a)$, we would like to invoke the following Poisson summation formula.

LEMMA 3.3. *Suppose that g and $\widehat{g}$ are both in $L^1(\mathbf{R})$. For any fixed integers d and a, we have*

$$\sum_{\substack{n\in\mathbf{Z}\\ n\equiv a \,(\mathrm{mod}\, d)}} g(n) = \frac{1}{d}\sum_{h\in\mathbf{Z}}\widehat{g}\left(\frac{h}{d}\right)\mathrm{e}\left(\frac{ha}{d}\right).$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

Note that

$$\mathscr{D}_3(\boldsymbol{N};q,a)=\sum_{\substack{a_1,a_2,a_3 \pmod q\\ a_1a_2a_3\equiv a \pmod q}}\prod_{1\leqslant i\leqslant 3}\sum_{n_i\equiv a_i \pmod q} V_i(n_i).$$

The following lemma can be obtained by applying Lemma 3.3 in different situations. The first identity follows from the applications of Poisson summation to each variable, and this is essentially [**6**, Corollary 2.3]. The second one is concluded by applying Poisson twice, and the third is produced by only one Poisson to n_3; both of them are complementary to (3.11).

LEMMA 3.4. *With the notation as above, we have*

$$\mathscr{D}_3(\boldsymbol{N};q,a)=\frac{1}{q^2}\sum_{m_1,m_2,m_3\in\mathbf{Z}}\widehat{V}_1\Big(\frac{m_1}{q}\Big)\widehat{V}_2\Big(\frac{m_2}{q}\Big)\widehat{V}_3\Big(\frac{m_3}{q}\Big)S_3(am_1,m_2,m_3;q) \tag{3.11}$$

$$=\frac{1}{q^{3/2}}\sum_{\substack{n_1,m_2,m_3\in\mathbf{Z}\\(n_1,q)=1}}V_1(n_1)\widehat{V}_2\Big(\frac{m_2}{q}\Big)\widehat{V}_3\Big(\frac{m_3}{q}\Big)S_2(an_1m_2,m_3;q) \tag{3.12}$$

$$=\frac{1}{q}\sum_{\substack{n_1,n_2,m_3\in\mathbf{Z}\\(n_1n_2,q)=1}}V_1(n_1)V_2(n_2)\widehat{V}_3\Big(\frac{m_3}{q}\Big)\mathrm{e}\Big(\frac{a\overline{n_1n_2}m_3}{q}\Big). \tag{3.13}$$

We would like to evaluate $\mathscr{D}_3(\boldsymbol{N};q,a)$ by virtue of (3.11), (3.12), and (3.13). The main term is expected to come from the terms with zero frequency. To do so, we define

$$\mathscr{M}_1(\boldsymbol{N})=\frac{1}{q^2}\sum_{m_1m_2m_3=0}\widehat{V}_1\Big(\frac{m_1}{q}\Big)\widehat{V}_2\Big(\frac{m_2}{q}\Big)\widehat{V}_3\Big(\frac{m_3}{q}\Big)S_3(am_1,m_2,m_3;q),$$

$$\mathscr{M}_2(\boldsymbol{N})=\frac{1}{q^{3/2}}\sum_{\substack{n_1\in\mathbf{Z}\\(n_1,q)=1}}\sum_{m_2m_3=0}V_1(n_1)\widehat{V}_2\Big(\frac{m_2}{q}\Big)\widehat{V}_3\Big(\frac{m_3}{q}\Big)S_2(an_1m_2,m_3;q),$$

$$\mathscr{M}_3(\boldsymbol{N})=\frac{\widehat{V}_3(0)}{q}\sum_{\substack{n_1,n_2\in\mathbf{Z}\\(n_1n_2,q)=1}}V_1(n_1)V_2(n_2),$$

and

$$\mathscr{E}_j(\boldsymbol{N};q,a)=\mathscr{D}_3(\boldsymbol{N};q,a)-\mathscr{M}_j(\boldsymbol{N}),\quad j=1,2,3. \tag{3.14}$$

For each $j\in\{1,2,3\}$, we may evaluate $\mathscr{M}_j$ asymptotically and each of them gives the expected main term $\mathscr{D}_3(\boldsymbol{N};q)/\varphi(q)$. The evaluations can be stated as follows.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

PROPOSITION 3.1. *For each $j \in \{1, 2, 3\}$, we have*

$$\mathscr{M}_j(\boldsymbol{N}) = \frac{1}{\varphi(q)}\mathscr{D}_3(\boldsymbol{N}; q) + O\bigg(\frac{X}{qN_j}\tau(q)(\log X)^{3B}\bigg).$$

Proof. From Poisson summation (Lemma 3.3), it follows that

$$\begin{aligned}\sum_{\substack{n_3\in\mathbf{Z}\\(n_3,q)=1}} V_3(n_3) &= \frac{1}{\sqrt{q}}\sum_{m_3\in\mathbf{Z}}\widehat{V}_3\bigg(\frac{m_3}{q}\bigg)S_2(0, m_3; q)\\ &= \frac{\varphi(q)}{q}\widehat{V}_3(0) + \frac{1}{\sqrt{q}}\sum_{|m_3|\geqslant 1}\widehat{V}_3\bigg(\frac{m_3}{q}\bigg)S_2(0, m_3; q),\end{aligned}$$

where $S_2(0, m; q)$ satisfies the bound $|S_2(0, m; q)| \leqslant (m, q)/\sqrt{q}$. From (3.6), the second summand is bounded by

$$\ll \frac{N_3(\log X)^{2B}}{q}\bigg(\sum_{m_3\leqslant q/N_3}(m_3, q) + \frac{q^2}{N_3^2}\sum_{m_3>q/N_3}\frac{(m_3, q)}{m_3^2}\bigg) \ll \tau(q)(\log X)^{2B}.$$

Hence, we conclude that

$$\widehat{V}_3(0) = \frac{q}{\varphi(q)}\sum_{\substack{n_3\in\mathbf{Z}\\(n_3,q)=1}} V_3(n_3) + O(\tau(q)(\log X)^{3B}),$$

where we have used the elementary fact $\varphi(q) \gg q/\log q$. It now follows that

$$\mathscr{M}_3(\boldsymbol{N}) = \frac{1}{\varphi(q)}\mathscr{D}_3(\boldsymbol{N}; q) + O\bigg(\frac{N_1N_2}{q}\tau(q)(\log X)^{3B}\bigg).$$

