

BILINEAR FORMS WITH KLOOSTERMAN SUMS AND MOMENTS OF TWISTED L -FUNCTIONS

DJORDJE MILIĆEVIĆ, XINHUA QIN, AND XIAOSHENG WU

ABSTRACT. We establish power-saving estimates for general bilinear forms with Kloosterman sums modulo arbitrary q , including when both variables are shorter than the Pólya–Vinogradov range. As an application, we obtain power-saving asymptotics for the second moment of (holomorphic or Maaß) modular L -functions twisted with Dirichlet characters to an arbitrary large admissible modulus q . The bounds obtained are independent of the Ramanujan–Petersson conjecture and remove all factorability conditions on q in the work of Blomer, Fouvry, Kowalski, Michel, Milićević, and Sawin.

1. INTRODUCTION

1.1. Bilinear forms with Kloosterman sums. Among the indispensable tools of analytic number theory are estimates on *bilinear forms* of type

$$(1.1) \quad \sum_m \sum_n \alpha_m \beta_n K(mn),$$

where $K : \mathbb{N} \rightarrow \mathbb{C}$ is a suitable arithmetic kernel depending on the particular situation and $\alpha = (\alpha_m)$ and $\beta = (\beta_n)$ are finitely supported sequences of coefficients on which available information is such that they need to be treated as essentially arbitrary beyond the size of $\|\alpha\|_2 = (\sum_m |\alpha_m|^2)^{1/2}$ and $\|\beta\|_2 = (\sum_n |\beta_n|^2)^{1/2}$. Of particular importance is the situation where $K : \mathbb{Z}/q\mathbb{Z} \rightarrow \mathbb{C}$ is a periodic kernel, often itself a “trace function” or otherwise a complete exponential or character sum modulo q , and coefficients are supported on intervals $m \leq M$, $n \leq N$ with M, N as small as possible compared to the modulus q . A major threshold for the lengths of the intervals is the Pólya–Vinogradov range $M, N \sim q^{\frac{1}{2}+\varepsilon}$; for M and N in shorter ranges, the completion method through the Fourier transform does not make the sums shorter.

Following approaches such as those we discuss in §2.3, at the heart of the proofs of such estimates are bounds on complete *sums of products* (such as our (4.9)), which are highly sensitive to the arithmetic structure of the modulus q . We refer the reader to [8, 14] for comprehensive surveys and several foundational results.

2010 *Mathematics Subject Classification.* Primary 11L07, Secondary 11D79, 11F66, 11L05, 11T23.

Key words and phrases. Kloosterman sum, bilinear form, asymptotic formula, L -functions, character twists, sums of products.

D. Milićević supported in part by the Simons Foundation Award MPS-TSM-00008085 and by the Charles Simonyi Endowment. X. Qin and X. Wu supported in part by the NSFC Grant 12271135, Anhui Provincial Natural Science Foundation Grant 2508085J005 and the Fundamental Research Funds for the Central Universities Grant JZ2025HGTG0254.

When studying automorphic forms and L -functions, the kernel $K(n) = \text{Kl}_2(n; q)$ given by the normalized Kloosterman sum

$$\text{Kl}_2(a; q) = \frac{1}{\sqrt{q}} \sum_{x \pmod{q}}^* e_q(ax + \bar{x})$$

often arises naturally through applications of the classical Kuznetsov, Petersson, and Voronoi formulas. Power-saving bounds breaking the Pólya–Vinogradov threshold on the bilinear form (1.1) with these Kloosterman sums as the kernel played a critical role in the proofs of power-saving asymptotics for the second moment of modular L -functions in the series of papers by Blomer, Fouvry, Kowalski, Michel, Milićević, and Sawin [2–4, 14], which treated (1.1) in critical ranges by disjoint methods, with [4] covering the case of suitably factorable q by Weyl differencing and p -adic methods and [14] covering prime q by deep algebro-geometric methods including the machinery of trace functions. From the perspective of the arithmetic structure of the modulus q , these can be thought of as extreme cases of prime and well-factorable (including possibly powerful) q , leaving unaddressed a number of intermediate regimes such as $q = q_1 q_2$ and $q = q_1^2$ (with q_j prime and $q_1 \asymp q_2$) that do not neatly follow either of the two paradigms. For applications, one is naturally interested in results that apply to all moduli q uniformly, and our first result establishes such a bound on (1.1).

Theorem 1.1. *Let q be a positive integer, $M, N \geq 1$, and let $\alpha = (\alpha_m)$, $\beta = (\beta_n)$ be two sequences supported respectively on $[1, M]$ and $[1, N]$. If the conditions*

$$(1.2) \quad 1 \leq M \leq Nq^{\frac{1}{4}}, \quad M^{\frac{7}{5}}N < q^{\frac{3}{2}}, \quad MN \leq q^{\frac{5}{4}}$$

are satisfied, then for any integer c coprime with q , we have

$$(1.3) \quad \sum_{m \leq M} \sum_{n \leq N} \alpha_m \beta_n \text{Kl}_2(cmn; q) \ll q^\varepsilon \|\alpha\|_2 \|\beta\|_2 (MN)^{\frac{1}{2}} \left(M^{-\frac{1}{2}} q^{\frac{1}{6}} + M^{-\frac{3}{25}} N^{-\frac{3}{10}} q^{\frac{1}{5}} + (MN)^{-\frac{3}{16}} q^{\frac{11}{64}} \right).$$

Remark 1.1. The principal novelty in Theorem 1.1 over [4, 14] is that it holds uniformly over all moduli q . As a baseline, applying Weil’s bound directly gives the trivial estimate $\|\alpha\|_2 \|\beta\|_2 (MN)^{\frac{1}{2}}$ for the bilinear form. A bit more refined analysis, following the Pólya–Vinogradov method, yields the improved bound (see [9, Theorem 1.17])

$$(1.4) \quad \sum_{m \leq M} \sum_{n \leq N} \alpha_m \beta_n \text{Kl}_2(cmn; q) \ll \|\alpha\|_2 \|\beta\|_2 (MN)^{\frac{1}{2}} \left(q^{-\frac{1}{4}} + M^{-\frac{1}{2}} + N^{-\frac{1}{2}} q^{\frac{1}{4}} \log q \right),$$

which requires at least $N \gg q^{\frac{1}{2}} \log^2 q$ to improve upon the trivial bound. To contrast, in the balanced range $M \sim N$, which is the most important one in many applications including our Theorem 1.2, the bound (1.3) is nontrivial already for $M \sim N \gg q^{\frac{10}{21} + \delta}$ (and saves $q^{-\frac{1}{100} + \varepsilon}$ in the Pólya–Vinogradov range $M, N \sim q^{1/2 + \varepsilon}$).

We point out that Pascadi [15] has simultaneously and independently obtained power-saving bounds on bilinear forms in Kloosterman sums and asymptotics for moments of twisted L -functions similar to our Theorems 1.1 and 1.2. The two papers are complementary, and the two substantially different methods perform slightly better in different ranges and for different types of moduli, both pleasingly achieving power savings for bilinear

sums of square-root length (and substantially beyond) and general moduli. The methods of the present paper (which uses algebraic geometry and builds on Kowalski–Michel–Sawin [14] and Blomer–Milićević [4]) perform better for general moduli q and remove the dependence on the Ramanujan–Petersson conjecture in Theorem 1.2, while the methods of [15] (which uses non-abelian Fourier analysis and builds on Shkredov [17, 18]) perform better and in longer ranges for specific classes of moduli and can handle more general ranges of variables.

1.2. Moments of twisted L -functions. Asymptotics for moments with a power-saving error term are crucial to the amplification and related analytic techniques in questions such as subconvexity, nonvanishing, and extreme values of L -functions; see [3] for a rich sample of applications such as these and their implications for distribution of analytic ranks, modular symbols, and more.

For f a holomorphic or Maaß cuspidal newform of level 1, a natural family arises when considering the central values $L(1/2, f \otimes \chi)$ twisted by all primitive characters to a large modulus q , which may be seen as the finite place analogue of the family of archimedean twists $L(1/2 + it, f)$ along the critical line. Following a breakthrough power-saving asymptotic of Young [19] for the fourth moment $L(1/2, \chi)^4$ (the Eisenstein series analogue), power-saving asymptotics for the twisted second moment (such as (1.6) below) for f_j holomorphic cusp forms (or Maaß cusp forms satisfying the Ramanujan–Petersson conjecture) were obtained via bounds on bilinear forms with Kloosterman sums such as our Theorem 1.1 in the cases of suitably factorable q [4] and prime q [14]. In addition to these significant factorability conditions, the applicability of these results was also hampered by the dependence on the Ramanujan–Petersson conjecture as well as the technical condition that q not be highly divisible by prime 2.

As an important application of Theorem 1.1 we obtain the following evaluation of the twisted second moment, which is free of any dependence on the Ramanujan–Petersson conjecture and any factorability conditions on the modulus q . Let $\varphi^*(q)$ denote the number of primitive Dirichlet characters modulo q ; such characters exist iff $q \not\equiv 2 \pmod{4}$, in which case we term q admissible and we have $\varphi^*(q) = q^{1+o(1)}$. As in [4], the leading constants in the asymptotic evaluation are expressed in terms of the finite Euler products

$$(1.5) \quad \begin{aligned} P(s) &= \prod_{p|q} \left(1 - \frac{\lambda_1(p^2)}{p^s} + \frac{\lambda_1(p^2)}{p^{2s}} - \frac{1}{p^{3s}} \right) \left(1 - \frac{1}{p^{2s}} \right)^{-1}, \\ Q(s) &= \prod_{p|q} \left(1 - \frac{\lambda_1(p)\lambda_2(p)}{p^s} + \frac{\lambda_1(p^2) + \lambda_2(p^2)}{p^{2s}} - \frac{\lambda_1(p)\lambda_2(p)}{p^{3s}} + \frac{1}{p^{4s}} \right) \left(1 - \frac{1}{p^{2s}} \right)^{-1}. \end{aligned}$$

Our evaluation of the twisted second moment over arbitrary admissible moduli q is then as follows.

Theorem 1.2. *For $j = 1, 2$, let f_j be both holomorphic or both Maaß cuspidal newforms of level 1 with Hecke eigenvalues $\lambda_j(n)$. If their root numbers satisfy $\varepsilon(f_1)\varepsilon(f_2) = 1$, then*

$$(1.6) \quad \frac{1}{\varphi^*(q)} \sum_{\chi \pmod{q}}^* L(1/2, f_1 \otimes \chi) \overline{L(1/2, f_2 \otimes \chi)} = \frac{2}{\zeta(2)} M(f_1, f_2, q) + O\left(q^{-\frac{1}{216} + \varepsilon}\right),$$

where

$$(1.7) \quad M(f_1, f_2, q) = \begin{cases} P(1)L(1, \text{sym}^2 f_1) \left(\log q + c + \frac{P'(1)}{P(1)} \right), & f_1 = f_2, \\ Q(1)L(1, f_1 \times f_2), & f_1 \neq f_2, \end{cases}$$

where $P(s)$, $Q(s)$ are finite Euler products shown in (1.5), and c is a constant depending only on f_1 (not on q).

The main terms in Theorem 1.2 feature leading coefficients $L(1, \text{sym}^2 f_1)$ and $L(1, f_1 \times f_2)$, which do not vanish by the lower bounds of Hoffstein–Lockhart [11], Ramakrishnan–Wang [16], and Brumley [5]. In light of $P(s) = (L_q(s, \text{sym}^2 f_1)/\zeta_q(2s))^{-1}$ and $Q(s) = (L_q(s, f_1 \times f_2)/\zeta_q(2s))^{-1}$, we have $P(1), Q(1) = \exp(O_\varepsilon((\log \log q)^\varepsilon))$ and $P'(1)/P(1) = O_\varepsilon((\log \log q)^{1+\varepsilon})$, so that the main terms in Theorem 1.2 are not far from a linear polynomial in $\log q$ or a constant depending on f_1 and f_2 ; see Remark 7.1.

1.3. Notation. We follow the standard convention where ε denotes an arbitrarily small positive constant that may vary at each occurrence. As is common, we write $e(z) = e^{2\pi iz}$; we denote $f = O(g)$ or $f \ll g$ to indicate that $|f| \leq Cg$ for some constant $C > 0$ which is allowed to depend on ε (with ε allowed to vary from line to line as explained above) and (in bounds pertaining to Theorem 1.2 and its proof) on the fixed cusp forms f_1 and f_2 , but is otherwise absolute/independent from all other parameters unless specifically indicated by a subscript, and we write $f \asymp g$ to mean that $f \ll g$ and $g \ll f$ both hold. For $d, q \in \mathbb{N}$, we denote by q_d the maximal factor of q such that $(q_d, d) = 1$.

2. SKETCHING THE TREATMENT FOR BILINEAR FORMS

Theorem 1.1 is proved by combining (and then optimizing) the results of Theorem 2.1 and Theorem 2.2, which provide power-saving estimates on the bilinear forms with Kloosterman sums according to the factorization of the modulus q . In §2.1, we state these two theorems; then, in §2.2, we show how they combine to prove Theorem 1.1, and in §2.3, we give a general overview of their proofs, which then constitute most of the rest of the paper (sections 3–6).

2.1. Estimates on bilinear forms in the factorable and rough regimes. In this section, we state Theorems 2.1 and 2.2.

When q admits a favorable factorization such that there exists a divisor $s \mid q$ satisfying $s \in [q^{\delta_1}, q^{\frac{1}{2}-\delta_2}]$ with positive constants δ_1, δ_2 not excessively small, we establish the estimate for the bilinear form via the following theorem.

Theorem 2.1. *Let $M, N \geq 1$ and $s, q \in \mathbb{N}$ with $s \mid q$. Given two sequences $\alpha = (\alpha_m)$ and $\beta = (\beta_n)$ supported respectively on $[M, 2M]$ and $[N, 2N]$, then for any integer c coprime with q , it holds that*

$$(2.1) \quad \sum_{M \leq m \leq 2M} \sum_{N \leq n \leq 2N} \alpha_m \beta_n \text{Kl}_2(cmn; q) \ll q^\varepsilon \|\alpha\|_2 \|\beta\|_2 (MN)^{\frac{1}{2}} \left(M^{-\frac{1}{2}} s^{\frac{1}{2}} + q^{-\frac{1}{4}} s^{\frac{1}{4}} + N^{-\frac{1}{2}} q^{\frac{1}{4}} s^{-\frac{1}{4}} \right).$$

4

When q lacks favorable factorization (i.e., is essentially a prime or a product of two, not necessarily distinct, similarly-sized primes), we adopt the decomposition:

$$(2.2) \quad q = q^\star \rho,$$

where $q^\star$ denotes either a prime number or a product of two primes. Let $p_{\min}$ represent the smallest prime factor of $q^\star$. This scenario is addressed by the subsequent theorem.

Theorem 2.2. *Let $q = q^\star \rho$ as defined in (2.2) with $p_{\min} \geq (q/\rho)^{\frac{1}{3}}$. For real numbers M and N satisfying*

$$(2.3) \quad 1 \leq M \leq Nq^{\frac{1}{4}}, \quad MN \leq q^{\frac{5}{4}}\rho^{-\frac{1}{4}},$$

let N be an integer interval of length $\lfloor N \rfloor$, and let $\alpha = (\alpha_m)$ and $\beta = (\beta_n)$ be complex sequences supported respectively on $[M, 2M]$ and N . Then for any integer c coprime with q , we have

$$(2.4) \quad \sum_{M \leq m \leq 2M} \sum_{n \in N} \alpha_m \beta_n \text{Kl}_2(cmn; q) \ll q^\varepsilon \|\alpha\|_2 \|\beta\|_2 (MN)^{\frac{1}{2}} \left(M^{-\frac{1}{2}} + p_{\min}^{-\frac{1}{2}} + (MN)^{-\frac{3}{16}} q^{\frac{11}{64}} \rho^{\frac{9}{64}} \right).$$

2.2. Proof of Theorem 1.1. In this section, we prove Theorem 1.1 by combining the results of Theorem 2.1 and Theorem 2.2.

Proof of Theorem 1.1. The estimate (1.3) becomes nontrivial under the condition

$$M^{\frac{2}{5}} N \gg q^{\frac{2}{3}},$$

a convention we maintain throughout the subsequent proof. For given q , we take

$$\rho = \prod_{p \leq q^{1/3}, p^k \parallel q} p^k.$$

When $\rho \neq q$, the component $q^\star$ exhibits one of two forms: either a single prime or a product of two primes, with the smallest prime factor $p_{\min} > q^{\frac{1}{3}}$. The critical threshold

$$\rho_0 = M^{\frac{12}{25}} N^{-\frac{4}{5}} q^{\frac{1}{5}}$$

emerges as the equilibrium point where the last terms in (2.1) and (2.4) coincide when $s = \rho = \rho_0$. Through systematic analysis of the constraints $MN^{-1} \ll q^{\frac{1}{4}}$, $M^{\frac{2}{5}} N \gg q^{\frac{2}{3}}$, and $MN \leq q^{\frac{5}{4}}$, we derive two refined bounds for ρ_0 such that

$$(2.5) \quad \rho_0 = (MN^{-1})^{\frac{4}{7}} (M^{\frac{2}{5}} N)^{-\frac{8}{35}} q^{\frac{1}{5}} \ll (MN^{-1})^{\frac{2}{5}} q^{\frac{19}{210}},$$

$$(2.6) \quad \rho_0 = N^{-\frac{2}{3}} (MN^{-1})^{\frac{23}{75}} (MN)^{\frac{13}{75}} q^{\frac{1}{5}} \ll N^{-\frac{2}{3}} q^{\frac{37}{75}}.$$

The proof strategy bifurcates based on the relative magnitude of ρ compared to $\max\{1, \rho_0\}$:

Case I: $\rho \leq \max\{1, \rho_0\}$. Condition (2.3) is ensured by (1.2). Applying Theorem 2.2 yields

$$\begin{aligned} \sum_{m \leq M} \sum_{n \in N} \alpha_m \beta_n \text{Kl}_2(cmn; q) &\ll q^\varepsilon \|\alpha\|_2 \|\beta\|_2 (MN)^{\frac{1}{2}} \left(M^{-\frac{1}{2}} + q^{-\frac{1}{6}} + (MN)^{-\frac{3}{16}} q^{\frac{11}{64}} \rho^{\frac{9}{64}} \right) \\ &\ll q^\varepsilon \|\alpha\|_2 \|\beta\|_2 (MN)^{\frac{1}{2}} \left(M^{-\frac{1}{2}} + q^{-\frac{1}{6}} + M^{-\frac{3}{25}} N^{-\frac{3}{10}} q^{\frac{1}{5}} + (MN)^{-\frac{3}{16}} q^{\frac{11}{64}} \right). \end{aligned}$$

Case II: $\rho > \max\{1, \rho_0\}$. Here, there exists an integer $s \mid \rho$ satisfying

$$(2.7) \quad \rho_0 \leq s \leq \max\{q^{\frac{1}{3}}, \rho_0^2\}.$$

To justify this, observe that if any prime $p \mid \rho$ satisfies $\rho_0 \leq p \leq q^{\frac{1}{3}}$, we set $s = p$. Otherwise, all prime factors of ρ are smaller than ρ_0 , and the existence of such s is obvious. Applying Theorem 2.1 with this s gives

$$\begin{aligned} \sum_{m \leq M} \sum_{n \in N} \alpha_m \beta_n \text{Kl}_2(cmn; q) \\ \ll q^\varepsilon \|\alpha\|_2 \|\beta\|_2 (MN)^{\frac{1}{2}} \left(M^{-\frac{1}{2}} \left(q^{\frac{1}{6}} + \rho_0 \right) + q^{-\frac{1}{4}} \left(q^{\frac{1}{12}} + \rho_0^{\frac{1}{2}} \right) + N^{-\frac{1}{2}} q^{\frac{1}{4}} \rho_0^{-\frac{1}{4}} \right). \end{aligned}$$

From (2.5) and (2.6), it follows that

$$M^{-\frac{1}{2}} \rho_0 \ll N^{-\frac{1}{2}} q^{\frac{1}{4}} \rho_0^{-\frac{1}{4}}, \quad q^{-\frac{1}{4}} \rho_0^{\frac{1}{2}} \ll N^{-\frac{1}{2}} q^{\frac{1}{4}} \rho_0^{-\frac{1}{4}},$$

which simplifies the bound to

$$\sum_{m \leq M} \sum_{n \in N} \alpha_m \beta_n \text{Kl}_2(cmn; q) \ll q^\varepsilon \|\alpha\|_2 \|\beta\|_2 (MN)^{\frac{1}{2}} \left(M^{-\frac{1}{2}} q^{\frac{1}{6}} + q^{-\frac{1}{6}} + M^{-\frac{3}{25}} N^{-\frac{3}{10}} q^{\frac{1}{5}} \right).$$

Combining the estimates from both cases completes the proof of Theorem 1.1. $\square$

2.3. Sketch of the treatments for Theorems 2.1 & 2.2. An application of the Cauchy–Schwarz inequality gives

$$(2.8) \quad \left| \sum_{m \leq m \leq 2M} \sum_{n \in N} \alpha_m \beta_n \text{Kl}_2(cmn; q) \right|^2 \ll \|\beta\|_2^2 \sum_{n \in N} \left| \sum_{M \leq m \leq 2M} \alpha_m \text{Kl}_2(cmn; q) \right|^2,$$

which reduces the problem to estimating the sum

$$(2.9) \quad \sum_{n \in N} \left| \sum_{M \leq m \leq 2M} \alpha_m \text{Kl}_2(cmn; q) \right|^2.$$

The analysis of (2.9) bifurcates according to the factorization structure of q , requiring fundamentally distinct methods. When q admits a favorable factorization, we employ “Weyl differencing” to reduce the modulus prior the deeper Kloosterman sums analysis. This approach is detailed in §3, adapting the methodology of [4, Section 9] while eliminating auxiliary conditions required in [4, Theorem 5]. When q has large prime factors, we effectively reduce the modulus to the product of the two largest prime factors. Through “shifted by ab ” trick, we transfer the problem to handling multiplicative completed exponential sums. For prime moduli, these were rigorously treated by Kowalski, Michel, and Sawin [14]. The most intricate case occurs for the product of two primes, where we establish bounds by enumerating rational points on a nearly zero-dimensional variety in $\mathbb{F}_p^{11}$.

3. BILINEAR FORMS FOR WELL-FACTORABLE MODULI

In this section, we give the proof of Theorem 2.1 where q admits a favorable factorization. Given a divisor s of q with suitable size, we first apply “Weyl differencing” to reduce the modulus, then obtain power savings through careful analysis of Kloosterman sums, where the saving depends critically on the parameter s . This follows the approach of the work of Blomer and Milićević [4, Theorem 5], who established the bound

$$(3.1) \quad \sum_{\substack{M \leq m \leq 2M \\ (m,q)=1}} \left| \sum_{K \leq k \leq 2K} \lambda(k) \text{Kl}_2(ckm; q) \right|^2 \ll (KMq)^\varepsilon \left(KMs + \frac{K^2 Ms^{\frac{1}{2}}}{q^{\frac{1}{2}}} + \frac{K^2 q^{\frac{1}{2}}}{s^{\frac{1}{2}}} \right)$$

under the restrictive assumptions that $\lambda(k) \leq 1$, $(q/s, 2) = 1$, and $(m, q) = 1$. These conditions prevent direct application to our setting.

The following theorem presents a refined version of (3.1), where we eliminate all technical constraints. For consistency with the literature, we maintain the original notation from [4, Theorem 5], while adopting the standard Kloosterman sums.

Theorem 3.1. *Let $M, K \geq 1$, $s, q \in \mathbb{N}$ with $s \mid q$. Let $\lambda : \mathbb{N} \rightarrow \mathbb{C}$ be an arithmetic function supported on $[K, 2K]$. For any integer c coprime with q , we have*

$$(3.2) \quad \sum_{M \leq m \leq 2M} \left| \sum_{K \leq k \leq 2K} \lambda(k) \text{Kl}_2(ckm; q) \right|^2 \ll (qKM)^\varepsilon \|\lambda\|_2 \left(Ms + \frac{KMs^{\frac{1}{2}}}{q^{\frac{1}{2}}} + \frac{Kq^{\frac{1}{2}}}{s^{\frac{1}{2}}} \right).$$

The removal of the coprimality condition $(m, q) = 1$ obviates the need to introduce a factor $r \mid q$ as in [4, Theorem 5], since it makes no essential difference in the context of Möbius inversion. We therefore set $r = q$ without loss of generality. Theorem 2.1 follows immediately upon substituting the estimate (3.2) into (2.8).

3.1. The case for high powers of 2. The condition $(q/s, 2) = 1$ affects the bound (3.1) only when q is divisible by extremely high powers of 2. To fill this gap, we establish an estimate as that in [4, Proposition 23] for high powers of 2.

3.1.1. Generalities about the 2-adic square root. We begin by noting the obvious fact that, whenever $x \equiv y \pmod{2^{k-1}}$, $x^2 \equiv y^2 \pmod{2^k}$ automatically. Thus, when taking square roots of $u \pmod{2^k}$, we should be looking at $x \pmod{2^{k-1}}$ such that $x^2 \equiv u \pmod{2^k}$.

We claim that, for every $k \geq 3$, and for every $u \equiv 1 \pmod{8}$, there exist exactly two values of $x \pmod{2^{k-1}}$ (which arise in two 2-adic towers) such that

$$x^2 \equiv u \pmod{2^k}.$$

This follows by a standard Hensel’s Lemma argument. The claim is clearly true for $k = 3$: $x \equiv \pm 1 \pmod{4}$ satisfy $x^2 \equiv 1 \pmod{8}$. Inductively, if, for some $k \geq 3$, we have a solution

$$x_0^2 \equiv u \pmod{2^k}, \quad \text{that is,} \quad x_0^2 = u + 2^k Y,$$

and we are looking to lift this to a solution $x_0 + 2^{k-1}t \pmod{2^k}$ ($t \in \{0, 1\}$) satisfying

$$(x_0 + 2^{k-1}t)^2 \equiv u \pmod{2^{k+1}},$$

in light of $2k - 2 \geq k + 1$ this is equivalent to

$$2^k x_0 t \equiv -2^k Y \pmod{2^{k+1}}, \quad t \equiv -Y \overline{x_0} \pmod{2},$$

and thus we have a unique lift $x_0 + 2^{k-1}t \pmod{2^k}$. The claim follows by induction.

We can now denote the two (exact) 2-adic square roots by $\pm u_{1/2}(x)$ with $u_{1/2}(x) \equiv 1 \pmod{4}$, and we simply write $x_{1/2}$ for $u_{1/2}(x) \pmod{2^k}$ with a sufficiently large value of k . The claim indicates, for $k \geq 3$ and $x \equiv y \equiv 1 \pmod{8}$,

$$(3.3) \quad x \equiv y \pmod{2^k} \Leftrightarrow x_{1/2} \equiv y_{1/2} \pmod{2^{k-1}}.$$

Lemma 3.2. *If $u, u' \equiv 1 \pmod{8}$, $u \equiv u' \pmod{2^{\lambda+1}}$, $t \equiv t' \pmod{2^\lambda}$, and $k \geq 3$, then*

$$(3.4) \quad (u + 2^k t)_{1/2} - (u' + 2^k t')_{1/2} \equiv u_{1/2} - u'_{1/2} + \overline{u_{1/2}} 2^{k-1} t - \overline{u'_{1/2}} 2^{k-1} t' \pmod{2^{2k+\lambda-3}}.$$

Proof. The most direct way to prove this is by comparing

$$(u + 2^k t)_{1/2} \quad \text{and} \quad \sum_{j=0}^n \binom{1/2}{j} u_{1/2}^{1-2j} (2^k t)^j.$$

Indeed, using the agreement of formal power series $((1+x)^{1/2})^2 = 1+x$, we have that

$$\begin{aligned} \text{ord}_2 \left[\left(\sum_{j=0}^n \binom{1/2}{j} u_{1/2}^{1-2j} (2^k t)^j \right)^2 - (u + 2^k t) \right] &\geq \min_{j_1+j_2 \geq n+1} \text{ord}_2 \left[\binom{1/2}{j_1} \binom{1/2}{j_2} (2^k)^{j_1+j_2} \right] \\ &= \min_{j_1+j_2 \geq n+1} [(k-1)(j_1+j_2) - \text{ord}_2(j_1!) - \text{ord}_2(j_2!)] \geq (n+1)(k-2) + 2, \end{aligned}$$

where we use the obvious bound $\text{ord}_2(j!) = \lfloor j/2 \rfloor + \lfloor j/4 \rfloor + \dots \leq j-1$. From this it follows that, for $k \geq 3$,

$$(u + 2^k t)_{1/2} \equiv \sum_{j=0}^n \binom{1/2}{j} u_{1/2}^{1-2j} (2^k t)^j \pmod{2^{(n+1)(k-2)+1}}.$$

From this we conclude that

$$\begin{aligned} (u + 2^k t)_{1/2} - (u' + 2^k t')_{1/2} \\ \equiv u_{1/2} - u'_{1/2} + \sum_{j=1}^n \binom{1/2}{j} 2^{kj} \left(u_{1/2}^{1-2j} t^j - u'_{1/2}^{1-2j} t'^j \right) \pmod{2^{(n+1)(k-2)+1}}. \end{aligned}$$

From $u \equiv u' \pmod{2^{\lambda+1}}$, we have that $u_{1/2} \equiv u'_{1/2} \pmod{2^\lambda}$, whence

$$\text{ord}_2 \left[\binom{1/2}{j} 2^{kj} \left(u_{1/2}^{1-2j} t^j - u'_{1/2}^{1-2j} t'^j \right) \right] \geq (k-2)j + \lambda + 1.$$

Thus, choosing $n \geq \lambda + 1$, we conclude that

$$\text{ord}_2 \left[(u + 2^k t)_{1/2} - (u' + 2^k t')_{1/2} - \left(u_{1/2} - u'_{1/2} \right) - \left(\overline{u_{1/2}} 2^{k-1} t - \overline{u'_{1/2}} 2^{k-1} t' \right) \right] \geq 2k + \lambda - 3,$$

as announced. $\square$

3.1.2. Kloosterman sum evaluation. For $s \geq 6$ and $2 \nmid x$, let

$$(3.5) \quad \tau(x, 2^s) := \begin{cases} \frac{1}{2\sqrt{2}} \sum_{t \pmod{4}} e\left(\frac{xt^2}{4}\right) = \frac{1+e(x/4)}{\sqrt{2}} \in \{e^{\pm i\pi/4}\}, & 2 \mid s, \\ \frac{1}{4} \sum_{t \pmod{8}} e\left(\frac{xt^2}{8}\right) = \frac{1+2e(x/8)+e(x/2)}{2} = e(x/8), & 2 \nmid s, \end{cases}$$

which depends only on the parity of s and the value of $x \pmod{8}$.

Lemma 3.3. For $s \geq 6$ and $2 \nmid a$, then $S(a, b; 2^s) = 0$ unless $a \equiv b \pmod{8}$, in which case

$$S(a, b; 2^s) = 2^{(s+1)/2} \sum_{\pm} \tau(\pm(ab)_{1/2}, 2^s) e\left(\frac{\pm 2(ab)_{1/2}}{2^s}\right).$$

Proof. We begin by noting the congruence

$$\frac{b}{x + p^k t} \equiv \frac{b}{x} - \frac{b}{x^2} p^k t + \frac{b}{x^3} p^{2k} t^2 \pmod{p^{3k}},$$

which is valid for all $k \geq 1$ and irrespective of the value of p (including $p = 2$). Thus, $f(x) = ax + b\bar{x}$ satisfies

$$f(x_0 + p^k t) \equiv f(x_0) + \left(a - \frac{b}{x_0^2}\right) \cdot p^k t + \frac{b}{x_0^3} \cdot p^{2k} t^2 \pmod{p^{3k}}.$$

For s even, we have by the standard stationary phase argument (see [12, Lemmas 12.2 & 12.3]) that

$$S(a, b; 2^s) = \sum_{\substack{x \pmod{2^s} \\ a-b/x^2 \equiv 0 \pmod{2^{s/2}}}^*} e\left(\frac{ax + b\bar{x}}{2^s}\right).$$

In particular, for $s \geq 6$ even,

$$S(m, n; 2^s) = 0 \quad \text{unless} \quad a \equiv b \pmod{8},$$

while in the case $a \equiv b \pmod{8}$ the stationary phase congruence $a - b/x^2 \equiv 0 \pmod{2^{s/2}}$ is equivalent to $x \equiv \pm(b\bar{a})_{1/2} \pmod{2^{s/2-1}}$ and

$$\begin{aligned} S(a, b; 2^s) &= \sum_{\pm} \sum_{t \pmod{2^{s/2+1}}} e\left(\frac{f(\pm(b\bar{a})_{1/2} + 2^{s/2-1}t)}{2^s}\right) \\ &= 2^{s/2-1} \sum_{\pm} e\left(\frac{\pm 2(ab)_{1/2}}{2^s}\right) \sum_{t \pmod{4}} e\left(\frac{\pm a_{1/2}^3 \bar{b}_{1/2} t^2}{4}\right) \\ &= 2^{(s+1)/2} \sum_{\pm} \tau(\pm(ab)_{1/2}, 2^s) e\left(\frac{\pm 2(ab)_{1/2}}{2^s}\right). \end{aligned}$$