On the other hand, from the inclusion–exclusion principle, we may write

$$\begin{aligned}\mathscr{M}_2(\boldsymbol{N}) = {}& \frac{\widehat{V}_2(0)}{q^2}\sum_{\substack{n_1,m_3\in\mathbf{Z}\\(n_1,q)=1}}\sum V_1(n_1)\widehat{V}_3\bigg(\frac{m_3}{q}\bigg)S(0, m_3; q)\\ &+ \frac{\widehat{V}_3(0)}{q^2}\sum_{\substack{n_1,m_2\in\mathbf{Z}\\(n_1,q)=1}}\sum V_1(n_1)\widehat{V}_2\bigg(\frac{m_2}{q}\bigg)S(an_1m_2, 0; q)\\ &- \frac{1}{q^2}\sum_{\substack{n_1\in\mathbf{Z}\\(n_1,q)=1}} V_1(n_1)\widehat{V}_2(0)\widehat{V}_3(0)S(0, 0; q).\end{aligned}$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

By Poisson summation, we derive that

$$\begin{aligned}\mathscr{M}_2(\boldsymbol{N}) &= \frac{\widehat{V}_2(0)}{q}\sum_{\substack{n_1,n_3\in\mathbf{Z}\\(n_1n_3,q)=1}}\sum V_1(n_1)V_3(n_3)+\frac{\widehat{V}_3(0)}{q}\sum_{\substack{n_1,n_2\in\mathbf{Z}\\(n_1n_2,q)=1}}\sum V_1(n_1)V_2(n_2)\\ &\quad-\widehat{V}_2(0)\widehat{V}_3(0)\frac{\varphi(q)}{q^2}\sum_{\substack{n_1\in\mathbf{Z}\\(n_1,q)=1}}V_1(n_1)\\ &=\frac{1}{\varphi(q)}\mathscr{D}_3(\boldsymbol{N};q)+O\Big(\frac{N_1N_3}{q}\tau(q)(\log X)^{3B}\Big).\end{aligned}$$

In a similar way, we can also conclude

$$\mathscr{M}_1(\boldsymbol{N})=\frac{1}{\varphi(q)}\mathscr{D}_3(\boldsymbol{N};q)+O\Big(\frac{N_2N_3}{q}\tau(q)(\log X)^{3B}\Big). \qquad \square$$

The next section will be devoted to the estimates for $\mathscr{E}_j(\boldsymbol{N};q,a)$, $j=1,2,3$.

§4. *Estimates for* $\mathscr{E}_j(\boldsymbol{N};q,a)$. The estimates for $\mathscr{E}_j(\boldsymbol{N};q,a)$ form the essential part of this paper. Roughly speaking, we will apply Lemma 2.3 to $\mathscr{E}_1(\boldsymbol{N};q,a)$ by re-grouping variables and this argument works well if the three variables are of comparable sizes. When there are reasonably long summations, we would apply Lemma 2.4 to $\mathscr{E}_1(\boldsymbol{N};q,a)$. On the other hand, the treatments to $\mathscr{E}_2(\boldsymbol{N};q,a)$ and $\mathscr{E}_3(\boldsymbol{N};q,a)$ are both trivial, and these are equivalent to the known levels of distributions for τ_2 and τ_1.

Put

$$N_j=X^{\mu_j},\quad j=1,2,3,$$

so that

$$\mu_1\leqslant\mu_2\leqslant\mu_3. \tag{4.1}$$

The following estimates for $\mathscr{E}_j(\boldsymbol{N};q,a)$ will work complementarily to conclude the final level in Theorem 1.1.

PROPOSITION 4.1. *With the notation as above, we have*

$$\mathscr{E}_1(\boldsymbol{N};q,a)\ll\frac{X}{q(\log X)^A}$$

if one of the following two conditions holds:

$$\frac{1}{2}\leqslant\theta<\frac{4}{7},\qquad\frac{(2\lambda-\kappa+2)\theta+\kappa-\lambda-1}{\kappa-\lambda+1}<\mu_3<2-3\theta; \tag{4.2a}$$

$$\frac{1}{2}\leqslant\theta<\frac{3}{5},\qquad\mu_2<2-3\theta,\qquad\mu_3>\frac{8}{3}\theta-1. \tag{4.2b}$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

PROPOSITION 4.2. *With the notation as above, we have*

$$\mathscr{E}_2(\boldsymbol{N}; q, a) \ll \frac{X}{q(\log X)^A}$$

if

$$\mu_2 + \mu_3 > \tfrac{3}{2}\theta. \tag{4.3}$$

PROPOSITION 4.3. *With the notation as above, we have*

$$\mathscr{E}_3(\boldsymbol{N}; q, a) \ll \frac{X}{q(\log X)^A}$$

if

$$\mu_3 > \theta. \tag{4.4}$$

We now start the proof of the above three propositions. Recall that

$$\mathscr{E}_1(\boldsymbol{N}; q, a) = \frac{1}{q^2} \sum\sum\sum_{m_1 m_2 m_3 \neq 0} \widehat{V}_1\left(\frac{m_1}{q}\right) \widehat{V}_2\left(\frac{m_2}{q}\right) \widehat{V}_3\left(\frac{m_3}{q}\right) S_3(am_1, m_2, m_3; q).$$

According to (3.6), the Fourier transform $\widehat{V}_i(\xi)$ is essentially supported in the interval $|\xi| < X^{\varepsilon} q N_i^{-1}$. By dyadic decomposition, it suffices to consider the sum

$$\Delta_1 := \frac{N_1 N_2 N_3}{q^2} \sum_{m_1 \sim M_1} \sum_{m_2 \sim M_2} \sum_{m_3 \sim M_3} \alpha_1(m_1)\alpha_2(m_2)\alpha_3(m_3) S_3(am_1, m_2, m_3; q),$$

where $0 < M_j < X^{\varepsilon} q N_j^{-1}$ for $1 \leqslant j \leqslant 3$, and we write

$$\alpha_j(m_j) = \frac{1}{N_j} \widehat{V}_j\left(\frac{m_j}{q}\right), \quad j = 1, 2, 3, \tag{4.5}$$

so that

$$\alpha_j(m_j) \ll (\log X)^B, \quad j = 1, 2, 3.$$

We may regard Δ_1 as a trilinear form of Kloosterman sums, and by re-grouping variables, we can apply the estimates for bilinear forms in Lemma 2.3 to produce non-trivial bounds. To do so, one has to transform $S_3(am_1, m_2, m_3; q)$ to the shape $S_3(am_1m_2m_3, 1, 1; q)$, for which we appeal to the Kuznetsov–Selberg-type identity as given in Lemma 2.1, getting

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

$$\Delta_1 = \frac{N_1N_2N_3}{q^2}\sum_{drs|q} d\mu(rs) \sum_{\substack{m_1\sim M_1, m_2\sim M_2, m_3\sim M_3\\ d|(m_1,m_2,m_3)\\ r|(m_1/d, m_2/d)\\ s|(m_1m_2/(dr)^2, m_3/d)}}\sum\sum$$
$$\times \alpha_1(m_1)\alpha_2(m_2)\alpha_3(m_3) S_3\left(\frac{\overline{rs}m_1m_2m_3}{d^3r^2s^2}, 1, 1; \frac{q}{drs}\right)$$
$$= \frac{N_1N_2N_3}{q^2}\sum_{drs|q} d\mu(rs) \sum_{\substack{m_1\sim M_1/dr, m_2\sim M_2/dr, m_3\sim M_3/ds\\ s|m_1m_2}}\sum\sum$$
$$\times \alpha_1(m_1dr)\alpha_2(m_2dr)\alpha_3(m_3ds)\mathrm{Kl}_3\left(\frac{\overline{rs}m_1m_2m_3}{s}, \frac{q}{drs}\right). \tag{4.6}$$

Define

$$\gamma(l, dr) = \sum_{m_1m_2=l}\sum \alpha_1(m_1dr)\alpha_2(m_2dr).$$

Therefore,

$$\Delta_1 = \frac{N_1N_2N_3}{q^2}\sum_{drs|q} d\mu(rs) \sum_{M_1M_2/(dr)^2s < l \leqslant 4M_1M_2/(dr)^2s} \sum_{m_3\sim M_3/ds}$$
$$\times \gamma(ls, dr)\alpha_3(m_3ds)\mathrm{Kl}_3\left(\overline{rs}lm_3, \frac{q}{drs}\right).$$

For fixed d, r, s with $drs|q$, we are now in a position to apply Lemma 2.3 to the inner sums over l, m_3 with

$$q \to \frac{q}{drs}, \qquad M \asymp \frac{M_1M_2}{(dr)^2s}, \qquad N \asymp \frac{M_3}{ds},$$

getting

$$\Delta_1 \ll \frac{N_1N_2N_3}{q^2}\sum_{drs|q} d\mu^2(rs)\left|\sum_l\sum_{m_3}\gamma(ls, dr)\alpha_3(m_3ds)\mathrm{Kl}_3\left(\overline{rs}lm_3, \frac{q}{drs}\right)\right|$$
$$\ll X^{\varepsilon}\frac{N_1N_2N_3}{q^2}\sum_{drs|q} d\mu^2(rs)\frac{M_1M_2M_3}{d^3(rs)^2}$$
$$\times\left\{\varepsilon_1\left(\frac{q}{drs}\right)^{\kappa/2}\left(\frac{M_1M_2}{(dr)^2s}\right)^{(\lambda-\kappa-1)/2} + \left(\frac{q}{drs}\right)^{-1/4} + \left(\frac{M_3}{ds}\right)^{-1/2}\right\},$$

where $\varepsilon_1 = 1$ if $M_1M_2/(dr)^2s < q/drs$ and vanishes otherwise. Therefore,

$$\Delta_1 \ll X^{\varepsilon}N_1N_2N_3(M_1M_2)^{(\lambda-\kappa+1)/2}M_3q^{\kappa/2-2}\sum_{drs|q}(dr)^{\kappa/2-\lambda-1}s^{-(\lambda+3)/2}$$
$$+ \frac{X^{\varepsilon}N_1N_2N_3M_1M_2M_3}{q^2}(q^{-1/4} + M_3^{-1/2})$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

$$\ll X^{\varepsilon}N_1N_2N_3(M_1M_2)^{(\lambda-\kappa+1)/2}M_3q^{\kappa/2-2} + \frac{X^{\varepsilon}N_1N_2N_3M_1M_2M_3}{q^2}(q^{-1/4}+M_3^{-1/2}).$$

Recall that $0 < M_j < X^{\varepsilon}qN_j^{-1}$ for $1 \leqslant j \leqslant 3$. Hence, we have

$$\Delta_1 \ll X^{\varepsilon}\{q^{\lambda-\kappa/2}X^{(\kappa-\lambda+1)/2}N_3^{(\lambda-\kappa-1)/2} + q^{3/4} + (qN_3)^{1/2}\}.$$

This would yield

$$\mathscr{E}_1(\boldsymbol{N}; q, a) \ll \frac{X}{q}(\log X)^{-A}$$

as long as

$$q^{\lambda-\kappa/2}X^{(\kappa-\lambda+1)/2}N_3^{(\lambda-\kappa-1)/2} + q^{3/4} + (qN_3)^{1/2} \ll \frac{X^{1-\varepsilon}}{q}.$$

This is implied by (4.2a) while noting the definitions of μ_3 and θ.

Applying Lemma 2.4 to the m_1, m_2-sums in (4.6), we get

$$\Delta_1 \ll X^{\varepsilon}\{q^{2/3} + (qN_2)^{1/2} + q^{1/3}(N_1N_2)^{1/2}\}.$$

This is acceptable if (4.2b) holds. This completes the proof of Proposition 4.1.

We now start from (3.12), which is obtained by applying Poisson summation twice. According to the decays of Fourier transforms in (3.6), it suffices to consider the sum

$$\Delta_2 := \frac{N_2N_3}{q^{3/2}} \sum_{\substack{n_1\in\mathbf{Z}\\(n_1,q)=1}} \sum_{m_2\sim M_2} \sum_{m_3\sim M_3} V_1(n_1)\alpha_2(m_2)\alpha_3(m_3)S_2(an_1m_2, m_3; q),$$

where $0 < M_j < X^{\varepsilon}qN_j^{-1}$ for $2 \leqslant j \leqslant 3$, and α_2, α_3 are defined by (4.5). Proposition 4.2 follows from Weil's bound for S_2. In fact, from (2.2) it follows that

$$\Delta_2 \ll \frac{X^{\varepsilon}N_1N_2N_3\tau(q)}{q^{3/2}} \sum_{m_2\sim M_2} \sum_{m_3\sim M_3} (m_2, m_3, q)^{1/2} \ll X^{1+\varepsilon}q^{1/2}(N_2N_3)^{-1}.$$

This yields

$$\mathscr{E}_2(\boldsymbol{N}; q, a) \ll \frac{X}{q(\log X)^A}$$

under the condition of (4.3).

Proposition 4.3 follows from trivial treatments to each summand.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

§5. *Proof of Theorem 1.1.* As argued in §3, it suffices to prove (3.10) uniformly for all triples $\boldsymbol{N} = (N_1, N_2, N_3)$, satisfying (3.8), (3.9), and (3.3) and for all integers a coprime to q satisfying $q \asymp X^{\theta}$.