For $s \geq 7$ odd, the same argument shows again that $S(a, b; 2^s) = 0$ unless $a \equiv b \pmod{8}$, in which case $S(a, b; 2^s)$ equals

$$\begin{aligned} \sum_{\substack{x \pmod{2^s} \\ a-b/x^2 \equiv 0 \pmod{2^{(s-1)/2}}}^*} e\left(\frac{ax + b\bar{x}}{2^s}\right) &= \sum_{\pm} \sum_{t \pmod{2^{(s+1)/2+1}}} e\left(\frac{f(\pm(b\bar{a})_{1/2} + 2^{(s-1)/2-1}t)}{2^s}\right) \\ &= 2^{(s+1)/2-2} \sum_{\pm} e\left(\frac{\pm 2(ab)_{1/2}}{2^s}\right) \sum_{t \pmod{8}} e\left(\frac{\pm a_{1/2}^3 \bar{b}_{1/2} t^2}{8}\right) \\ &= 2^{(s+1)/2} \sum_{\pm} \tau(\pm(ab)_{1/2}, 2^s) e\left(\frac{\pm 2(ab)_{1/2}}{2^s}\right). \end{aligned}$$

This completes the proof. □

3.1.3. *Estimation of complete sums.* For $s \geq 2$ and $\epsilon \in \{\pm 1\}$, we will need estimates on the sum

$$\mathcal{S}_d^\epsilon(h, k_1, k_2, 2^s) := \sum_{m \pmod{2^s}}^* S^\epsilon(k_1, md, 2^s) \overline{S^\epsilon(k_2, md, 2^s)} e\left(-\frac{hm}{2^s}\right),$$

where

$$(3.6) \quad S^\epsilon(m, n, 2^s) = \begin{cases} 2^{(s+1)/2} \tau(\epsilon(mn)_{1/2}, 2^s) e\left(\frac{2\epsilon(ab)_{1/2}}{2^s}\right), & m \equiv n \pmod{8}, \\ 0, & \text{otherwise.} \end{cases}$$

Lemma 3.4. *For $2 \nmid k_1 k_2 d$, $s \geq 2$, and $\epsilon \in \{\pm 1\}$, we have*

$$\mathcal{S}_d^\epsilon(h, k_1, k_2, 2^s) \ll 2^{\frac{3s}{2}} (k_1 - k_2, h, 2^s)^{\frac{1}{2}} \delta_{(k_1 - k_2, 2^s) | 8h}.$$

Proof. Since the estimate is trivial for small s , we may assume that $s \geq 4$. We need to estimate

$$\begin{aligned} \mathcal{S}_d^\epsilon(h, k_1, k_2, 2^s) &= 2^{(s+1)} \sum_{\substack{m \pmod{2^s} \\ m \equiv k_1 \equiv k_2 \pmod{8}}} \tau(\epsilon(mk_1)_{1/2}, 2^s) \overline{\tau(\epsilon(mk_2)_{1/2}, 2^s)} \\ &\quad \times e\left(\frac{2\epsilon((mk_1)_{1/2} - (mk_2)_{1/2}) - \bar{d}hm}{2^s}\right). \end{aligned}$$

We write $(k_1 - k_2, 2^s) = 2^r$ with $r \geq 3$, $t = \lfloor (s+r)/2 \rfloor - 3$, and $m = m_1 + 2^{s-t}m_2$. Recall that $\tau(x, 2^s)$ depends only on x modulo 8. Applying Lemma 3.2 with $\lambda = r - 1$ and $k = s - t \geq 3$, we have

$$\begin{aligned} \mathcal{S}_d^\epsilon(h, k_1, k_2, 2^s) &= 2^{(s+1)} \sum_{\substack{m_1 \pmod{2^{s-t}} \\ m_1 \equiv k_1 \equiv k_2 \pmod{8}}} \tau(\epsilon(m_1 k_1)_{1/2}, 2^s) \overline{\tau(\epsilon(m_1 k_2)_{1/2}, 2^s)} \\ &\quad \times e\left(\frac{2\epsilon((m_1 k_1)_{1/2} - (m_1 k_2)_{1/2}) - \bar{d}hm_1}{2^s}\right) \\ &\quad \times \sum_{m_2 \pmod{2^t}} e\left(\frac{\epsilon m_2 (k_1 \overline{(m_1 k_1)_{1/2}} - k_2 \overline{(m_1 k_2)_{1/2}}) - \bar{d}hm_2}{2^t}\right), \end{aligned}$$

observing $2k + \lambda - 3 = 2s + r - 2t - 4 > s$. The m_2 -sum vanishes unless

$$(3.7) \quad k_1 \overline{(m_1 k_1)_{1/2}} - k_2 \overline{(m_1 k_2)_{1/2}} - \epsilon \bar{d}h \equiv 0 \pmod{2^t},$$

in which case it is equal to 2^t . Through (3.3), the condition $(k_1 - k_2, 2^s) = 2^r$ yields

$$(3.8) \quad (k_1 \overline{(m_1 k_1)_{1/2}} - k_2 \overline{(m_1 k_2)_{1/2}}, 2^s) = 2^{r-1}.$$

For $3 \leq r \leq s - 6$, we have $r - 1 < t$, and thus $2^{r-1} \parallel h$, derived from (3.7) and (3.8). Writing $h = h' 2^{r-1}$, we reduce (3.7) to

$$(3.9) \quad 2^{-(r-1)} (k_1 \overline{(m_1 k_1)_{1/2}} - k_2 \overline{(m_1 k_2)_{1/2}}) \equiv \epsilon \bar{d}h' \pmod{2^{t-r+1}}.$$

By introducing a fixed $u \equiv k_1 \pmod{8}$, we note

$$k_1 \overline{(m_1 k_1)_{1/2}} - k_2 \overline{(m_1 k_2)_{1/2}} \in \left\{ \pm \left(\frac{(k_1 u)_{1/2}}{(m_1 u)_{1/2}} - \frac{(k_2 u)_{1/2}}{(m_1 u)_{1/2}} \right) \right\},$$

and then reduce (3.9) to

$$\bar{d}h'(m_1u)_{1/2} \equiv \pm 2^{-r+1} ((k_1u)_{1/2} - (k_2u)_{1/2}) \pmod{2^{t-r+1}},$$

whence, there are at most $O(1)$ solutions $m_1 \pmod{2^{t-r+1}}$. Therefore, there are at most $O(2^{s-2t+r-1})$ contributing values of m_1 , whence

$$|\mathcal{S}_d^\epsilon(h, k_1, k_2, 2^s)| \ll 2^{s+1} 2^{s-2t+r-1} 2^t \ll 2^{\frac{3s}{2}} 2^{\frac{r}{2}},$$

in agreement with the statement of the Theorem.

For large $r \geq s - 5$, we have $r - 3 \leq t \leq r - 1$. Combining (3.7) and (3.8) yields $2^t \mid h$, and then a trivial estimate shows

$$|\mathcal{S}_d^\epsilon(h, k_1, k_2, 2^s)| \ll 2^{2s} \ll 2^{\frac{3s}{2}} 2^{\frac{t}{2}} \ll 2^{\frac{3s}{2}} (k_1 - k_2, h, 2^s)^{\frac{1}{2}}.$$

This completes the proof of the lemma. $\square$

3.2. Proof of Theorem 3.1. Since Lemmas 3.3 & 3.4 cover the $p = 2$ case, unsettled in Lemma 22 and Proposition 23 of [4], the argument of [4] then applies equally to these q containing extremely high powers of 2. This removes the condition $(q/s, 2) = 1$.

We next establish (3.2) while retaining the coprimality condition $(m, q) = 1$. The proof of this part largely follows [4, Theorem 5], with necessary adjustments to [4, Section 9] detailed below.

To account for $\lambda(k)$, we decouple it from $b_1^\epsilon(k)$ and adjust the definition [4, (9.9)] to

$$b_1(k) = \lambda(k) \sum_{\epsilon \in \{\pm 1\}^e} b_1^\epsilon(k),$$

where $b_1^\epsilon(k)$ retains its original definition in [4, (9.9)] but isolates $\lambda(k)$ as a separate factor.

The primary modification occurs at [4, (9.12)], which becomes

$$\Sigma \ll r^\epsilon H r_2^2 \sum_{\epsilon \in \{\pm 1\}^e} \sum_{(m, q)=1} w\left(\frac{m}{M}\right) \sum_{\substack{K \leq k_1, k_2 \leq 2K \\ k_1 \equiv k_2 \pmod{H r_2} \\ (k_1 k_2, r_1^\#)=1}} \lambda(k_1) \overline{\lambda(k_2)} b_1^\epsilon(k) \overline{b_1^\epsilon(k)},$$

with $r = r_1 r_2$, $(r_1, r_2) = 1$, $r_1^\#$ denoting the squarefull part of r_1 , and $\text{rad}(r_1^\#)$ its squarefree kernel, while H is a parameter satisfying

$$H \mid \frac{r_1^\#}{\text{rad}(r_1^\#)}.$$

Diagonal terms ($k_1 = k_2$) contribute

$$\ll (rM)^\epsilon H r_2^2 M r_1 \sum_{K \leq k \leq 2K} |\lambda(k)|^2 \ll (rKM)^\epsilon H r_2^2 M K r_1.$$

By symmetry in k_1 and k_2 , we apply

$$\lambda(k_1) \overline{\lambda(k_2)} \ll |\lambda(k_1)|^2 + |\lambda(k_2)|^2$$

to off-diagonal terms. Employing Möbius inversion to eliminate $(m, q) = 1$ yields

$$(3.10) \quad \Sigma \ll (rKM)^\varepsilon Hr_2^2 MKr_1 + r^\varepsilon Hr_2^2 \sum_{d|q} \sum_{\substack{K \leq k_1 \neq k_2 \leq 2K \\ k_1 \equiv k_2 \pmod{Hr_2} \\ (k_1 k_2, r_1^\#) = 1}} |\lambda(k_1)|^2 \left| \sum_{\epsilon \in \{\pm 1\}^e} \Sigma_d^\epsilon(k_1, k_2, r_1) \right|,$$

where $\Sigma_d^\epsilon(k_1, k_2, r_1)$ remains as defined in [4].

The second modification concerns the bound for $\sum_{\epsilon \in \{\pm 1\}^e} \Sigma_d^\epsilon(k_1, k_2, r_1)$. From the last formula in [4, p.502], it is shown that

$$\sum_{\epsilon \in \{\pm 1\}^e} \Sigma_d^\epsilon(k_1, k_2, r_1) \ll \sum_{\rho \mid \left(\frac{r_1}{r_1^\#}, d\right)} \frac{(k_1, \rho)(k_2, \rho)}{\rho^{\frac{1}{2}}} \frac{M}{d} r_1^{\frac{1}{2} + \varepsilon} (k_1 - k_2, r_1)^{\frac{1}{2}} \left(1 + \frac{dr_1}{M(k_1 - k_2, \frac{r_1^\#}{\text{rad}(r_1^\#)})\rho} \right).$$

Using

$$(k_1, \rho)(k_2, \rho) = (k_1 k_2, \rho)(k_1, k_2, \rho) \leq (k_1 k_2, \rho)(k_1 - k_2, \rho)$$

and $(\rho, H) = 1$ since $(\rho, r_1^\#) = 1$, we have

$$\sum_{\epsilon \in \{\pm 1\}^e} \Sigma_d^\epsilon(k_1, k_2, r_1) \ll \sum_{\substack{\rho \mid (d, r_1) \\ (\rho, H) = 1}} \frac{(k_1 - k_2, \rho)}{\rho^{\frac{1}{2}}} \left(Mr_1^{\frac{1}{2} + \varepsilon} (k_1 - k_2, r_1)^{\frac{1}{2}} + r_1^{\frac{3}{2} + \varepsilon} \frac{(k_1 - k_2, r_1)^{\frac{1}{2}}}{\left(k_1 - k_2, \frac{r_1^\#}{\text{rad}(r_1^\#)}\right)} \right).$$

Noting $\rho \mid r_1$ and $(\rho, Hr_2) = 1$, we have

$$\begin{aligned} & \sum_{K \leq k \leq 2K} |\lambda(k)|^2 \sum_{0 \neq l \ll \frac{K}{Hr_2}} \left| \sum_{\epsilon \in \{\pm 1\}^e} \Sigma_d^\epsilon(k, k + lHr_2, r_1) \right| \\ & \ll (dr_1)^\varepsilon \sum_{K \leq k \leq 2K} |\lambda(k)|^2 \sum_{0 \neq l \ll \frac{K}{Hr_2}} \sum_{\substack{\rho \mid (d, r_1) \\ (\rho, H) = 1}} \frac{(l, \rho)}{\rho^{\frac{1}{2}}} \left(Mr_1^{\frac{1}{2}} (l, r_1)^{\frac{1}{2}} H^{\frac{1}{2}} + r_1^{\frac{3}{2}} \frac{(l, r_1)^{\frac{1}{2}} H^{\frac{1}{2}}}{H} \right) \\ & \ll (dr_1)^\varepsilon \sum_{K \leq k \leq 2K} |\lambda(k)|^2 \sum_{0 \neq l \ll \frac{K}{Hr_2}} \left(Mr_1^{\frac{1}{2}} (l, r_1) H^{\frac{1}{2}} + r_1^{\frac{3}{2}} \frac{(l, r_1)}{H^{\frac{1}{2}}} \right) \\ & \ll (dr_1)^\varepsilon \frac{K^2}{Hr_2} MH^{\frac{1}{2}} r_1^{\frac{1}{2}} + (dr_1)^\varepsilon \frac{K^2}{Hr_2} \frac{r_1^{\frac{3}{2}}}{H^{\frac{1}{2}}}. \end{aligned}$$

Substituting this into (3.10) replicates the bound [4, (9.14)] for Σ . The remainder of the proof aligns with [4], yielding (3.2) with the coprimality condition $(m, q) = 1$ still present.

We finally eliminate $(m, q) = 1$ through a detailed analysis of Kloosterman sums. Let $d = (m, q)$, and for $p \mid d$ write $q = p^j q_p$. The twisted multiplicativity of Kloosterman sums gives

$$\text{Kl}_2(ckm; q) = \text{Kl}_2(c\bar{q}_p^2 km; p^j) \text{Kl}_2(c\bar{p}^{2j} km; q_p).$$

If $2 \mid d$, Lemma 3.3 shows that the Kloosterman sum vanishes unless q is divisible by only a small power of 2 (not exceeding 2^5). Since a small power of 2 can be removed at no cost, we may assume $(d, 2) = 1$ in the following analysis. For an odd prime $p \mid d$,

the Kloosterman sum $\text{Kl}_2(c\bar{q}_p^2 km; p^j)$ vanishes unless $j = 1$, in which case, it reduces to a Ramanujan sum. In other words,

$$\text{Kl}_2(c\bar{q}_p^2 km; p^j) = \begin{cases} -p^{-\frac{1}{2}}, & \text{if } j = 1; \\ 0, & \text{otherwise.} \end{cases}$$

Thus $\text{Kl}_2(ckm; q)$ vanishes unless $q = dq_d$ with d square-free, in which case,

$$\text{Kl}_2(ckm; q) = \frac{\mu(d)}{d^{\frac{1}{2}}} \text{Kl}_2(c\bar{d}^2 km; q_d).$$

Substituting this into (2.9) shows

$$\begin{aligned} \sum_{M \leq m \leq 2M} \left| \sum_{K \leq k \leq 2K} \lambda(k) \text{Kl}_2(ckm; q) \right|^2 &= \sum_{d|q} \frac{\mu^2(d)}{d} \sum_{\substack{M/d \leq m \leq 2M/d \\ (m, q_d)=1}} \left| \sum_{K \leq k \leq 2K} \lambda(k) \text{Kl}_2(c\bar{d} km; q_d) \right|^2 \\ &\ll (qKM)^\varepsilon \|\lambda\|_2 \left(Ms + \frac{KM s^{\frac{1}{2}}}{q^{\frac{1}{2}}} + \frac{Kq^{\frac{1}{2}}}{s^{\frac{1}{2}}} \right), \end{aligned}$$

which establishes the lemma.

4. BILINEAR FORMS FOR MODULI WITH LARGE PRIME FACTORS

From this section, we start our proof of Theorem 2.2.

4.1. Initial treatment for Theorem 2.2 with “shifted by ab ” trick. We expand the square in (2.9) to obtain

$$(4.1) \quad \sum_{N \leq n \leq 2N} \left| \sum_{M \leq m \leq 2M} \alpha_m \text{Kl}_2(cmn; q) \right|^2 = \sum_{d|q} \mathcal{T}_d(\alpha, 1_N),$$

where

$$\mathcal{T}_d(\alpha, 1_N) = \sum_{\substack{M \leq m_1, m_2 \leq 2M \\ (m_1 - m_2, q) = d}} \alpha_{m_1} \bar{\alpha}_{m_2} \sum_{n \in N} \text{Kl}_2(cm_1 n; q) \text{Kl}_2(cm_2 n; q).$$

A trivial estimate with the inequality $\alpha_{m_1} \bar{\alpha}_{m_2} \ll |\alpha_{m_1}|^2 + |\alpha_{m_2}|^2$ shows

$$(4.2) \quad \mathcal{T}_d(\alpha, 1_N) \ll \left(1 + \frac{M}{d}\right) N \|\alpha\|_2^2,$$

which will be employed when $(d, q^\star) > 1$. For the remaining case $d \mid \rho$, we apply Vinogradov’s “shifted by ab ” summation approach and analyze it through three distinct scenarios:

- $q^\star = p$ is a prime;
- $q^\star = p_1 p_2$ is a product of two distinct primes;
- $q^\star = p^2$ is a square of a prime.