First, we show all possible cases for (μ_1, μ_2, μ_3) (and thus for $\boldsymbol{N} = (N_1, N_2, N_3)$) can be covered by Propositions 4.1 and 4.2. We now assume

$$\mu_2 + \mu_3 \leqslant \tfrac{3}{2}\theta, \quad \mu_3 \leqslant \theta; \tag{5.1}$$

otherwise, one can employ Propositions 4.2 or 4.3.

Furthermore, we assume (4.2b) is not valid, and it suffices to show that the resultant μ_2, μ_3 fall into the area given by (4.2a).

Case 1: $\mu_2 \geqslant 2 - 3\theta$. From (5.1) it follows that

$$\mu_3 \leqslant \tfrac{3}{2}\theta - (2 - 3\theta) < 2 - 3\theta \leqslant \mu_2$$

if $\frac{1}{2} < \theta < \frac{8}{15}$, in which case we have a contradiction to (4.1). We thus find that $\mu_2 \geqslant 2 - 3\theta$ does not hold if $\frac{1}{2} < \theta < \frac{8}{15}$.

Case 2: $\mu_2 < 2 - 3\theta$ *and* $\mu_3 \leqslant \frac{8}{3}\theta - 1$. Note that

$$\mu_3 \leqslant \tfrac{8}{3}\theta - 1 < 2 - 3\theta$$

if $\frac{1}{2} < \theta < \frac{9}{17}$.

It remains to consider the restriction

$$\frac{(2\lambda - \kappa + 2)\theta + \kappa - \lambda - 1}{\kappa - \lambda + 1} < \mu_3,$$

and we would like to show this is redundant by choosing suitable (κ, λ). Note that $\mu_3 \geqslant \frac{1}{3}$. For $(\kappa, \lambda) = (\frac{1}{2}, \frac{1}{2})$, we have

$$\frac{(2\lambda - \kappa + 2)\theta + \kappa - \lambda - 1}{\kappa - \lambda + 1} = \frac{5}{2}\theta - 1,$$

which is clearly smaller than $\frac{1}{3}$ if $\frac{1}{2} < \theta < \frac{8}{15}$.

Therefore, we may conclude

$$\mathscr{E}_j(\boldsymbol{N}; q, a) \ll \frac{X}{q(\log X)^A}$$

for $j = 1, 2, 3$ as long as $\frac{1}{2} < \theta < \frac{9}{17}$. To conclude Theorem 1.1, it remains to show, in view of Proposition 3.1 and the above estimate, that

$$\frac{X}{qN_j}\tau(q)(\log X)^{3B} \ll \frac{X}{q(\log X)^A}, \quad j = 1, 2, 3 \tag{5.2}$$

holds while applying Propositions 4.1, 4.2, or 4.3. In fact, we will show that N_j is at least a certain positive power of X. Thanks to Heath-Brown [**11**], we may assume $\frac{1}{2} + \frac{1}{82} \leqslant \theta < \frac{9}{17}$.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

The case $j = 3$ is obvious since $N_3 \geqslant X^{1/3-\varepsilon}$ due to (3.8) and (3.9).

While applying Proposition 4.2, one requires that $N_3 \ll X^{\theta}$ and $N_2 N_3 \gg X^{3/2\theta}$, which in fact implies $N_2 \gg X^{1/2\theta} \gg X^{1/4}$, and thus gives (5.2) for $j = 2$.

It remains to consider the case $j = 1$. Under the condition (4.2a) and assuming that $\mu_2 + \mu_3 \leqslant \frac{3}{2}\theta$, one has

$$\mu_1 \geqslant 1 - (\mu_2 + \mu_3) - B\frac{\log\log X}{\log X} \geqslant 1 - \frac{3}{2}\theta - B\frac{\log\log X}{\log X} > 1 - \frac{3}{2}\cdot\frac{11}{17} = \frac{1}{34},$$

provided that X is large enough. On the other hand, under the condition (4.2b) and assuming that $\mu_3 \leqslant \theta$, one has similarly

$$\mu_1 \geqslant 1 - (2 - 3\theta) - \theta - B\frac{\log\log X}{\log X} = 2\theta - 1 - B\frac{\log\log X}{\log X} \geqslant \frac{1}{42}.$$

Each condition implies the lower bound $N_1 \gg X^{1/42}$, which yields (5.2) for $j = 1$, completing the proof of Theorem 1.1.

Acknowledgements. This work is supported in part by NSFC (No. 11601413). I am grateful to the referee for valuable comments and suggestions, which led to a more polished version of this paper.

A. *Appendix. Kuznetsov–Selberg-type identity for Kloosterman sums.* Let us recall the classical Kuznetsov–Selberg identity

$$S_2(m, n; q) = \sum_{d|(m,n,q)} \sqrt{d}\, S_2(mnd^{-2}, 1; qd^{-1}).$$

Since Kuznetsov's proof is based on the trace formula, it is interesting to ask whether automorphic forms are really necessary. In fact, there are at least two existing proofs, both of which are supposed to be elementary. The first one was given by Matthes [**17**] using the Chinese remainder theorem and Möbius inversion; and the second is due to Andersson [**1**], however the theory of Hecke operators was applied. Here we give a completely elementary and direct proof, which only requires the orthogonality of additive characters.

A.1. *An elementary proof of (2.4).* Our method works for the following slightly more general sum

$$\Xi_k(m, n; q) = \frac{1}{\sqrt{q}} \sum_{\substack{x,y \,(\mathrm{mod}\, q) \\ xy \equiv k \,(\mathrm{mod}\, q)}}\sum \mathrm{e}\left(\frac{mx + ny}{q}\right).$$

In particular, we have $\Xi_1(m, n; q) = S_2(m, n; q)$. We shall prove

$$\Xi_k(m, n; q) = \sum_{d|(m,n,q)} \sqrt{d}\, S_2(mnd^{-2}, k; qd^{-1}) \tag{A.1}$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

and

$$\Xi_k(m,n;q)=\sum_{d|(m,k,q)}\sqrt{d}S_2(mkd^{-2},n;qd^{-1}). \tag{A.2}$$

Thus, (2.4) follows by taking $k=1$ in (A.1).

Remark A.1. One may see from (A.1) and (A.2) that $\Xi_k(m,n;q)$ is stable under permutations among m,n,k. This is at least observed by Andersson [**1**].