For given $A, B \geq 1$ satisfying

$$2B < q, \quad AB \leq N, \quad AM < q,$$

we have

$$\begin{aligned}\mathcal{T}_d(\alpha, 1_N) &= \frac{1}{A^*B} \sum_{\substack{A \leq a \leq 2A \\ B \leq b \leq 2B}}^* \sum_{\substack{M \leq m_1, m_2 \leq 2M \\ (m_1 - m_2, q) = d}} \alpha_{m_1} \bar{\alpha}_{m_2} \\ &\quad \times \sum_{n+ab \in N} \text{Kl}_2(acm_1(\bar{a}n + b); q) \text{Kl}_2(acm_2(\bar{a}n + b); q),\end{aligned}$$

where A^* counts the integers $a \in [A, 2A]$ coprime to q . By performing the variable substitutions

$$am_1 \rightarrow l_1, \quad am_2 \rightarrow l_2, \quad \bar{a}n \rightarrow r$$

and introducing a smooth weight function $W(x) \geq 0$, which has compact support in $[1/2, 5]$ with $W^j(x) \ll 1$ for $j \geq 0$ and takes the value of 1 in $[1, 4]$, we obtain via the method of [10, p. 116] that

$$\begin{aligned}\mathcal{T}_d(\alpha, 1_N) &\ll \frac{q^\varepsilon}{AB} \sum_{r \pmod{q}} \sum_{(l_1 - l_2, q) = d} \sum_{l_1, l_2} \nu(r, l_1, l_2) W\left(\frac{l_1}{AM}\right) W\left(\frac{l_2}{AM}\right) \\ &\quad \times \left| \sum_{B \leq b \leq 2B} \gamma_b \text{Kl}_2(cl_1(r + b); q) \text{Kl}_2(cl_2(r + b); q) \right|,\end{aligned}$$

where the coefficient $\nu(r, l_1, l_2)$ aggregates contributions from

$$\nu(r, l_1, l_2) = \sum_{\substack{A \leq a \leq 2A, \\ am_1 = l_1, am_2 = l_2, \bar{a}n \equiv r \pmod{q}}}^* \sum_{M \leq m_1, m_2 \leq 2M} \cdots \sum_{n \in N'} |\alpha_{m_1}| |\alpha_{m_2}|.$$

Here $N' \supset N$ denotes an extended interval of length $2N$, and γ_b are some complex coefficients with $|\gamma_b| \leq 1$. The following bounds hold

$$\sum_r \sum_{(l_1 - l_2, q) = d} \sum_{l_1, l_2} \nu(r, l_1, l_2) W\left(\frac{l_1}{AM}\right) W\left(\frac{l_2}{AM}\right) \ll AN \|\alpha\|_1^2 \ll AMN \|\alpha\|_2^2$$

and

$$\begin{aligned}\sum_r \sum_{(l_1 - l_2, q) = d} \sum_{l_1, l_2} \nu(r, l_1, l_2)^2 W\left(\frac{l_1}{AM}\right) W\left(\frac{l_2}{AM}\right) &\ll \sum_{\substack{am_1 = a'm'_1 \\ am_2 = a'm'_2 \\ an \equiv a'n' \pmod{q}}} \cdots \sum |\alpha_{m_1} \alpha_{m'_1} \alpha_{m_2} \alpha_{m'_2}| \\ &\ll q^\varepsilon \sum_{a, m_1, m_2, n} |\alpha_{m_1}|^2 |\alpha_{m_2}|^2 \ll q^\varepsilon AN \|\alpha\|_2^4,\end{aligned}$$

since the cross terms satisfy

$$|\alpha_{m_1} \alpha_{m'_1} \alpha_{m_2} \alpha_{m'_2}| \leq (|\alpha_{m_1}|^2 + |\alpha_{m'_1}|^2) (|\alpha_{m_2}|^2 + |\alpha_{m'_2}|^2)$$

and m_i, m'_i are determined by each other up to $O(q^\varepsilon)$ possibilities. The congruence $a'n \equiv an'$ restricts n' to at most two values in N' due to its length $\leq 2q$. Applying the Hölder inequality yields

$$(4.3) \quad \mathcal{T}_d(\alpha, 1_N) \ll \frac{q^\varepsilon}{AB} (AN)^{\frac{3}{4}} M^{\frac{1}{2}} \|\alpha\|_2^2 \left(\sum_b |\Sigma^d(b, AM; q)| \right)^{\frac{1}{4}},$$

where the key exponential sum is

$$(4.4) \quad \Sigma^d(\mathbf{b}, AM; q) = \sum_{r \pmod{q}} \sum_{(l_1 - l_2, q) = d} \sum w\left(\frac{l_1}{AM}\right) w\left(\frac{l_2}{AM}\right) \prod_{i=1}^2 \prod_{j=1}^4 \text{Kl}_2(cl_i(r + b_j); q).$$

Remark 4.1. The Vinogradov’s “shifted by ab ” technique employed here constitutes a refinement of methods established in prior research [2] and [10]. A novel element lies in the incorporation of a smooth test function, which facilitates the treatment of the exponential sum via Poisson summation formula rather than the discrete Plancherel identity in subsequent section. This approach enables potential savings in the frequency sum estimation.

4.2. Reduction to multiplicative completed exponential sums. Let $\mathbb{B}$ denote the set of all 4-tuples satisfying

$$\mathbb{B} = \{\mathbf{b} = (b_1, b_2, b_3, b_4) : 1 \leq b_i \leq B\}.$$

For a prime p , define $\mathcal{V}^\Delta$ as the affine variety of 4-tuples

$$\mathbf{b} = (b_1, b_2, b_3, b_4) \in \mathbb{F}_p^4$$

characterized by the condition that for each index $i \in \{1, 2, 3, 4\}$, the cardinality

$$|\{j = 1, 2, 3, 4 \mid b_j = b_i\}|$$

must be even. We further define $\mathbb{B}_p^\Delta$ as the subset of integer tuples $\mathbf{b} \in \mathbb{B}$ satisfying

$$\mathbf{b} \pmod{p} \in \mathcal{V}^\Delta.$$

A direct computation yields the trivial bound

$$(4.5) \quad \Sigma^d(\mathbf{b}, AM; q) \ll A^2 M^2 q$$

for arbitrary $\mathbf{b} \in \mathbb{B}$ and divisors $d \mid q$.

To obtain refined estimates for $\Sigma^d(\mathbf{b}, AM; q)$, we employ distinct analytical approaches. The primary method involves evaluating the r -summation for fixed (l_1, l_2) pairs. For $\mathbf{l} = (l_1, l_2)$, we define the completed exponential sum

$$\mathfrak{R}(\mathbf{b}, \mathbf{l}; q) = \sum_{r \pmod{q}} \prod_{i=1}^2 \prod_{j=1}^4 \text{Kl}_2(l_i(r + b_j); q).$$

Through the twisted multiplicativity of Kloosterman sums and application of the Chinese remainder theorem, $\mathfrak{R}(\mathbf{b}, \mathbf{l}; q)$ inherits the following multiplicative property.

Lemma 4.1 (Multiplicative property). *Let $q = q_1 q_2$ with $(q_1, q_2) = 1$. Then for arbitrary $\mathbf{b}$ and $\mathbf{l}$, we have*

$$\mathfrak{R}(\mathbf{b}, \mathbf{l}; q) = \mathfrak{R}(\mathbf{b}, \bar{q}_2^2 \mathbf{l}; q_1) \mathfrak{R}(\mathbf{b}, \bar{q}_1^2 \mathbf{l}; q_2).$$

Proof. The twisted multiplicativity of Kloosterman sums implies the decomposition

$$\text{Kl}_2(l_i(r + b_j); q) = \text{Kl}_2(l_i(r + b_j) \bar{q}_2^2; q_1) \text{Kl}_2(l_i(r + b_j) \bar{q}_1^2; q_2).$$

The claimed identity then follows from the Chinese remainder theorem. $\square$

A direct computation yields the uniform trivial bound

$$(4.6) \quad \Re(\mathbf{b}, \mathbf{l}; q) \ll q,$$

valid for all $\mathbf{b}$ and $\mathbf{l}$.

Lemma 4.2. *For tuples $\mathbf{b} \notin \mathbb{B}_p^\Delta$ and $\mathbf{l} = (l_1, l_2)$ satisfying $l_1 \not\equiv l_2 \pmod{p}$, we have*

$$(4.7) \quad \Re(\mathbf{b}, \mathbf{l}; p) \ll p^{\frac{1}{2}}.$$

Proof. When $l_1 l_2 \not\equiv 0 \pmod{p}$, the result follows from [14, Lemma 2.5]. In the remaining cases where some l_i vanishes in $\mathbb{F}_p$, the Kloosterman sum reduces to a Ramanujan sum, such that

$$\text{Kl}_2(l_i(r + b_j); p) = -p^{-\frac{1}{2}},$$

The bound (4.7) then follows by estimating the remaining terms trivially. $\square$

The subsequent analysis requires more sophisticated techniques to extract savings from averaging over l_1 and l_2 . By applying the Poisson summation formula to transform this average into completed sums, we obtain

$$(4.8) \quad \Sigma^d(\mathbf{b}, AM; q) \ll \frac{(AM)^2}{q^2} \sum_{\mathbf{h}} |\mathfrak{S}(\mathbf{b}, \bar{\mathbf{c}}\mathbf{h}, d; q)|,$$

where the completed exponential sum is defined as

$$(4.9) \quad \mathfrak{S}(\mathbf{b}, \mathbf{h}, d; q) = \sum_{\substack{r, s_1, s_2 \pmod{q} \\ (s_1 - s_2, q) = d}} \sum_{i=1}^2 \sum_{j=1}^4 e_q(h_i s_i) \text{Kl}_2(s_i(r + b_j); q)$$

with frequency parameters $\mathbf{h} = (h_1, h_2)$ satisfying

$$h_i \ll H = \left(\frac{q}{AM} \right)^{1+\varepsilon}.$$

This completed exponential sum possesses the following multiplicative property.

Lemma 4.3 (Multiplicative property). *Let $q = q_1 q_2$ with $(q_1, q_2) = 1$ and $d = d_1 d_2$ where $d_i = (d, q_i)$. Then for arbitrary $\mathbf{b}$ and $\mathbf{h}$, we have*

$$\mathfrak{S}(\mathbf{b}, \mathbf{h}, d; q) = \mathfrak{S}(\mathbf{b}, q_2 \mathbf{h}, d_1; q_1) \mathfrak{S}(\mathbf{b}, q_1 \mathbf{h}, d_2; q_2).$$

Proof. By the twisted multiplicativity of Kloostermann sums, the reciprocity law of exponent functions and the Chinese remainder theorem, we have

$$\begin{aligned} \mathfrak{S}(\mathbf{b}, \mathbf{h}, d; q) &= \sum_{\substack{r_1, s_1, s_2 \pmod{q_1} \\ (s_1 - s_2, q_1) = d_1}} \sum_{i=1}^2 \sum_{j=1}^4 e_{q_1}(h_i s_i \bar{q}_2) \text{Kl}_2(s_i(r_1 + b_j) \bar{q}_2^2; q_1) \\ &\quad \times \sum_{\substack{r_2, t_1, t_2 \pmod{q_2} \\ (t_1 - t_2, q_2) = d_2}} \sum_{i=1}^2 \sum_{j=1}^4 e_{q_2}(h_i t_i \bar{q}_1) \text{Kl}_2(t_i(r_2 + b_j) \bar{q}_1^2; q_2). \end{aligned}$$

After making the variable substitutions

$$s_i \bar{q}_2^2 \rightarrow s_i, \quad t_i \bar{q}_1^2 \rightarrow t_i,$$

we obtain the multiplicative relation. $\square$

For $d \mid \rho$, Lemma 4.3 yields the factorization

$$(4.10) \quad \Xi(\mathbf{b}, \mathbf{h}, d; q) = \Xi(\mathbf{b}, q^* \mathbf{h}, d; \rho) \Xi(\mathbf{b}, \rho \mathbf{h}, 1; q^*).$$

The first component admits the trivial bound

$$(4.11) \quad \Xi(\mathbf{b}, q^* \mathbf{h}, d; \rho) \ll \rho^3,$$

while the evaluation of $\Xi(\mathbf{b}, \mathbf{h}, 1; q^*)$ reduces, via the multiplicative relation, to analyzing the prime-power cases $\Xi(\mathbf{b}, \mathbf{h}, 1; p)$ and $\Xi(\mathbf{b}, \mathbf{h}, 1; p^2)$.

Lemma 4.4. *For $\mathbf{b} \notin \mathbb{B}_p^\Delta$, we have*

$$(4.12) \quad \Xi(\mathbf{b}, \mathbf{h}, 1; p) \ll p^{\frac{5}{2}}.$$

Proof. We begin by observing the preliminary bound

$$\Xi(\mathbf{b}, \mathbf{h}, 1; p) \ll \sum_{\substack{s_1, s_2 \pmod{p} \\ (s_1 - s_2, p) = 1}} |\Re(\mathbf{b}, \mathbf{s}; p)|,$$

where $\mathbf{s} = (s_1, s_2)$. For each fixed pair $\mathbf{s}$, Lemma 4.2 establishes $\Re(\mathbf{b}, \mathbf{s}; p) \ll p^{\frac{1}{2}}$ under the hypothesis $\mathbf{b} \notin \mathbb{B}_p^\Delta$. Then the trivial estimate for the double summation over s_1, s_2 gives the required bound. $\square$

Lemma 4.5. *For sufficiently large prime p , there exists a codimension 1 subvariety $\mathcal{V}^{bad} \subset \mathbb{F}_p^4$ containing $\mathcal{V}^\Delta$, with degree bounded independently of p , such that for all $\mathbf{b}$ satisfying $\mathbf{b} \pmod{p} \notin \mathcal{V}^{bad}$ and arbitrary $\mathbf{h}$, we have*

$$(4.13) \quad \Xi(\mathbf{b}, \mathbf{h}, 1; p) \ll p^{\frac{3}{2}}.$$

Proof. If we restrict the sum in the definition (4.9) to the terms with $s_1 s_2 \neq 0$, then the bound follows from [14, Theorem 2.6], combining with the argument just before it. For the remaining terms, in which one of $s_i = 0$, the corresponding four Kloosterman sums reduce to

$$\text{Kl}_2(s_i(r + b_j); p) = -p^{-\frac{1}{2}},$$

which shows that their contribution is at most $O(p)$. $\square$

Theorem 4.6. *For $\mathbf{b} \notin \mathbb{B}_p^\Delta$, we have*

$$(4.14) \quad \Xi(\mathbf{b}, \mathbf{h}, 1; p^2) \ll \begin{cases} p^4 & \text{if } \mathbf{h} \pmod{p} \equiv (0, 0), \\ p^4 & \text{if } (\mathbf{b}, \mathbf{h}) \in \mathcal{V}_4^{bad} \times \mathcal{V}_2^{bad}(\mathbf{b}), \\ p^3 & \text{otherwise.} \end{cases}$$

Here $\mathcal{V}_4^{bad} \subset \mathbb{F}_p^4$ is a variety defined by a homogeneous polynomial of bounded degree, while for any given $\mathbf{b}$, the variety $\mathcal{V}_2^{bad}(\mathbf{b}) \subset \mathbb{F}_p^2$ is defined by $O(1)$ linear homogeneous polynomials.

Remark 4.2. Note that bounds in (4.11) – (4.13) are independent of the parameter vector $\mathbf{h}$, whereas the estimates in (4.14) exhibit homogeneous dependence on $\mathbf{h}$. Consequently, these bounds remain valid under the substitution $\mathbf{h} \rightarrow c\mathbf{h}$ for any integer c satisfying $(c, p) = 1$. This substitution principle is particularly significant when applying the multiplicative decomposition from Lemma 4.3, as it ensures the preservation of estimate quality under scaling transformations of the frequency parameters. The uniform bounds thus obtained are essential for handling the exponential sums that arise in the subsequent analytic arguments.

We defer the proof of Theorem 4.6 to the following section and proceed to establish Theorem 2.2 assuming the validity of Theorem 4.6. To begin with, we define the required notation. For large prime p , we define the exceptional set

$$\mathbb{B}_p^{bad} = \{\mathbf{b} \in \mathbb{B} \mid \mathbf{b} \pmod{p} \in \mathcal{V}^{bad}\} \setminus \mathbb{B}_p^\Delta,$$

where $\mathcal{V}^{bad}$ is as defined in Lemma 4.5. This allows us to partition the parameter space into three disjoint subsets such that

$$\mathbb{B} = \mathbb{B}_p^\Delta \cup \mathbb{B}_p^{bad} \cup \mathbb{B}_p^{gen}.$$

Since $\mathcal{V}^{bad}$ has a finite degree, independently of p , we have

$$|\mathbb{B}_p^\Delta| \ll B^2 \left(1 + \frac{B^2}{p^2}\right), \quad |\mathbb{B}_p^{bad}| \ll B^3 \left(1 + \frac{B}{p}\right), \quad |\mathbb{B}_p^{gen}| \ll B^4.$$

The proof of Theorem 2.2 will be carried out in three cases depending on the factorization of $q^\star$. In all cases, we employ the following parameter choices

$$(4.15) \quad A = M^{-\frac{1}{2}} N^{\frac{1}{2}} q^{\frac{1}{8}} \rho^{\frac{3}{8}}, \quad B = M^{\frac{1}{2}} N^{\frac{1}{2}} q^{-\frac{1}{8}} \rho^{-\frac{3}{8}}.$$

For $MN \leq q^{\frac{5}{4}} \rho^{-\frac{1}{4}}$, it is easy to check the inequalities

$$(4.16) \quad AM < q, \quad AB = N, \quad B \leq q^{\frac{1}{2}} \rho^{-\frac{1}{2}} = (q^\star)^{\frac{1}{2}},$$

which we present here for easy of reference.

4.2.1. *Proof of Theorem 2.2 for $q^\star = p$.* Through the multiplicative property (Lemma 4.1), we derive from (4.4) that

$$(4.17) \quad \Sigma^d(\mathbf{b}, AM; q) \ll (AM)^2 |\Re(\mathbf{b}, \bar{p}^2 \mathbf{l}; \rho) \Re(\mathbf{b}, \bar{p}^2 \mathbf{l}; p)|.$$

For $\mathbf{b} \notin \mathbb{B}_p^\Delta$, we apply the trivial bound (4.6) to the first exponential sum and (4.7) to the second one, obtaining

$$(4.18) \quad \Sigma^d(\mathbf{b}, AM; q) \ll (AM)^2 (q\rho)^{\frac{1}{2}}.$$

On the other hand, applying the multiplicative property (Lemma 4.3) into (4.8), we arrive at

$$(4.19) \quad \Sigma^d(\mathbf{b}, AM; q) \ll \frac{(AM)^2}{q^2} \sum_{\mathbf{h}} |\Im(\mathbf{b}, p\mathbf{h}, d; \rho) \Im(\mathbf{b}, \rho\mathbf{h}, 1; p)|.$$

For $\mathbf{b} \in \mathbb{B}_p^{gen}$, we bound the first exponential sum via (4.11) and the second one via (4.13), getting

$$(4.20) \quad \Sigma^d(\mathbf{b}, AM; q) \ll q^\varepsilon (q\rho)^{\frac{3}{2}}.$$

To combine all cases, we apply the trivial bound (4.5) for $\mathbf{b} \in \mathbb{B}_p^\Delta$, the bound (4.18) for $\mathbf{b} \in \mathbb{B}_p^{bad}$ and finally the bound (4.20) for $\mathbf{b} \in \mathbb{B}_p^{gen}$. For $p = q^\star \geq B$ (see (4.16)), this gives

$$(4.21) \quad \sum_b |\Sigma^d(\mathbf{b}, AM; q)| \ll q^\varepsilon \left(A^2 B^2 M^2 q + A^2 B^3 M^2 (q\rho)^{\frac{1}{2}} + B^4 (q\rho)^{\frac{3}{2}} \right).$$

By using parameters A, B from (4.15) under $MN \leq q^{\frac{5}{4}} \rho^{-\frac{1}{4}}$, the first and third terms on the right-hand side of (4.21) are equal to $(MN)^2 q$, and the second term is

$$(MN)^{\frac{5}{2}} q^{\frac{3}{8}} \rho^{\frac{1}{8}} \ll (MN)^2 q.$$

We conclude

$$\sum_b |\Sigma^d(\mathbf{b}, AM; q)| \ll (MN)^2 q^{1+\varepsilon},$$

and consequently by (4.3)

$$(4.22) \quad \begin{aligned} \mathcal{T}_d(\alpha, 1_N) &\ll \frac{q^\varepsilon}{N} (AN)^{\frac{3}{4}} M^{\frac{1}{2}} \|\alpha\|_2^2 (MN)^{\frac{1}{2}} q^{\frac{1}{4}} \\ &= \|\alpha\|_2^2 (MN)^{\frac{5}{8}} q^{\frac{11}{32} + \varepsilon} \rho^{\frac{9}{32}}. \end{aligned}$$

Substituting this into (4.1) with the trivial bound (4.2) for $p \mid d$, we arrive at

$$\sum_{N \leq n \leq 2N} \left| \sum_{M \leq m \leq 2M} \alpha_m \text{Kl}_2(cmn; q) \right|^2 \ll q^\varepsilon \|\alpha\|_2^2 MN \left(M^{-1} + p^{-1} + (MN)^{-\frac{3}{8}} q^{\frac{11}{32}} \rho^{\frac{9}{32}} \right),$$

which establishes (2.4) through (2.8).

4.2.2. *Proof of Theorem 2.2 for $q^\star = p_1 p_2$, $p_1 < p_2$.* In this case, we decompose $\mathbb{B}$ into seven disjoint subsets such that

$$\mathbb{B} = \mathbb{B}^\Delta \cup \mathbb{B}^{bad} \cup \mathbb{B}^{gen} \cup \mathbb{B}_1 \cup \mathbb{B}_2 \cup \mathbb{B}'_1 \cup \mathbb{B}'_2,$$

where the components are defined as

$$\mathbb{B}^\Delta = \mathbb{B}_{p_1}^\Delta \cap \mathbb{B}_{p_2}^\Delta, \quad \mathbb{B}^{bad} = \mathbb{B}_{p_1}^{bad} \cap \mathbb{B}_{p_2}^{bad}, \quad \mathbb{B}^{gen} = \mathbb{B}_{p_1}^{gen} \cap \mathbb{B}_{p_2}^{gen},$$

and

$$\mathbb{B}_1 = \mathbb{B}_{p_1}^\Delta \setminus \mathbb{B}_{p_2}^\Delta, \quad \mathbb{B}_2 = \mathbb{B}_{p_2}^\Delta \setminus \mathbb{B}_{p_1}^\Delta, \quad \mathbb{B}'_1 = (\mathbb{B}_{p_1}^{bad} \cap \mathbb{B}_{p_2}^{gen}), \quad \mathbb{B}'_2 = (\mathbb{B}_{p_2}^{bad} \cap \mathbb{B}_{p_1}^{gen}).$$

Note $\mathbb{B}_j = \emptyset$ when $p_j > B$. For $p_1 p_2 = q^\star \geq B$, it is easy to see

$$|\mathbb{B}^\Delta| \ll B^2, \quad |\mathbb{B}^{bad}| \ll B^3 \left(1 + \frac{B}{p_2} \right), \quad |\mathbb{B}^{gen}| \ll B^4,$$

and

$$|\mathbb{B}_j| \ll \frac{B^4}{p_j^2}, \quad |\mathbb{B}'_j| \ll B^3 \left(1 + \frac{B}{p_j} \right).$$

Applying Lemmas 4.1 & 4.3 yields the analogues of (4.17) and (4.19) such that

$$(4.23) \quad \Sigma^d(\mathbf{b}, AM; q) \ll (AM)^2 \rho \left| \Re(\mathbf{b}, (\overline{\rho p_2})^2 \mathbf{l}; p_1) \Re(\mathbf{b}, (\overline{\rho p_1})^2 \mathbf{l}; p_2) \right|,$$

$$(4.24) \quad \Sigma^d(\mathbf{b}, AM; q) \ll \frac{(AM)^2}{q^2} \rho^3 \sum_h |\Im(\mathbf{b}, p_2 \rho \mathbf{h}, d; p_1) \Im(\mathbf{b}, p_1 \rho \mathbf{h}, 1; p_2)|.$$

For $\mathbf{b} \in \mathbb{B}^{bad}$, applying (4.7) to both exponential sums in (4.23) yields the estimate (4.18).

For $\mathbf{b} \in \mathbb{B}^{gen}$, applying (4.13) to both exponential sums in (4.24) reproduces the estimate (4.20).

For $\mathbf{b} \in \mathbb{B}_j$, we employ the trivial bound for the exponential sum modulo p_j in (4.23) and (4.7) for the other sum, which gives

$$(4.25) \quad \Sigma^d(\mathbf{b}, AM; q) \ll (AM)^2 (q\rho)^{\frac{1}{2}} p_j^{\frac{1}{2}}.$$

For $\mathbf{b} \in \mathbb{B}'_j$, we apply (4.12) to the exponential sum modulo p_j in (4.24) and (4.13) to the other sum, which shows

$$\Sigma^d(\mathbf{b}, AM; q) \ll q^\varepsilon (q\rho)^{\frac{3}{2}} p_j.$$

Combining this with (4.18) yields

$$(4.26) \quad \Sigma^d(\mathbf{b}, AM; q) \ll q^\varepsilon \min \left\{ (AM)^2 (q\rho)^{\frac{1}{2}}, (q\rho)^{\frac{3}{2}} p_j \right\}.$$

To consolidate all cases, we apply the trivial bound (4.5) for $\mathbf{b} \in \mathbb{B}^\Delta$, the bound (4.18) for $\mathbf{b} \in \mathbb{B}^{bad}$, the bound (4.20) for $\mathbf{b} \in \mathbb{B}^{gen}$, the bound (4.25) for $\mathbf{b} \in \mathbb{B}_j$, and finally the bound (4.26) for $\mathbf{b} \in \mathbb{B}'_j$. This yields the comprehensive estimate

$$\begin{aligned} \sum_b |\Sigma^d(\mathbf{b}, AM; q)| &\ll q^\varepsilon \left\{ A^2 B^2 M^2 q + A^2 B^3 M^2 (q\rho)^{\frac{1}{2}} \left(1 + \frac{B}{p_2} \right) + B^4 (q\rho)^{\frac{3}{2}} \right. \\ &\quad \left. + \sum_{j=1,2} A^2 B^4 M^2 (q\rho)^{\frac{1}{2}} p_j^{-\frac{3}{2}} + \sum_{j=1,2} \min \left\{ A^2 B^3 M^2 (q\rho)^{\frac{1}{2}}, B^3 (q\rho)^{\frac{3}{2}} p_j \right\} \left(1 + \frac{B}{p_j} \right) \right\}. \end{aligned}$$

For $p_2 > (q/\rho)^{\frac{1}{2}} > p_1 > (q/\rho)^{\frac{1}{3}}$, it is easy to see that

$$A^2 B^3 M^2 (q\rho)^{\frac{1}{2}} \frac{B}{p_2} + \sum_{j=1,2} A^2 B^4 M^2 (q\rho)^{\frac{1}{2}} p_j^{-\frac{3}{2}} \ll A^2 B^4 M^2 \rho,$$

$$\sum_{j=1,2} \min \left\{ A^2 B^3 M^2 (q\rho)^{\frac{1}{2}}, B^3 (q\rho)^{\frac{3}{2}} p_j \right\} \left(1 + \frac{B}{p_j} \right) \ll A^2 B^3 M^2 (q\rho)^{\frac{1}{2}} + B^4 (q\rho)^{\frac{3}{2}}.$$

Consequently, we conclude that

$$(4.27) \quad \sum_b |\Sigma^d(\mathbf{b}, AM; q)| \ll q^\varepsilon \left\{ A^2 B^2 M^2 q + A^2 B^3 M^2 (q\rho)^{\frac{1}{2}} + B^4 (q\rho)^{\frac{3}{2}} + A^2 B^4 M^2 \rho \right\},$$

where all terms except the last also appear in (4.21). With A, B from (4.15) and $MN \ll q^{\frac{5}{4}} \rho^{-\frac{1}{4}}$, the last term satisfies

$$(MN)^3 q^{-\frac{1}{4}} \rho^{\frac{1}{4}} \ll (MN)^2 q.$$

Now we also arrive at

$$\sum_b |\Sigma^d(\mathbf{b}, AM; q)| \ll (MN)^2 q^{1+\varepsilon},$$

reconfirming (2.4) through identical arguments as before.

4.2.3. *Proof of Theorem 2.2 for $q^\star = p^2$.* Adopting notations from Theorem 4.6, we define

$$\mathbb{B}^{bad} = \{\mathbf{b} \in \mathbb{B} \mid \mathbf{b} \pmod{p} \in \mathcal{V}_4^{bad} \setminus \mathbb{B}_p^\Delta, \quad \mathbb{H}^{bad} = \{\mathbf{h} \in \mathbb{H} \mid \mathbf{h} \pmod{p} \in \mathcal{V}_2^{bad}(\mathbf{b})\}.$$

Given the finite degrees of $\mathcal{V}_2^{bad}(\mathbf{b})$ and $\mathcal{V}_4^{bad}$, we have

$$|\mathbb{B}^{bad}| \ll B^3 \left(1 + \frac{B}{p}\right), \quad |\mathbb{H}^{bad}| \ll H \left(1 + \frac{H}{p}\right).$$

Applying Lemma 4.3's multiplicative relation to (4.8) yields

$$\Sigma^d(\mathbf{b}, AM; q) \ll \frac{(AM)^2}{q^2} \rho^3 \sum_{\mathbf{h}} |\mathfrak{S}(\mathbf{b}, \rho \bar{c} \mathbf{h}, 1; p^2)|.$$

Bounding the exponential sum via Theorem 4.6 gives

$$\begin{aligned} \sum_{\mathbf{b} \in \mathbb{B} \setminus \mathbb{B}_p^\Delta} \Sigma^d(\mathbf{b}, AM; q) &\ll \frac{(AM)^2}{q^2} \rho^3 \left(B^4 p^4 \left(1 + \frac{H^2}{p^2}\right) + B^3 H p^4 \left(1 + \frac{B}{p}\right) \left(1 + \frac{H}{p}\right) + B^4 H^2 p^3 \right) \\ &\ll \frac{(AM)^2}{q^2} \rho^3 (B^4 p^4 + B^3 H p^4 + B^4 H^2 p^3) \\ &\ll q^\varepsilon (A^2 B^4 M^2 \rho + AB^3 M q \rho + B^4 (q \rho)^{\frac{3}{2}}). \end{aligned}$$

Combining this with the trivial bound (4.5) for $\mathbf{b} \in \mathbb{B}_p^\Delta$ produces

$$\begin{aligned} (4.28) \quad \sum_b |\Sigma^d(\mathbf{b}, AM; q)| &\ll q^\varepsilon \left(A^2 B^2 M^2 q \left(1 + \frac{B^2}{p^2}\right) + A^2 B^4 M^2 \rho + AB^3 M q \rho + B^4 (q \rho)^{\frac{3}{2}} \right) \\ &\ll q^\varepsilon (A^2 B^2 M^2 q + A^2 B^4 M^2 \rho + B^4 (q \rho)^{\frac{3}{2}}), \end{aligned}$$

where $AB^3 M q \rho$ is absorbed by the first and last terms. Noting that all terms in (4.28) appear in (4.27), we also have

$$\sum_b |\Sigma^d(\mathbf{b}, AM; q)| \ll (MN)^2 q^{1+\varepsilon},$$

available for A, B in (4.15). This establishes (2.4), concluding the proof of Theorem 2.2.

5. PROOF OF THEOREM 4.6

To establish an optimal bound for $\mathfrak{S}(\mathbf{b}, \mathbf{h}, 1; p^2)$, we require point counting on a (generically zero-dimensional) variety in $\mathbb{F}_p^{11}$. For given $\mathbf{b} \in \mathbb{F}_p^4$ and $\mathbf{h} \in \mathbb{F}_p^2$, we define $\mathcal{V}_{11}$ as the variety of 11-tuples

$$(5.1) \quad \mathbf{v} = (u, x_1, x_2, x_{11}, \dots, x_{14}, x_{21}, \dots, x_{24}) \in \mathbb{F}_p^{11}$$

satisfying the following eleven equations

$$(5.2a) \quad x_i x_{ij} = (u + b_j) \bar{x}_{ij}, \text{ for } 1 \leq i \leq 2, 1 \leq j \leq 4;$$

$$(5.2b) \quad \sum_{j=1}^4 x_{ij} + h_i = 0, \text{ for } 1 \leq i \leq 2;$$

$$(5.2c) \quad \sum_{i=1}^2 \sum_{j=1}^4 \bar{x}_{ij} = 0.$$

Let $\mathcal{K}(\mathbf{b}, \mathbf{h}; p)$ denote the number of $\mathbf{v} \in \mathcal{V}_{11}$ with

$$(5.3) \quad x_1 \neq x_2, \quad x_i, x_{ij}, u + b_j \in \mathbb{F}_p^\times$$

for $1 \leq i \leq 2, 1 \leq j \leq 4$.