Using the orthogonality of additive characters to detect the congruence condition, we have

$$\begin{aligned}\Xi_k(m,n;q)&=\frac{1}{q^{3/2}}\sum_{a\,(\mathrm{mod}\,q)}\sum_{x,y\,(\mathrm{mod}\,q)}\mathrm{e}\left(\frac{x(m+ay)+ny-ka}{q}\right)\\&=\frac{1}{q^{3/2}}\sum_{d|q}\sum_{\substack{a\,(\mathrm{mod}\,q/d)\\(a,q/d)=1}}\sum_{x,y\,(\mathrm{mod}\,q)}\mathrm{e}\left(\frac{x(m+ady)+ny-kad}{q}\right).\end{aligned}$$

We evaluate the x-sum by orthogonality, getting

$$\Xi_k(m,n;q)=\frac{1}{\sqrt{q}}\sum_{d|q}\sum_{\substack{a\,(\mathrm{mod}\,q/d)\\(a,q/d)=1}}\sum_{\substack{y\,(\mathrm{mod}\,q)\\ady\equiv -m\,(\mathrm{mod}\,q)}}\mathrm{e}\left(\frac{ny-kad}{q}\right).$$

For each fixed $d|q$, from the condition $ady\equiv -m\,(\mathrm{mod}\,q)$ it follows that $d|m$. In a similar way, we also conclude that $d|n$. We thus have

$$\begin{aligned}\Xi_k(m,n;q)&=\frac{1}{\sqrt{q}}\sum_{d|(m,n,q)}\sum_{\substack{a\,(\mathrm{mod}\,q/d)\\(a,q/d)=1}}\sum_{\substack{y\,(\mathrm{mod}\,q)\\ y\equiv -\bar{a}m/d\,(\mathrm{mod}\,q/d)}}\mathrm{e}\left(\frac{ny-kad}{q}\right)\\&=\frac{1}{\sqrt{q}}\sum_{d|(m,n,q)}d\sum_{\substack{a\,(\mathrm{mod}\,q/d)\\(a,q/d)=1}}\mathrm{e}\left(\frac{\bar{a}mn/d+kad}{q}\right)\\&=\sum_{d|(m,n,q)}\sqrt{d}S_2(mnd^{-2},k;qd^{-1}).\end{aligned}$$

This establishes (A.1).

On the other hand, we have

$$\begin{aligned}\Xi_k(m,n;q)&=\frac{1}{\sqrt{q}}\sum_{d|(k,q)}\sum_{\substack{y\,(\mathrm{mod}\,q)\\(y,q)=d}}\mathrm{e}\left(\frac{ny}{q}\right)\sum_{\substack{x\,(\mathrm{mod}\,q)\\ xy\equiv k\,(\mathrm{mod}\,q)}}\mathrm{e}\left(\frac{mx}{q}\right)\\&=\frac{1}{\sqrt{q}}\sum_{d|(k,q)}\sum_{\substack{y\,(\mathrm{mod}\,q/d)\\(y,q/d)=1}}\mathrm{e}\left(\frac{ny}{q/d}\right)\sum_{\substack{x\,(\mathrm{mod}\,q)\\ xyd\equiv k\,(\mathrm{mod}\,q)}}\mathrm{e}\left(\frac{mx}{q}\right).\end{aligned}$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

For each y with $(y, q/d) = 1$ and $d|(k, q)$, the innermost sum can be rewritten as

$$\sum_{\substack{x \,(\mathrm{mod}\, q) \\ xy \equiv k/d \,(\mathrm{mod}\, q/d)}} \mathrm{e}\left(\frac{mx}{q}\right) = \sum_{\substack{x \,(\mathrm{mod}\, q) \\ x \equiv \overline{y}k/d \,(\mathrm{mod}\, q/d)}} \mathrm{e}\left(\frac{mx}{q}\right).$$

Write $x = \overline{y}k/d + l \cdot q/d$ with $l \,(\mathrm{mod}\, d)$, the above exponential sum is

$$\sum_{l \,(\mathrm{mod}\, d)} \mathrm{e}\left(\frac{m(\overline{y}k/d + l \cdot q/d)}{q}\right) = \mathrm{e}\left(\frac{m\overline{y}k/d}{q}\right) \sum_{l \,(\mathrm{mod}\, d)} \mathrm{e}\left(\frac{ml}{d}\right).$$

By orthogonality of additive characters, this must vanish unless $d|m$, in which case it is equal to

$$d \cdot \mathrm{e}\left(\frac{\overline{y}mkd^{-2}}{q/d}\right).$$

Hence, we have

$$\begin{aligned}\Xi_k(m, n; q) &= \frac{1}{\sqrt{q}} \sum_{d|(m,k,q)} d \sum_{\substack{y \,(\mathrm{mod}\, q/d) \\ (y, q/d)=1}} \mathrm{e}\left(\frac{ny + \overline{y}mkd^{-2}}{q/d}\right) \\ &= \sum_{d|(m,k,q)} \sqrt{d} S_2(mkd^{-2}, n; qd^{-1})\end{aligned}$$

as claimed in (A.2).

A.2. *Extension to rank three Kloosterman sums.* We now prove (2.5) following the above strategy. Using the orthogonality of additive characters to detect the congruence condition, we have

$$\begin{aligned}&S_3(m, n, l; q) \\ &= \frac{1}{q^2} \sum_{a \,(\mathrm{mod}\, q)} \sum_{x,y \,(\mathrm{mod}\, q)}\sum \mathrm{e}\left(\frac{x(m + ayz) + ny + lz - a}{q}\right) \\ &= \frac{1}{q^2} \sum_{d|q} \sum_{\substack{a \,(\mathrm{mod}\, q/d) \\ (a, q/d)=1}} \sum_{x,y \,(\mathrm{mod}\, q)}\sum \mathrm{e}\left(\frac{x(m + adyz) + ny + lz - ad}{q}\right).\end{aligned}$$

We evaluate the x-sum by orthogonality, getting

$$S_3(m, n, l; q) = \frac{1}{q} \sum_{d|q} \sum_{\substack{a \,(\mathrm{mod}\, q/d) \\ (a, q/d)=1}} \sum_{\substack{y,z \,(\mathrm{mod}\, q) \\ adyz \equiv -m \,(\mathrm{mod}\, q)}}\sum \mathrm{e}\left(\frac{ny + lz - ad}{q}\right).$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