Since $\mathcal{V}_{11}$ is defined by eleven polynomials in $\mathbb{F}_p^{11}$, it generically constitutes a zero-dimensional variety. Thus, we expect $\mathcal{K}(\mathbf{b}, \mathbf{h}; p) \ll 1$ to hold except in degenerate cases. We make this precise in the following key lemma.

Lemma 5.1. *For $\mathbf{b} \in \mathbb{F}_p^4$ and $\mathbf{h} \in \mathbb{F}_p^2$ with $\mathbf{b} \notin \mathcal{V}^\Delta$, we have*

$$(5.4) \quad \mathcal{K}(\mathbf{b}, \mathbf{h}; p) \ll \begin{cases} p & \text{if } \mathbf{h} = (0, 0), \\ p & \text{if } (\mathbf{b}, \mathbf{h}) \in \mathcal{V}_4^{bad} \times \mathcal{V}_2^{bad}(\mathbf{b}), \\ 1 & \text{otherwise.} \end{cases}$$

Here $\mathcal{V}_4^{bad} \subset \mathbb{F}_p^4$ is a variety defined by a homogeneous polynomial of bounded degree, while for any given $\mathbf{b}$, the variety $\mathcal{V}_2^{bad}(\mathbf{b}) \subset \mathbb{F}_p^2$ is defined by $O(1)$ linear homogeneous polynomials.

We now proceed to establish Theorem 4.6 conditional on Lemma 5.1, deferring the proof of Lemma 5.1 to the subsequent section.

Proof of Theorem 4.6. Note that $\text{Kl}_2(s_i(r + b_j); p^2)$ vanishes unless $(s_i(r + b_j), p) = 1$, a convention maintained through this proof. Recall the representation

$$\text{Kl}_2(s_i(r + b_j); p^2) = \frac{1}{p} \sum_{x \pmod{p^2}}^* e_{p^2}(s_i x + (r + b_j) \bar{x}).$$

Through the parameterization

$$s_i = x_i(1 + y_i p), \quad r = u + v p, \quad x = x_{ij}(1 + y_{ij} p),$$

where $x_{ij}, y_{ij}, x_i, y_i, u, v$ are defined modulo p with $(x_i(u + b_j)x_{ij}, p) = 1$, we have

$$\begin{aligned} \text{Kl}_2(s_i(r + b_j); p^2) &= \frac{1}{p} \sum_{x_{ij} \pmod{p}}^* e_{p^2}(x_i x_{ij} + (u + b_j) \bar{x}_{ij}) e_p(x_i y_i x_{ij} + v \bar{x}_{ij}) \\ &\quad \times \sum_{y_{ij} \pmod{p}} e_p(y_{ij}(x_i x_{ij} - (u + b_j) \bar{x}_{ij})). \end{aligned}$$

The y_{ij} summation vanishes unless $p \mid (x_i x_{ij} - (u + b_j) \bar{x}_{ij})$, reducing to p when nonvanishing. Consequently,

$$\text{Kl}_2(s_i(r + b_j); p^2) = \sum_{\substack{x_{ij} \pmod{p} \\ x_i x_{ij} \equiv (u + b_j) \bar{x}_{ij} \pmod{p}}}^* e_{p^2}(x_i x_{ij} + (u + b_j) \bar{x}_{ij}) e_p(x_i y_i x_{ij} + v \bar{x}_{ij}).$$

Substituting this and the identity

$$e_{p^2}(h_i s_i) = e_{p^2}(h_i x_i) e_p(h_i x_i y_i)$$

into $\mathfrak{S}(\mathbf{b}, \mathbf{h}, 1; p^2)$ shows

$$(5.5) \quad \begin{aligned} \mathfrak{S}(\mathbf{b}, \mathbf{h}, 1; p^2) = & \sum_{u \pmod{p}} \sum_{\substack{x_i, x_{ij}, (i=1,2; j=1,2,3,4) \\ x_i x_{ij} \equiv (u + b_j) \bar{x}_{ij} \pmod{p} \\ x_1 \neq x_2}}^* \cdots \sum_{x_1 \neq x_2}^* e_{p^2} \left(\sum_{i=1}^2 \left(h_i x_i + \sum_{j=1}^4 (x_i x_{ij} + (u + b_j) \bar{x}_{ij}) \right) \right) \\ & \sum_{y_1, y_2 \pmod{p}} e_p \left(\sum_{i=1}^2 x_i y_i \left(\sum_{j=1}^4 x_{ij} + h_i \right) \right) \sum_{v \pmod{p}} e_p \left(v \sum_{i=1}^2 \sum_{j=1}^4 \bar{x}_{ij} \right), \end{aligned}$$

where the product of the sums over y_i and v vanish unless

$$\sum_{j=1}^4 x_{ij} + h_i \equiv 0 \pmod{p} \quad \text{and} \quad \sum_{i=1}^2 \sum_{j=1}^4 \bar{x}_{ij} \equiv 0 \pmod{p}.$$

Applying this to (5.5) yields

$$\mathfrak{S}(\mathbf{b}, \mathbf{h}, 1; p^2) \ll p^3 \mathcal{K}(\mathbf{b}, \mathbf{h}; p),$$

which, combined with Lemma 5.1, establishes Theorem 4.6. $\square$

6. COUNTING RATIONAL POINTS ON THE VARIETY

This section is devoted to the proof of Lemma 5.1. Our strategy is to reduce, in several steps, the problem of counting the vectors $\mathbf{v}$ on the variety $\mathcal{V}_{11}$ (subject to the additional conditions (5.3)), first to the problem of counting the corresponding triples (u, x_{11}, x_{21}) , and then just to counting the corresponding values of u . We analyze the relations for u through polynomial substitution, which generically admits $O(1)$ solutions. Degeneracy occurs precisely when both $\mathbf{b}$ and $\mathbf{h}$ lie on the varieties defined by homogeneous polynomials of bounded degree as indicated in the statement of Lemma 5.1.

We remark that the statement of Lemma 5.1 is trivially true for primes p of bounded size, and that therefore, throughout this section, we may (and will) assume that p is larger than a certain absolute lower bound, so that various absolute positive integer constants appearing in our arguments will be units modulo p .

6.1. Reduction to solutions of polynomials. From (5.2a), we establish the relations

$$(6.1) \quad \frac{x_{ij}}{x_{i1}} = \pm \left(\frac{u + b_j}{u + b_1} \right)^{\frac{1}{2}} \quad \text{and} \quad \frac{x_{2j}}{x_{1j}} = \pm \left(\frac{x_1}{x_2} \right)^{\frac{1}{2}}$$

for $1 \leq i \leq 2, 1 \leq j \leq 4$, where we fix an arbitrary branch of the square root from $\mathbb{F}_p^{\times 2} \rightarrow \mathbb{F}_p^{\times}$, and merely using the square root notation indicates that the argument lies

in $\mathbb{F}_p^{\times 2}$ (and thus the square root is defined) without additional notice. Substituting these relations into (5.2b) and (5.2c) produces the system

$$(6.2a) \quad x_{11}f_1(u) + h_1 = 0;$$

$$(6.2b) \quad x_{21}f_2(u) + h_2 = 0;$$

$$(6.2c) \quad \bar{x}_{11}g_1(u) + \bar{x}_{21}g_2(u) = 0,$$

where the functions are defined as

$$(6.3) \quad f_i(u) = Q_i\left(1, \left(\frac{u+b_2}{u+b_1}\right)^{\frac{1}{2}}, \left(\frac{u+b_3}{u+b_1}\right)^{\frac{1}{2}}, \left(\frac{u+b_4}{u+b_1}\right)^{\frac{1}{2}}\right),$$

$$(6.4) \quad g_i(u) = Q_i\left(1, \left(\frac{u+b_1}{u+b_2}\right)^{\frac{1}{2}}, \left(\frac{u+b_1}{u+b_3}\right)^{\frac{1}{2}}, \left(\frac{u+b_1}{u+b_4}\right)^{\frac{1}{2}}\right)$$

with $Q_i(x, y, z, w) \in \{x \pm y \pm z \pm w\}$.

Let $\mathcal{K}_1(\mathbf{b}, \mathbf{h}; p)$ denote the number of triples $(u, x_{11}, x_{21}) \in \mathbb{F}_p^3$ satisfying one of the $O(1)$ instances of the system (6.2a)–(6.2c), with f_i, g_i as in (6.3)–(6.4). We have proved that, for every tuple $\mathbf{v} \in \mathcal{V}_{11}$ counted by $\mathcal{K}(\mathbf{b}, \mathbf{h}; p)$ and labeled as in (5.1), the triple (u, x_{11}, x_{21}) is among the triples counted by $\mathcal{K}_1(\mathbf{b}, \mathbf{h}; p)$. On the other hand, the relations (5.2a) and (6.1) demonstrate that, for given parameters $\mathbf{h}, \mathbf{b}$, all remaining variables in $\mathbf{v}$ are determined by the triple (u, x_{11}, x_{21}) up to $O(1)$ choices. From this it follows that

$$\mathcal{K}(\mathbf{b}, \mathbf{h}; p) \asymp \mathcal{K}_1(\mathbf{b}, \mathbf{h}; p),$$

and it suffices to bound the latter count.

Moreover, for a given u , the value of x_{11} and x_{21} can be resolved from (6.2a)–(6.2c), except in the degenerate case where at least two of $f_1(u), f_2(u), g_1(u)$, and $g_2(u)$ vanish simultaneously. The following subsection §6.2 is devoted to determining u for which this happens, with analysis proceeding in two cases. When either $h_1 = 0$ or $h_2 = 0$, u will be determined to within $O(1)$ choices by the corresponding $f_i(u) = 0$. For non-zero h_1 and h_2 , substituting (6.2a) and (6.2b) into (6.2c) yields

$$\bar{h}_1 f_1(u) g_1(u) + \bar{h}_2 f_2(u) g_2(u) = 0,$$

which will similarly determine u up to $O(1)$ choices outside degenerate cases. We put everything together and prove Lemma 5.1 in §6.3.

6.2. Lemmas for counting solutions of polynomials.

Lemma 6.1. *For any rational function $Q(x, y, z, w)$ in four variables, the product*

$$F(x, y, z, w) = \prod Q(\pm x^{\frac{1}{2}}, \pm y^{\frac{1}{2}}, \pm z^{\frac{1}{2}}, \pm w^{\frac{1}{2}})$$

taken over all sign combinations remains a rational function of x, y, z, w .

Proof. We first establish the polynomial case. Consider the auxiliary polynomial

$$P(x, y, z, w) = \prod Q(\pm x, \pm y, \pm z, \pm w).$$

The key observation is the invariance property

$$P(\epsilon_1 x, \epsilon_2 y, \epsilon_3 z, \epsilon_4 w) = P(x, y, z, w) \quad \text{for any } \epsilon_i = \pm 1.$$

This symmetry implies that all monomials in $P(x, y, z, w)$ must contain even powers of each variable; indeed, this is clear from $P(x, y, z, w) = (1/16) \sum P(\pm x, \pm y, \pm z, \pm w)$. Therefore, the substitution $x \rightarrow x^{\frac{1}{2}}, y \rightarrow y^{\frac{1}{2}}, z \rightarrow z^{\frac{1}{2}}, w \rightarrow w^{\frac{1}{2}}$ yields a well-defined polynomial

$$F(x, y, z, w) = P(x^{\frac{1}{2}}, y^{\frac{1}{2}}, z^{\frac{1}{2}}, w^{\frac{1}{2}}).$$

The rational case follows by considering numerator and denominator polynomial separately. $\square$

Lemma 6.2. *For any given $\mathbf{b} \notin \mathcal{V}^\Delta$, both each one of the congruence equations*

$$(6.5) \quad f_i(u) \equiv 0 \pmod{p},$$

$$(6.6) \quad g_i(u) \equiv 0 \pmod{p}$$

possess $O(1)$ solutions.

Proof. Applying Lemma 6.1, we eliminate all half-powers through the symmetric polynomial construction

$$\begin{aligned} F(x, y, z, w) &= \prod Q_i(\pm x^{\frac{1}{2}}, \pm y^{\frac{1}{2}}, \pm z^{\frac{1}{2}}, \pm w^{\frac{1}{2}}) \\ &= 64xyzw - \left((x + y - z - w)^2 - 4(xy + zw) \right)^2. \end{aligned}$$

For equation (6.5), solutions must satisfy

$$F(u + b_1, u + b_2, u + b_3, u + b_4) = 0,$$

which, by a straightforward calculation, simplifies to the linear equation¹

$$c_1 u + c_0 = 0$$

with the coefficient

$$c_1 = 8(b_1 + b_2 - b_3 - b_4)(b_1 - b_2 + b_3 - b_4)(b_1 - b_2 - b_3 + b_4)$$

and c_0 a certain explicit degree 6 form in b_i . We claim that, for $\mathbf{b} \notin \mathcal{V}^\Delta$, we cannot have $c_1 = 0$ and $c_0 = 0$ simultaneously. The symmetry of $F(u + b_1, u + b_2, u + b_3, u + b_4)$ in b_i produces the symmetry for the coefficients c_0 and c_1 . Without loss of generality, we apply $b_1 + b_2 - b_3 - b_4 = 0$ in place of $c_1 = 0$ to the expression of c_0 , and it follows that

$$c_0 = -16(b_1 - b_3)^2(b_2 - b_3)^2 \neq 0$$

unless $\mathbf{b} \in \mathcal{V}^\Delta$. Thus the two coefficients cannot vanish simultaneously, yielding at most one solution for u .

For (6.6), the transformed equation

$$(6.7) \quad F\left((u + b_1)^{-1}, (u + b_2)^{-1}, (u + b_3)^{-1}, (u + b_4)^{-1}\right) = 0$$

reduces to a ninth-degree equation

$$c_9 u^9 + c_8 u^8 + \cdots = 0,$$

where the leading coefficient

$$c_9 = -8(b_1 + b_2 - b_3 - b_4)(b_1 - b_2 + b_3 - b_4)(b_1 - b_2 - b_3 + b_4).$$

¹This equation as well as other equations in the next lemma can be calculated by math softwares such as *Mathematica* easily.

Applying $b_1 + b_2 - b_3 - b_4 = 0$ to the expression of c_8 gives

$$c_8 = 48(b_1 - b_3)^2(b_2 - b_3)^2 \neq 0$$

for $\mathbf{b} \notin \mathcal{V}^\Delta$. The fact that c_9 and c_8 cannot vanish simultaneously yields at most nine solutions for u . □

Lemma 6.3. *Given the function definitions f_i, g_i from (6.3) and (6.4) with distinct $Q_1 \neq Q_2$, for $\mathbf{b} \notin \mathcal{V}^\Delta$, the equation*

$$(6.8) \quad \bar{h}_1 f_1(u) g_1(u) + \bar{h}_2 f_2(u) g_2(u) = 0$$

admits $O(1)$ solutions except when $(\mathbf{b}, \mathbf{h}) \in \mathcal{V}_4^{bad} \times \mathcal{V}_2^{bad}(\mathbf{b})$. Here $\mathcal{V}_4^{bad} \subset \mathbb{F}_p^4$ denotes an algebraic variety determined by a homogeneous polynomial of bounded degree, and $\mathcal{V}_2^{bad}(\mathbf{b}) \subset \mathbb{F}_p^2$ represents a variety specified by $O(1)$ linear homogeneous polynomials for given $\mathbf{b}$.

Proof. Building upon Lemma 6.2, we focus our analysis on the non-degenerate case where $f_1(u)f_2(u)g_1(u)g_2(u) \neq 0$. To eliminate half-powers terms, we apply Lemma 6.3 by constructing the symmetric polynomial

$$F(x, y, z, w) = \prod Q(\pm x^{\frac{1}{2}}, \pm y^{\frac{1}{2}}, \pm z^{\frac{1}{2}}, \pm w^{\frac{1}{2}}),$$

where the rational function

$$Q(x, y, z, w) = \bar{h}_1 Q_1(x, y, z, w) Q_1(x^{-1}, y^{-1}, z^{-1}, w^{-1}) + \bar{h}_2 Q_2(x, y, z, w) Q_2(x^{-1}, y^{-1}, z^{-1}, w^{-1}).$$

Consequently, all solutions to equation (6.8) must satisfy the transformed equation

$$(6.9) \quad F(u + b_1, u + b_2, u + b_3, u + b_4) = 0.$$

When ignoring variables orders, the expression for Q (and consequently F) is determined exclusively by the parity of sign differences between $Q_1, Q_2 \in \{x \pm y \pm z \pm w\}$. This leads us to partition equation (6.9) into two cases according to this sign difference count.

Case I: two different signs. Without loss of generality, we consider the specific configuration $Q_1(x, y, z, w) = x + y + z + w$ and $Q_2(x, y, z, w) = x + y - z - w$ in this case. A direct calculation yields

$$Q = (\bar{h}_1 + \bar{h}_2) \left(4 + \frac{x}{y} + \frac{y}{x} + \frac{z}{w} + \frac{w}{z} \right) + (\bar{h}_1 - \bar{h}_2) \left(\frac{x}{z} + \frac{z}{x} + \frac{x}{w} + \frac{w}{x} + \frac{y}{z} + \frac{z}{y} + \frac{y}{w} + \frac{w}{y} \right),$$

where $\bar{h}_1 + \bar{h}_2$ and $\bar{h}_1 - \bar{h}_2$ are not simultaneously zero.