For each fixed $d|q$, from $adyz \equiv -m \pmod q$ it follows that $d|m$. In a similar way, we also conclude that $d|(n,l)$. We thus have

$$S_3(m,n,l;q) = \frac{1}{q}\sum_{d|(m,n,l,q)}\sum_{\substack{a \,(\mathrm{mod}\, q/d)\\(a,q/d)=1}}\sum_{\substack{y,z \,(\mathrm{mod}\, q)\\ yz\equiv -\overline{a}m/d \,(\mathrm{mod}\, q/d)}}\mathrm{e}\left(\frac{ny+lz-ad}{q}\right)$$
$$= \frac{1}{\sqrt{q}}\sum_{d|(m,n,l,q)} d^{3/2}\sum_{\substack{a \,(\mathrm{mod}\, q/d)\\(a,q/d)=1}}\mathrm{e}\left(\frac{a}{q/d}\right)\Xi_{m/d}(\overline{a}n/d, l/d; q/d).$$

Applying (A.1) twice, we may derive that

$$S_3(m,n,l;q) = \frac{1}{\sqrt{q}}\sum_{d|(m,n,l,q)}\sum_{r|(n/d,l/d,q/d)} d^{3/2}r^{1/2}$$
$$\times \sum_{\substack{a \,(\mathrm{mod}\, q/d)\\(a,q/d)=1}}\mathrm{e}\left(\frac{a}{q/d}\right)S_2(\overline{a}nl/(dr)^2, m/d; q/(dr))$$
$$= \frac{1}{\sqrt{q}}\sum_{d|(m,n,l,q)}\sum_{r|(n/d,l/d,q/d)}\sum_{s|(nl/(dr)^2,m/d,q/dr)} d^{3/2}(rs)^{1/2}$$
$$\times \sum_{\substack{a \,(\mathrm{mod}\, q/d)\\(a,q/d)=1}}\mathrm{e}\left(\frac{a}{q/d}\right)S_2\left(\frac{\overline{a}mnl}{d^3r^2s^2}, 1; \frac{q}{drs}\right).$$

Note that q is squarefree, and thus $(rs, q/(drs)) = 1$. From the Chinese remainder theorem, the last sum over a can be expressed as

$$\left(\frac{q}{drs}\right)^{1/2}\mu(rs)S_3\left(\frac{\overline{rs}mnl}{d^3r^2s^2}, 1, 1; \frac{q}{drs}\right),$$

from which we conclude that

$$S_3(m,n,l;q) = \sum_{d|(m,n,l,q)}\sum_{r|(n/d,l/d,q/d)}\sum_{s|(nl/(dr)^2,m/d,q/dr)}$$
$$\times d\mu(rs)S_3\left(\frac{\overline{rs}mnl}{d^3r^2s^2}, 1, 1; \frac{q}{drs}\right).$$

This establishes (2.5).

B. *Appendix. Estimate for two-dimensional exponential sums.* Let p be an odd prime and $\mathcal{F}$ a middle-extension sheaf on $\mathbf{A}^1_{\mathbf{F}_p}$, pointwise pure of weight 0 and geometrically irreducible. Suppose T is the trace function of $\mathcal{F}$ in the sense of Katz [**14**, §7.3.7]. For $\mathbf{l} = (l_1, l_2)$ and $\mathbf{h} = (h_1, h_2)$, we consider the double sum

$$\Pi(\mathbf{l}, \mathbf{h}; p) = \sideset{}{^*}\sum_{x,y \,(\mathrm{mod}\, p)}\sideset{}{^*}\sum T(xy)\overline{T((x+l_1)(y+l_2))}\mathrm{e}\left(\frac{h_1x+h_2y}{p}\right).$$

The following estimate presents a square-root cancellation when $\mathcal{F}$ is a Kloosterman sheaf of rank three.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

THEOREM B.1. *For $T(x) = \mathrm{Kl}_3(x, p)$, we have*

$$|\Pi(\mathbf{l}, \mathbf{h}; p)| \ll p \cdot (h_1, h_2, l_1, l_2, p).$$

Proof. The bound is trivial if $p \mid (h_1, h_2, l_1, l_2)$. We now assume $p \nmid (h_1, h_2, l_1, l_2)$. For fixed $u, v \pmod p$, define

$$H(u, v) = \sideset{}{^*}\sum_{\substack{x,y \pmod p \\ xy \equiv u \pmod p \\ (x+l_1)(y+l_2) \equiv v \pmod p}} \sideset{}{^*}\sum \mathrm{e}\left(\frac{h_1 x + h_2 y}{p}\right),$$

so that

$$\Pi(\mathbf{l}, \mathbf{h}; p) = \sum_{u,v \pmod p} \sum T(u)\overline{T(v)}H(u, v).$$

By the Parseval (discrete Plancherel) identity, we have

$$\Pi(\mathbf{l}, \mathbf{h}; p) = \sum_{s,t \pmod p} \sum \widehat{T}(s)\overline{\widehat{T}(t)}\widehat{H}(-s, t),$$

where

$$\widehat{T}(s) = \frac{1}{\sqrt{p}} \sum_{u \pmod p} T(u)\mathrm{e}\left(\frac{-su}{p}\right)$$

and

$$\widehat{H}(s, t) = \frac{1}{p} \sum_{u,v \pmod p} \sum H(u, v)\mathrm{e}\left(\frac{-su - tv}{p}\right).$$

After a change of variable, one may write

$$H(u, v) = \sideset{}{^*}\sum_{\substack{a \pmod p \\ l_2 a + l_1 u\bar{a} \equiv v - u - l_1 l_2 \pmod p}} \mathrm{e}\left(\frac{h_1 a + h_2 u\bar{a}}{p}\right),$$

from which and the orthogonality of additive characters, it follows that

$$\begin{aligned}\widehat{H}(s, t) &= \frac{1}{p} \sum_{u \pmod p} \sideset{}{^*}\sum_{a \pmod p} \mathrm{e}\left(\frac{h_1 a + h_2 u\bar{a}}{p}\right)\mathrm{e}\left(\frac{-su - t(l_2 a + l_1 u\bar{a} + u + l_1 l_2)}{p}\right) \\ &= \sideset{}{^*}\sum_{\substack{a \pmod p \\ (h_2 - tl_1)\bar{a} \equiv s + t \pmod p}} \mathrm{e}\left(\frac{(h_1 - tl_2)a - tl_1 l_2}{p}\right).\end{aligned}$$

This yields

$$\widehat{H}(s, t) = \mathrm{e}\left(\frac{(h_1 - tl_2)(h_2 - tl_1)\overline{(s + t)} - tl_1 l_2}{p}\right)$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

if $s+t \not\equiv 0 \pmod p$; and otherwise, $\widehat{H}(s,t)$ vanishes unless $h_2 \equiv tl_1 \pmod p$, in which case the variable a runs freely over $(\mathbf{Z}/p\mathbf{Z})^\times$. Therefore,

$$\Pi(\mathbf{l},\mathbf{h};p) = \Pi_1(\mathbf{l},\mathbf{h};p) + \Pi_2(\mathbf{l},\mathbf{h};p),$$

where

$$\Pi_1(\mathbf{l},\mathbf{h};p) = \sum_{\substack{s,t \pmod p \\ s\not\equiv t \pmod p}}\sum \widehat{T}(s)\overline{\widehat{T}(t)}\mathrm{e}\left(\frac{(h_1-tl_2)(h_2-tl_1)\overline{(t-s)}-tl_1l_2}{p}\right)$$

and

$$\Pi_2(\mathbf{l},\mathbf{h};p) = \sum_{\substack{t \pmod p \\ h_2\equiv tl_2 \pmod p}} |\widehat{T}(t)|^2 \sideset{}{^*}\sum_{a \pmod p} \mathrm{e}\left(\frac{(h_1-tl_2)a-tl_1l_2}{p}\right).$$

Note that

$$\Pi_2(\mathbf{l},\mathbf{h};p) = (p-1)\sum_{\substack{t \pmod p \\ h_2\equiv tl_1 \pmod p \\ h_1\equiv tl_2 \pmod p}} |\widehat{T}(t)|^2\mathrm{e}\left(\frac{-tl_1l_2}{p}\right) + O(p) \ll p$$

since at least one of h_1, h_2, l_1, l_2 is coprime to p.

It remains to prove

$$\Pi_1(\mathbf{l},\mathbf{h};p) \ll p.$$

From now on, we assume $T = \mathrm{Kl}_3$ so that

$$\widehat{T}(x) = \begin{cases} \mathrm{Kl}_2(\overline{x},p), & p\nmid x, \\ 0, & p \mid x. \end{cases}$$

Hence,

$$\Pi_1(\mathbf{l},\mathbf{h};p) = \sqrt{p}\sum_{s \pmod p}^* \mathrm{Kl}_2(\overline{s},p)W(s,\mathbf{l},\mathbf{h};p),$$

where

$$W(s,\mathbf{l},\mathbf{h};p) = \frac{1}{\sqrt{p}}\sum_{\substack{t \pmod p \\ t\not\equiv 0, s \pmod p}} \mathrm{Kl}_2(\overline{t},p)\mathrm{e}\left(\frac{(h_1-tl_2)(h_2-tl_1)\overline{(t-s)}-tl_1l_2}{p}\right).$$

Given a tuple $(\mathbf{l},\mathbf{h}) = (l_1,l_2,h_1,h_2) \in \mathbf{Z}^4$, we would like to show $s \mapsto W(s,\mathbf{l},\mathbf{h};p)$ is a *good* trace function up to some controllable tails. To do so, we would like to appeal to the following quantitative characterization due to Fouvry, Kowalski and Michel [4, Corollary 5.3].

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

PROPOSITION B.1. *Let $\mathbf{F}_p$ be a finite field and $\ell \neq p$ a prime. Let $f \in \mathbf{F}_p(X, Y)$ be a rational function given by $f = f_1/f_2$ with $f_i \in \mathbf{F}_p[X, Y]$ coprime polynomials, and assume that $\mathfrak{c}(f) := 1 + \deg(f_1) + \deg(f_2) < p$. Let $\mathcal{F}$ be a middle-extension sheaf on $\mathbf{A}^1_{\mathbf{F}_p}$ which is pointwise pure of weight 0 and f-disjoint in the sense of* [4, Definition 5.2].

Then there exists an absolute constant $A \geqslant 1$, independent of f and $\mathcal{F}$, such that for all $x \in \mathbf{F}_p$, we have

$$\frac{1}{\sqrt{p}} \sum_{\substack{y \in \mathbf{F}_p \\ f_2(x,y) \neq 0}} t_{\mathcal{F}}(y)\psi(f(x, y)) = -t_0(x) + t_1(x) + t_2(x),$$

where t_0 is the trace function of a middle-extension sheaf $\mathcal{G}_0$ of weight 0 on $\mathbf{A}^1_{\mathbf{F}_p}$ with

$$\mathfrak{c}(\mathcal{G}_0) \leqslant (2\mathfrak{c}(f)\mathfrak{c}(\mathcal{F}))^A$$

the function t_1 is zero except for a set of values of $x \in \mathbf{F}_p$ of size at most $(2\mathfrak{c}(f)\mathfrak{c}(\mathcal{F}))^A$, and it satisfies

$$\|t_1\|_\infty \leqslant (2\mathfrak{c}(f)\mathfrak{c}(\mathcal{F}))^A,$$

while the function t_2 satisfies

$$\|t_2\|_\infty \leqslant (2\mathfrak{c}(f)\mathfrak{c}(\mathcal{F}))^A p^{-1/2}.$$

Remark B.1. It is useful to make a comment on the f-disjointness here. According to [4, Definition 5.2], a middle-extension sheaf $\mathcal{F}$ is called f-disjoint, if $H^2_c(\mathbf{A}^1 \times \overline{\mathbf{F}}_p, \mathcal{F} \otimes \mathcal{L}_x)$ vanishes for all $x \in \overline{\mathbf{F}}_p$, where $\mathcal{L}_x = j_x^* \mathcal{L}_{\psi(f(X,Y))}$ with ψ an additive character and $j_x : \{x\} \times \mathbf{A}^1 \hookrightarrow \mathbf{A}^2$ being the closed immersion. Therefore, $\mathcal{F}$ is always f-disjoint if it is irreducible of rank at least two (e.g. Kloosterman sheaves in one or more variables).

Taking $\mathcal{F} = [-1]^*\mathcal{K}\ell_2$ and $f(x, y) = (h_1 - yl_2)(h_2 - yl_1)/(y - x) - yl_1l_2$ with

$$f_1(x, y) = (h_1 - yl_2)(h_2 - yl_1) - (y - x)yl_1l_2, \qquad f_2(x, y) = y - x$$

in Proposition B.1, we then find f_1 and f_2 are coprime mod p since $p \nmid (l_1, l_2, h_1, h_2)$. From Proposition B.1 we conclude that

$$W(s, \mathbf{l}, \mathbf{h}; p) = -W_0(s, \mathbf{l}, \mathbf{h}; p) + W_1(s, \mathbf{l}, \mathbf{h}; p) + W_2(s, \mathbf{l}, \mathbf{h}; p),$$