When $\bar{h}_1 + \bar{h}_2 = 0$, expansion of (6.9) yields a 24th-degree polynomial

$$c_{24}u^{24} + c_{23}u^{23} + \dots = 0$$

with the leading coefficient

$$c_{24} = 1048576 \left((b_1 - b_2)^2 + (b_3 - b_4)^2 \right)^4.$$

Thus, u admits at most 24 distinct solutions unless $\mathbf{b}$ satisfies $(b_1 - b_2)^2 + (b_3 - b_4)^2 = 0$.

For $\bar{h}_1 - \bar{h}_2 = 0$, simplification of (6.9) gives an 8th-degree polynomial

$$c_8 u^8 + c_7 u^7 \dots = 0,$$

where

$$c_8 = (b_1 - b_2)^8(b_1 + b_2 - b_3 - b_4)^8(b_3 - b_4)^8.$$

Hence, u has at most 8 solutions unless $(b_1 - b_2)(b_1 + b_2 - b_3 - b_4)(b_3 - b_4) = 0$.

When $(\bar{h}_1 + \bar{h}_2)(\bar{h}_1 - \bar{h}_2) \neq 0$, we define

$$(6.10) \quad t = \frac{\bar{h}_1 + \bar{h}_2}{\bar{h}_1 - \bar{h}_2},$$

where $t \notin \{0, \pm 1\}$. Treating t and $\mathbf{b}$ as parameters, a direct calculation shows that (6.9) reduces to a 24th-degree polynomial equation

$$c_{24}u^{24} + c_{23}u^{23} + c_{22}u^{22} + \cdots = 0$$

with the leading coefficient

$$c_{24} = 1048576t^8(t^2 - 1)^2 \left(((b_1 - b_2)^2 + (b_3 - b_4)^2)^2 t^2 - 4(b_1 - b_2)^2(b_3 - b_4)^2 \right)^2.$$

For $\mathbf{b} \notin \mathcal{V}^\Delta$ and $t \neq 0, \pm 1$, not both the t^2 and constant term coefficients in the final factor above can vanish simultaneously, and so the condition $c_{24} = 0$ holds precisely when $\mathbf{h}$ satisfies

$$(6.11) \quad t = \frac{\bar{h}_1 + \bar{h}_2}{\bar{h}_1 - \bar{h}_2} = \pm \frac{2(b_1 - b_2)(b_3 - b_4)}{(b_1 - b_2)^2 + (b_3 - b_4)^2}.$$

Substituting (6.11) into subsequent coefficients yields $c_{23} = 0$ and

$$c_{22} = 4294967296 \frac{(b_1 - b_2)^{12}(b_3 - b_4)^{12}}{(b_1 - b_2)^2 + (b_3 - b_4)^2} \\ \times (b_1 + b_2 - b_3 - b_4)^2(b_1 - b_2 + b_3 - b_4)^6(b_1 - b_2 - b_3 + b_4)^6.$$

Consequently, both c_{24} and c_{22} vanish simultaneously if and only if

$$(b_1 + b_2 - b_3 - b_4)(b_1 - b_2 + b_3 - b_4)(b_1 - b_2 - b_3 + b_4) = 0$$

and $\mathbf{h}$ satisfies (6.11). Otherwise, the variable u admits at most 24 distinct values.

Case II: one different sign. In this case, we take $Q_1(x, y, z, w) = x + y + z + w$ and $Q_2(x, y, z, w) = x + y + z - w$ for example. A direct calculation shows

$$Q = (\bar{h}_1 + \bar{h}_2) \left(4 + \frac{x}{y} + \frac{y}{x} + \frac{x}{z} + \frac{z}{x} + \frac{y}{z} + \frac{z}{y} \right) + (\bar{h}_1 - \bar{h}_2) \left(\frac{x}{w} + \frac{w}{x} + \frac{y}{w} + \frac{w}{y} + \frac{z}{w} + \frac{w}{z} \right),$$

where $\bar{h}_1 + \bar{h}_2$ and $\bar{h}_1 - \bar{h}_2$ also cannot vanish simultaneously.

After substituting this into (6.9), a simplification yields

$$40960000u^{32} + c_{31}u^{31} + \cdots = 0$$

when $\bar{h}_1 + \bar{h}_2 = 0$, and it reduces to

$$5308416u^{32} + c_{31}u^{31} + \cdots = 0$$

when $\bar{h}_1 - \bar{h}_2 = 0$. Consequently, we establish that u admits at most 32 possible values when $(\bar{h}_1 + \bar{h}_2)(\bar{h}_1 - \bar{h}_2) = 0$.

Given $(\bar{h}_1 + \bar{h}_2)(\bar{h}_1 - \bar{h}_2) \neq 0$, we employ the parameter t as defined in (6.10). A direct calculation then simplifies (6.9) to a 32nd-degree polynomial equation

$$c_{32}u^{32} + c_{31}u^{31} + \cdots = 0$$

where the leading coefficient satisfies

$$c_{32} = 65536(t^2 - 1)^6(25t^2 - 9)^2 \neq 0$$

except when $t = \pm 3/5$. Substituting $t = \pm 3/5$ into subsequent coefficients yields $c_{32} = \cdots = c_{29} = 0$, with

$$c_{28} = c \left(b_1^2 - 6b_1b_2 + b_2^2 - 6b_1b_3 - 6b_2b_3 + b_3^2 + 10b_1b_4 + 10b_2b_4 + 10b_3b_4 - 15b_4^2 \right)^2,$$

where $c = 618475290624/6103515625$. Consequently, u admits at most 32 solutions except when $\mathbf{h}$ and $\mathbf{b}$ simultaneously satisfy

$$5(\bar{h}_1 + \bar{h}_2) = \pm 3(\bar{h}_1 - \bar{h}_2)$$

and

$$b_1^2 - 6b_1b_2 + b_2^2 - 6b_1b_3 - 6b_2b_3 + b_3^2 + 10b_1b_4 + 10b_2b_4 + 10b_3b_4 - 15b_4^2 = 0.$$

We thereby establish the lemma. $\square$

Lemma 6.4. *For any vector $\mathbf{b} = (b_1, b_2, b_3, b_4)$ satisfying the congruence relations*

$$(6.12) \quad b_1 + b_2 \equiv b_3 + b_4 \pmod{p},$$

$$(6.13) \quad \bar{b}_1 + \bar{b}_2 \equiv \bar{b}_3 + \bar{b}_4 \pmod{p},$$

the squared tuple $(b_1^2, b_2^2, b_3^2, b_4^2)$ lies in the variety $\mathcal{V}^\Delta$ modulo p .

Proof. When $b_1 + b_2 \equiv b_3 + b_4 \equiv 0 \pmod{p}$, the conclusion follows immediately. Otherwise, we deduce from (6.13) the relation

$$\bar{b}_1\bar{b}_2(b_1 + b_2) \equiv \bar{b}_1 + \bar{b}_2 \equiv \bar{b}_3 + \bar{b}_4 \equiv \bar{b}_3\bar{b}_4(b_3 + b_4) \pmod{p},$$

implying $b_1b_2 \equiv b_3b_4 \pmod{p}$. Combining with (6.12), this establishes

$$(b_1 - b_2)^2 \equiv (b_3 - b_4)^2 \pmod{p}.$$

Consequently, either

$$b_1 - b_2 + b_3 - b_4 \equiv 0 \pmod{p},$$

or

$$b_1 - b_2 - b_3 + b_4 \equiv 0 \pmod{p}$$

must be satisfied. The lemma follows by combining these with (6.12). $\square$

6.3. Proof of Lemma 5.1. Lemma 6.4 along with a moment's reflection shows that the system

$$\begin{cases} f_i(u) = 0, \\ g_i(u) = 0 \end{cases}$$

has no solution when $\mathbf{b} \notin \mathcal{V}^\Delta$, implying $f_i(u)$ and $g_i(u)$ cannot both vanish in this case. We analyze $\mathcal{K}_1(\mathbf{b}, \mathbf{h}; p)$ by examining four cases classified according to the values of h_i .

When $h_1 = h_2 = 0$, we obtain $f_1(u) = f_2(u) = 0$ and thus $g_1(u)g_2(u) \neq 0$. By Lemma 6.2 and equation (6.2c), the variable u can take at most $O(1)$ values, while x_{11}, x_{21} are uniquely determined by each other; thus, in this case,

$$\mathcal{K}_1(\mathbf{b}, \mathbf{h}; p) \ll p.$$

When $h_1 = 0 \neq h_2$, we have $f_1(u) = 0$ with $f_2(u)g_1(u) \neq 0$. Lemma 6.2 shows that u can take at most $O(1)$ values, while x_{11}, x_{21} are uniquely determined by equations (6.2b) and (6.2c). Consequently,

$$\mathcal{K}_1(\mathbf{b}, \mathbf{h}; p) \ll 1.$$

The symmetric case $h_2 = 0 \neq h_1$ yields an identical bound.

When $h_1 h_2 \neq 0$ with $g_1(u)g_2(u) = 0$, the non-degeneracy condition $f_1(u)f_2(u) \neq 0$ is automatically satisfied. Lemma 6.2 restricts u to $O(1)$ possibilities, while equations (6.2a) and (6.2b) uniquely determine x_{11} and x_{21} . This yields the bound

$$\mathcal{K}_1(\mathbf{b}, \mathbf{h}; p) \ll 1.$$

Given $h_1 h_2 \neq 0$ and $g_1(u)g_2(u) \neq 0$, the variables x_{11} and x_{21} are uniquely determined by equations (6.2a) and (6.2b) for fixed $u, \mathbf{h}$ and $\mathbf{b}$. To determine u , we substitute (6.2a) and (6.2b) into (6.2c), yielding

$$\bar{h}_1 f_1(u)g_1(u) + \bar{h}_2 f_2(u)g_2(u) = 0.$$

For $x_1 \neq x_2$ with $g_1(u)g_2(u) \neq 0$, equations (6.1) and (6.2c) ensure $g_1(u) \neq g_2(u)$, which implies $Q_1 \neq Q_2$. By Lemma 6.3, u admits at most $O(1)$ values unless $(\mathbf{b}, \mathbf{h}) \in \mathcal{V}_4^{bad} \times \mathcal{V}_2^{bad}(\mathbf{b})$. Consequently, we obtain the bound

$$\mathcal{K}_1(\mathbf{b}, \mathbf{h}; p) \ll \begin{cases} p, & \text{if } (\mathbf{b}, \mathbf{h}) \in \mathcal{V}_4^{bad} \times \mathcal{V}_2^{bad}(\mathbf{b}); \\ 1, & \text{otherwise.} \end{cases}$$

Keeping in mind that $\mathcal{K}(\mathbf{b}, \mathbf{h}; p) \asymp \mathcal{K}_1(\mathbf{b}, \mathbf{h}; p)$, this completes the proof via exhaustive case analysis.

7. EVALUATION OF THE MOMENT

This section presents the asymptotic evaluation of moments for twisted L -function as stated in Theorem 1.2. We begin by performing several reduction steps and auxiliary estimates in §7.1, and we combine everything into a proof of Theorem 1.2 in §7.2.

7.1. Reduction steps. Following established methodology (see [2, Section 6.1]), we outline the key steps below.

Using the approximate functional equation from ([12, Theorem 5.3]), we express the central values as a convergent series

$$\begin{aligned} L(1/2, f_1 \otimes \chi) \overline{L(1/2, f_2 \otimes \chi)} &= \sum_{m,n \geq 1} \sum \frac{\lambda_1(m) \lambda_2(n) \chi(m) \bar{\chi}(n)}{(mn)^{\frac{1}{2}}} V\left(\frac{mn}{q^2}\right) \\ &\quad + \varepsilon(f_1, f_2, \chi) \sum_{m,n \geq 1} \sum \frac{\lambda_1(m) \lambda_2(n) \bar{\chi}(m) \chi(n)}{(mn)^{\frac{1}{2}}} V\left(\frac{mn}{q^2}\right), \end{aligned}$$

where $\varepsilon(f_1, f_2, \chi) = \pm 1$ denotes the root number, and $V(x)$ is a smooth function with rapid decay for $x \gg q^\varepsilon$. Character orthogonality yields that the average over primitive characters reduces to combinations of

$$\frac{1}{\varphi^*(q)} \sum_{d|q} \mu\left(\frac{q}{d}\right) \frac{\varphi(d)}{(MN)^{\frac{1}{2}}} \sum_{\substack{m \equiv \pm n \pmod{d} \\ (mn, q)=1}} \lambda_1(m) \lambda_2(n) V\left(\frac{mn}{q^2}\right).$$

The main term in (1.6) originates from the diagonal term $m = n$, computable via Mellin inversion and contour shifting (see [4, Section 3.1]). This calculation requires no assumptions (such as factorability of q or bounds toward the Ramanujan–Petersson conjecture) beyond those already stated in Theorem 1.2. The off-diagonal contribution requires bounding

$$(7.1) \quad \mathcal{B}^\pm(M, N) := \frac{1}{\varphi^*(q)} \sum_{d|q} \mu\left(\frac{q}{d}\right) \frac{\varphi(d)}{(MN)^{\frac{1}{2}}} \sum_{\substack{m \equiv \pm n \pmod{d} \\ (mn, q)=1 \\ m \neq n}} \lambda_1(m) \lambda_2(n) W_1\left(\frac{m}{M}\right) W_2\left(\frac{n}{N}\right)$$

for $N \geq M$ and $MN \leq q^{2+\varepsilon}$, where W_1, W_2 are test functions supported on $[1, 2]$ with derivatives satisfying $W_1^{(j)}, W_2^{(j)} \ll_j q^\varepsilon$.

To begin with, we establish two crucial estimates for $\mathcal{B}^\pm(M, N)$. Some of our estimates will be in terms of an admissible exponent $\theta \geq 0$ toward the Ramanujan–Petersson conjecture for Hecke–Maaß newforms f of arbitrary level ℓ , that is, an exponent such that the corresponding Hecke eigenvalues satisfy

$$(7.2) \quad \lambda_f(n) \ll n^{\theta+\varepsilon};$$

currently, $\theta = \frac{7}{64}$ is known to be admissible by the work of Kim–Sarnak [13]. Of course, the corresponding bound $O_\varepsilon(n^\varepsilon)$ is trivial for Eisenstein series and known for holomorphic newforms by the work of Deligne [6].

First, we have the trivial bound

$$(7.3) \quad \mathcal{B}^\pm(M, N) \ll N^\theta \frac{(MN)^{\frac{1}{2}}}{q} (MNq)^\varepsilon$$

for $N \geq M \geq 1$, derived via the Cauchy–Schwarz inequality using

$$\lambda_2(n) \ll n^{\theta+\varepsilon}, \quad \sum_{m \leq M} |\lambda_1(m)|^2 \ll M.$$

The second estimate applies specifically to balanced cases where M and N are comparable in size.

Lemma 7.1. *For $N \geq M \geq 1$ and $MN \leq q^{2+\varepsilon}$, we have*

$$(7.4) \quad \mathcal{B}^\pm(M, N) \ll \left(\frac{N}{M}\right)^{\frac{1}{4}} q^{-\frac{1}{4}+\varepsilon} + \left(\frac{N}{M}\right)^{\frac{1}{2}} q^{-\frac{1}{2}+\varepsilon} + q^{-\frac{1}{2}+2\theta+\varepsilon}.$$

Proof. We will prove Lemma 7.1 by refining the arguments of [4] to remove all dependence on the Ramanujan–Petersson conjecture. For ease of reference, we recall the notations from [4, (3.4)–(3.8)]:

$$\begin{aligned} S_{N,M,d,q} &= \frac{d}{(MN)^{1/2}} \sum_{\substack{m \equiv n \pmod{d} \\ (mn, q)=1 \\ m \neq n}} \lambda_1(m) \lambda_2(n) W_1\left(\frac{m}{M}\right) W_2\left(\frac{n}{N}\right), \\ \mathcal{D}(\ell_1, \ell_2, h, N, M) &= \sum_{\ell_1 n - \ell_2 m = h} \lambda_1(m) \lambda_2(n) W_1\left(\frac{\ell_2 m}{M}\right) W_2\left(\frac{\ell_1 n}{N}\right), \\ \mathcal{S}(\ell_1, \ell_2, d, N, M) &= \sum_r \mathcal{D}(\ell_1, \ell_2, rd, N, M). \end{aligned}$$

Bounds on $S_{N,M,d,q}$ directly lead to corresponding estimates on $\mathcal{B}^\pm(M, N)$. In particular, the estimate (7.4) follows from the combination of the following two bounds:

$$(7.5) \quad S_{N,M,d,q} \ll (qN)^\varepsilon \left(\frac{(Nq)^{\frac{1}{2}}}{M^{\frac{1}{2}}} + \frac{N^{\frac{3}{4}}}{M^{\frac{1}{4}}} + \frac{N^{\frac{1}{4}} q^{\frac{3}{4}}}{M^{\frac{1}{4}}} + N^{\frac{1}{2}} q^{\frac{1}{4}} \right) \quad (N \geq 20M),$$

$$(7.6) \quad S_{N,M,d,q} \ll \frac{q^{\theta+\varepsilon} N^{1+\theta}}{M^{\frac{1}{2}}} \quad (N \geq M).$$

These bounds were originally formulated and proved in [4, (3.11), (3.12)] (the latter without the q^θ factor) under the Deligne’s bound for $\lambda_j(n)$, $j = 1, 2$. We now prove (7.5) and (7.6) independently of the Ramanujan–Petersson conjecture.