where, for some absolute $C > 0$, $s \mapsto W_0(s, \mathbf{l}, \mathbf{h}; p)$ is the trace function of a middle-extension sheaf of weight 0 on $\mathbf{A}^1_{\mathbf{F}_p}$ with the conductor at most C, the function $s \mapsto W_1(s, \mathbf{l}, \mathbf{h}; p)$ is zero except for a set $\mathcal{D}$ of values of $s \in \mathbf{F}_p$ of size $|\mathcal{D}| \leqslant C$,

$$\|W_1\|_\infty \leqslant C, \qquad \|W_2\|_\infty \leqslant Cp^{-1/2}.$$

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

Therefore,

$$\begin{aligned}\Pi_1(\mathbf{l},\mathbf{h};\,p) &= -\sqrt{p}\sideset{}{^*}\sum_{s\,(\mathrm{mod}\,p)}\mathrm{Kl}_2(\overline{s},\,p)W_0(s,\mathbf{l},\mathbf{h};\,p)\\ &\quad+\sqrt{p}\sum_{s\in\mathcal{D}}\mathrm{Kl}_2(\overline{s},\,p)W_1(s,\mathbf{l},\mathbf{h};\,p)+O(p)\\ &= -\sqrt{p}\sideset{}{^*}\sum_{s\,(\mathrm{mod}\,p)}\mathrm{Kl}_2(\overline{s},\,p)W_0(s,\mathbf{l},\mathbf{h};\,p)+O(p).\end{aligned}$$

Since $\mathrm{Kl}_2(\overline{x},\,p)$ has a singularity at $x=0$, however

$$\begin{aligned}W(0,\mathbf{l},\mathbf{h};\,p) &= \frac{1}{\sqrt{p}}\mathrm{e}\left(\frac{-(h_1l_1+h_2l_2)}{p}\right)\sideset{}{^*}\sum_{t\,(\mathrm{mod}\,p)}\mathrm{Kl}_2(t,\,p)\mathrm{e}\left(\frac{h_1h_2t}{p}\right)\\ &= \mathrm{e}\left(\frac{-(h_1l_1+h_2l_2)}{p}\right)\left\{\sum_{h_1h_2x+1\equiv 0\,(\mathrm{mod}\,p)}\mathrm{e}\left(\frac{x}{p}\right)+\frac{1}{p}\right\},\end{aligned}$$

which yields $W_0(x,\mathbf{l},\mathbf{h};\,p)$ has no singularity at $x=0$. We then find the associated sheaves with $\mathrm{Kl}_2(\overline{x},\,p)$ and $W_0(x,\mathbf{l},\mathbf{h};\,p)$ have no geometrically irreducible components that are geometrically isomorphic. The Riemann hypothesis of Deligne (see [2] or [8, Theorem 4.1]) then yields

$$\Pi_1(\mathbf{l},\mathbf{h};\,p)\ll p$$

with an absolute implied constant. This completes the proof of Theorem B.1 by combining the above arguments. □

References

1. J. Andersson, Summation formulae and zeta functions, *Doctoral Dissertation*, Stockholm University, 2006.
2. P. Deligne, La conjecture de Weil, II. *Publ. Math. Inst. Hautes Études Sci.* **52** (1980), 137–252.
3. É. Fouvry, Sur le problème des diviseurs de Titchmarsh. *J. Reine Angew. Math.* **357** (1985), 51–76.
4. É. Fouvry, E. Kowalski and Ph. Michel, On the conductor of cohomological transforms. *Preprint*, 2013, arXiv:1310.3603 [math.NT].
5. É. Fouvry, E. Kowalski and Ph. Michel, Algebraic trace functions over the primes. *Duke Math. J.* **163** (2014), 1683–1736.
6. É. Fouvry, E. Kowalski and Ph. Michel, On the exponent of distribution of the ternary divisor function. *Mathematika* **61** (2015), 121–144.
7. É. Fouvry, E. Kowalski and Ph. Michel, Algebraic twists of modular forms and Hecke orbits. *Geom. Funct. Anal.* **25** (2015), 580–657.
8. É. Fouvry, E. Kowalski and Ph. Michel, Trace functions over finite fields and their applications. In *Colloquium de Giorgi, 2013 and 2014*, Scuola Normale Superiore Pisa (2015), 7–35.
9. J. B. Friedlander and H. Iwaniec, Incomplete Kloosterman sums and a divisor problem (with an appendix by B. J. Birch and E. Bombieri). *Ann. of Math. (2)* **121** (1985), 319–350.
10. S. W. Graham and G. Kolesnik, *Van der Corput Method of Exponential Sums* (*London Mathematical Society Lecture Note Series* **126**), Cambridge University Press (Cambridge, 1991).
11. D. R. Heath-Brown, The divisor function $d_3(n)$ in arithmetic progressions. *Acta Arith.* **47** (1986), 29–56.
12. C. Hooley, An asymptotic formula in the theory of numbers. *Proc. Lond. Math. Soc. (3)* **7** (1957), 396–413.
13. A. J. Irving, The divisor function in arithmetic progressions to smooth moduli. *Int. Math. Res. Not. IMRN* **15** (2015), 6675–6698.

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press

14. N. M. Katz, *Exponential Sums and Differential Equations* (*Annals of Math. Studies* **124**), Princeton University Press (1990).
15. N. V. Kuznetsov, The Petersson conjecture for cusp forms of weight zero and the Linnik conjecture. Sums of Kloosterman sums. *Mat. Sb.* **111(153)** (1980), 334–383; 479.
16. Yu. V. Linnik, All large numbers are sums of a prime and two squares (A problem of Hardy and Littlewood). II. *Mat. Sb.* **53(95)** (1961), 3–38.
17. R. Matthes, An elementary proof of a formula of Kuznecov for Kloosterman sums. *Results Math.* **18** (1990), 120–124.
18. D. H. L. Polymath, New equidistribution estimates of Zhang type. *Algebra Number Theory* **8** (2014), 2067–2199.
19. A. Selberg, *Lectures on Sieves* (*Collected Papers* **II**), Springer (1991), 65–247.
20. R. A. Smith, On n-dimensional Kloosterman sums. *J. Number Theory* **11** (1979), 324–343.
21. R. A. Smith, A generalization of Kuznietsov's identity for Kloosterman sums. *C. R. Math. Rep. Acad. Sci. Canada* **2**(6) (1980), 315–320.
22. J. Wu and P. Xi, Arithmetic exponent pairs for algebraic trace functions and applications. *Preprint*, 2016, arXiv:1603.07060 [math.NT].

Ping Xi,
School of Mathematics and Statistics,
Xi'an Jiaotong University,
Xi'an 710049,
P.R. China
E-mail: ping.xi@xjtu.edu.cn

https://doi.org/10.1112/S0025579318000220 Published online by Cambridge University Press