The sum $S_{N,M,d,q}$ is related to the other two sums above (which are in turn treated by the shifted convolution sum methods) in [4, (3.10)]. Using (7.2) in place of Deligne’s bound in [4, (3.10)] introduces an additional $(f_1 f_2)^{\theta+\varepsilon}$ factor in the final expression:

$$(7.7) \quad S_{N,M,d,q} \ll \frac{d}{(NM)^{\frac{1}{2}}} \sum_{\substack{g_1 | f_1 | q \\ g_2 | f_2 | q}} (f_1 f_2)^{\theta+\varepsilon} |\mathcal{S}(f_1 g_1, f_2 g_2, d, N, M)|.$$

We begin by proving (7.5). Crucial in deriving its precedent, [4, (3.11)], is the upper bound on $\mathcal{S}(\ell_1, \ell_2, d, N, M)$ for $N \geq 20M$ in [4, Proposition 8], whose proof we note is entirely independent of the Ramanujan–Petersson conjecture. Indeed, up to the intermediate bound [4, (8.17)],

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

the proof proceeds by shifted convolution sum methods and the Hecke eigenvalues $\lambda_j(n)$ are only estimated on average using the bounds of Rankin–Selberg and Wilton. To obtain [4, Proposition 8], the extraneous factor $(\ell_1 \ell_2, d)^{1/2}$ must be removed. This is accomplished in the argument below [4, (8.17)], which uses the same decompositions as our (7.10) and (7.11) below, where $(\ell'_1 \ell'_2, d) = 1$, whence $(\ell'_1 g \ell'_2 h, d') = (gh, d') \mid gh$, and using only the trivial bounds $\lambda_j(n) \ll n^{1/2}$ suffices to complete the proof of [4, Proposition 8] because the combined contribution of each term $N^u M^v / d^w$ in the above bound is seen to be

$$(7.8) \quad \ll (gh)^{1/2} \left(\frac{\delta_1 \delta_2}{gh} \right)^{\frac{1}{2}} \frac{(N/\delta \delta_1 \delta_2 \tilde{\ell})^u (M/\delta \delta_1 \delta_2 \tilde{\ell})^v}{(d/\delta \delta_1 \delta_2)^w} \ll \frac{(\delta_1 \delta_2)^{\frac{1}{2}}}{(\delta \delta_1 \delta_2)^{u+v-w}} \frac{1}{\tilde{\ell}^{u+v}} \frac{N^u M^v}{d^w} \ll \frac{N^u M^v}{d^w}$$

in light of $u + v \geq w + \frac{1}{2}$.

Now, [4, Proposition 8] can be further refined: namely, under identical notations, we have that, for $N \geq 20M$ and still without any recourse to the Ramanujan–Petersson conjecture,

$$(7.9) \quad \mathcal{S}(\ell_1, \ell_2, d, N, M) \ll (dN)^\varepsilon \left(\frac{N}{d^{\frac{1}{2}}} + \frac{N^{\frac{5}{4}} M^{\frac{1}{4}}}{d(\ell_1 \ell_2)^{\frac{1}{4}}} + \frac{N^{\frac{3}{4}} M^{\frac{1}{4}}}{d^{\frac{1}{4}}} + \frac{NM^{\frac{1}{2}}}{d^{\frac{3}{4}}} \right).$$

Indeed, this follows by simply retaining the factor $(\ell_1 \ell_2)^{\frac{1}{4}}$ between [4, (8.13)] and [4, (8.17)].

With the refined estimate (7.9) at our disposal, we now return to (7.7) and set $\ell_1 = f_1 g_1$, $\ell_2 = f_2 g_2$. We again apply the decomposition as in [4, Section 8.5]

$$(7.10) \quad \ell_1 = \ell'_1 \tilde{\ell} \delta \delta_1, \quad \ell_2 = \ell'_2 \tilde{\ell} \delta \delta_2, \quad d = d' \delta \delta_1 \delta_2,$$

where $\delta = (d, \ell_1, \ell_2)$, $\delta_1 = (d, \ell_1)/\delta$, $\delta_2 = (d, \ell_2)/\delta$, $\tilde{\ell} = (\ell_1, \ell_2)/\delta$. Using the Hecke relation as in [4, Section 8.5] and (7.2) yields

$$(7.11) \quad S_{N,M,d,q} \ll \frac{d}{(NM)^{\frac{1}{2}}} \sum_{\substack{g_1 | f_1 | q \\ g_2 | f_2 | q}} (f_1 f_2 \delta_1 \delta_2)^{\theta + \varepsilon} \sum_{g | \delta_2} \sum_{h | \delta_1} \left| \mathcal{S} \left(\ell'_1 g, \ell'_2 h, d', \frac{N}{\delta \delta_1 \delta_2 \tilde{\ell}}, \frac{M}{\delta \delta_1 \delta_2 \tilde{\ell}} \right) \right|.$$

Estimating the innermost term $\mathcal{S}$ using (7.9) and handling the resulting terms as in (7.8), we obtain

$$S_{N,M,d,q} \ll (qN)^\varepsilon \sum_{f_1, f_2 | q} \frac{(f'_1 f'_2)^\theta}{(\delta \delta_1 \delta_2)^{\frac{1}{2} - 2\theta} \tilde{\ell}^{1 - 2\theta}} \left(\frac{(dN)^{\frac{1}{2}}}{M^{\frac{1}{2}}} + \frac{N^{\frac{3}{4}}}{(\ell'_1 \ell'_2)^{\frac{1}{4}} M^{\frac{1}{4}}} + d^{\frac{3}{4}} \frac{N^{\frac{1}{4}}}{M^{\frac{1}{4}}} + d^{\frac{1}{4}} N^{\frac{1}{2}} \right),$$

where $f'_j = f_j / (f_j, \tilde{\ell} \delta \delta_j)$ for $j = 1, 2$. It is easy to see that $f'_j \mid \ell'_j$ and $f'_j \mid q/d$, which yield $(f'_1, f'_2) = 1$ and thus $(f'_1 f'_2)^\theta \leq \min\{(q/d)^\theta, (\ell'_1 \ell'_2)^\theta\}$. Using that $\theta \leq 1/4$, dropping the denominator $(\delta \delta_1 \delta_2)^{1/2 - 2\theta} \tilde{\ell}^{1 - 2\theta}$, and using, for every $u \geq \frac{1}{4}$, $\sum_{d|q} (q/d)^\theta d^u \ll q^{u+\varepsilon}$, we conclude (7.5).

We now turn our attention to proving (7.6). The bound [4, (3.12)] is a trivial estimate of [4, (3.10)] using the uniform individual bound (quoted there as [4, (3.9)])

$$(7.12) \quad \mathcal{D}(\ell_1, \ell_2, h, N, M) \ll (NMq)^\varepsilon (N + M)^{\frac{1}{2} + \theta},$$

which is established in [1, Theorem 1.3] independently of the Ramanujan–Petersson conjecture. Let $\ell_1, \ell_2, h \in \mathbb{N}$ with $\delta = (\ell_1, \ell_2)$. Observing from the definition that

$$(7.13) \quad \mathcal{D}(\ell_1, \ell_2, h, N, M) = \mathcal{D}\left(\frac{\ell_1}{\delta}, \frac{\ell_2}{\delta}, \frac{h}{\delta}, \frac{N}{\delta}, \frac{M}{\delta}\right),$$

we refine (7.12) to

$$\mathcal{D}(\ell_1, \ell_2, h, N, M) \ll (NMq)^\varepsilon \left(\frac{N+M}{\delta}\right)^{\frac{1}{2}+\theta}.$$

Applying this to (7.7) yields

$$S_{N,M,d,q} \ll \frac{d}{(NM)^{\frac{1}{2}}} q^\varepsilon \frac{N}{d} \sum_{g_1|f_1|q} \sum_{g_2|f_2|q} (f_1 f_2)^\theta \left(\frac{N+M}{(f_1 g_1, f_2 g_2)}\right)^{\frac{1}{2}+\theta} \ll \frac{q^{\theta+\varepsilon} N^{1+\theta}}{M^{\frac{1}{2}}}.$$

This proves (7.6), and, as already commented, (7.5) and (7.6) combined imply the statement of the lemma. $\square$

For unbalanced terms, we assume $N \geq q^{\frac{3}{2}-\delta}$ for some small $\delta > 0$. Applying the Voronoi summation formula to the n -sum transform the problem to evaluating bilinear sums as in Theorem 1.1, where both variables have lengths comparable to $q^{\frac{1}{2}}$.

Our next Lemma provides a requisite Voronoi formula, available in [7, Lemma 2.2] or [2, Lemma 2.3].

Lemma 7.2 (Voronoi formula). *Let $q \in \mathbb{N}$, $c \in \mathbb{Z}$ with $(c, q) = 1$, and let V be a smooth compactly supported function. Then*

$$(7.14) \quad \sum_n \lambda(n) V\left(\frac{n}{N}\right) e\left(\frac{cn}{q}\right) = \frac{N}{q} \sum_{\pm} \sum_n \lambda(n) \mathring{V}_{\pm}\left(\frac{nN}{q^2}\right) e\left(\pm \frac{\bar{c}n}{q}\right),$$

where the Hankel-type transforms $\mathring{V}_{\pm} : (0, \infty) \rightarrow \mathbb{C}$ are given by

$$\mathring{V}_{\pm}(y) = \int_0^\infty V(x) \mathcal{J}_{\pm}(4\pi\sqrt{xy}) dx$$

with the Bessel kernels satisfying

$$(7.15) \quad \mathcal{J}_+(x) = 2\pi i^k J_{k-1}(x), \quad \mathcal{J}_-(x) = 0$$

for f a cuspidal holomorphic newform of weight k , and

$$(7.16) \quad \mathcal{J}_+(x) = \frac{\pi i}{\sinh(\pi t)} (J_{2i\kappa}(x) - J_{-2i\kappa}(x)), \quad \mathcal{J}_-(x) = 4 \cosh(\pi\kappa) K_{2i\kappa}(x)$$

for f a cuspidal Maaß newform with spectral parameter κ .

The Hankel-type transforms $\mathring{V}_{\pm}(y)$ are well-known to be of Schwartz class. Since the Selberg conjecture of exceptional eigenvalues is known for $\mathrm{SL}_2(\mathbb{Z})$, the bound (see [2, Lemma 2.4])

$$y^j \mathring{V}_{\pm}(y) \ll_{i,j} (1+y)^{\frac{j}{2}} \left(1+y^{\frac{1}{2}} q^{-\varepsilon}\right)^{-i}$$

holds for all $i, j \geq 0$ and $y > 0$, when f is a cuspidal newform (holomorphic or Maaß) of level one and $V^{(j)}(y) \ll_j q^{j\varepsilon}$. In particular, we have that $\mathring{V}_{\pm}(y) \ll_j y^{-j}$ for $y \geq q^\varepsilon$.

7.2. Proof of Theorem 1.2. In this section, we combine all the preparatory steps from §7.1 and the input from Theorem 1.1 to prove Theorem 1.2.

Proof of Theorem 1.2. Recall that $\eta = \frac{1}{216}$. As explained in §7.1 (see the reduction to (7.1)), the proof of Theorem 1.2 reduces to establishing the bound

$$(7.17) \quad \mathcal{B}^\pm(M, N) \ll q^{-\eta+\varepsilon}.$$

We parametrize the exponents as

$$M = q^u, \quad N = q^v, \quad v^* = 2 - v.$$

From (7.3) and (7.4), the desired bound (7.17) follows immediately if either $u + v \leq 2 - 2\eta - 2\theta v$ or $v - u \leq 1 - 4\eta$. Thus, we may restrict our consideration to the complementary regime

$$2 - 2\eta - 2\theta v \leq u + v \leq 2 + \varepsilon, \quad 1 - 4\eta \leq v - u,$$

which yields the key constraints

$$(7.18) \quad u \leq 1/2 + 2\eta + \varepsilon, \quad u - \varepsilon \leq v^* \leq 1/2 + \frac{3\theta + 6\eta}{2 + 2\theta}, \quad u + v^* \leq 1 + 4\eta.$$

After rearrangement in (7.1), we derive

$$\mathcal{B}^\pm(M, N) = \frac{1}{\varphi^*(q)} \sum_{d|q} \mu\left(\frac{q}{d}\right) \frac{\varphi(d)}{(MN)^{\frac{1}{2}}} \sum_{(m,q)=1} \lambda_1(m) W_1\left(\frac{m}{M}\right) \sum_{\substack{n \equiv \pm m \pmod{d} \\ (n,q)=1}} \lambda_2(n) W_2\left(\frac{n}{N}\right).$$

Given $(m, q) = 1$ and $n \equiv \pm m \pmod{d}$, the condition $(n, q) = 1$ is equivalent to $(n, q_d) = 1$. Applying Möbius inversion and the Hecke relation yields

$$\begin{aligned} \sum_{\substack{n \equiv \pm m \pmod{d} \\ (n,q)=1}} \lambda_2(n) W_2\left(\frac{n}{N}\right) &= \sum_{f|q_d} \mu(f) \sum_{fn \equiv \pm m \pmod{d}} \lambda_2(fn) W_2\left(\frac{fn}{N}\right) \\ &= \sum_{g|f|q_d} \mu(f) \mu(g) \lambda_2\left(\frac{f}{g}\right) \sum_{n \equiv \pm \overline{f}gm \pmod{d}} \lambda_2(n) W_2\left(\frac{fgn}{N}\right). \end{aligned}$$

We detect the condition $n \equiv \pm \overline{f}gm \pmod{d}$ with primitive additive characters modulo r for $r | d$, and then apply the Voronoi summation formula (Lemma 7.2) to the n -sum, reducing it to the sum of four sums of the form

$$\frac{1}{d} \sum_{r|d} \frac{N}{fgr^{\frac{1}{2}}} \sum_{n \geq 1} \lambda_2(n) \mathring{W}_{2,\pm}\left(\frac{n}{N^*}\right) \text{Kl}_2(\pm \overline{f}gmn; r),$$

where $N^* = fgr^2/N \ll q^{v^*}$ since $fr \leq q$, and the two $\pm$ signs vary independently of each other. Substituting this back into $\mathcal{B}^\pm(M, N)$ gives

$$\mathcal{B}^\pm(M, N) \ll \sum_{r|q} \sum_{f|g|q_r} \frac{q^\varepsilon f^\theta r^{\frac{1}{2}}}{q(fgMN^*)^{\frac{1}{2}}} \left| \sum_{m \leq M} \sum_{n \leq N^*} \alpha_m \beta_n \text{Kl}_2(\pm \overline{f}gmn; r) \right|$$

with coefficients

$$\alpha_m = \lambda_1(m) W_1\left(\frac{m}{M}\right), \quad \beta_n = \lambda_2(n) \mathring{W}_{2,\pm}\left(\frac{n}{N^*}\right).$$

Trivial summation via Weil's bound for Kloosterman sums produces

$$\mathcal{B}^\pm(M, N) \ll (M/N)^{\frac{1}{2}} f^\theta r^{\frac{3}{2}} q^{-1+\varepsilon} \leq r^{\frac{3}{2}-\theta} q^{-\frac{3}{2}+\theta+2\eta+\varepsilon},$$

establishing (7.17) unless

$$(7.19) \quad r \geq q^{1-\frac{6}{3-2\theta}\eta}.$$

Given η 's small magnitude, (7.19) combined with (7.18) ensures the condition (1.2) in Theorem 1.1. Thus, we derive from Theorem 1.1 and (7.18) that

$$\begin{aligned} \mathcal{B}^\pm(M, N) &\ll \sum_{r|q} \frac{r^{\frac{1}{2}}(MN^*)^{\frac{1}{2}}}{q} \left(M^{-\frac{1}{2}} r^{\frac{1}{6}} + M^{-\frac{3}{25}} (N^*)^{-\frac{3}{10}} r^{\frac{1}{5}} + (MN^*)^{-\frac{3}{16}} r^{\frac{11}{64}} \right) \\ &\ll q^{-\frac{1}{3}+\frac{1}{2}v^*+\varepsilon} + q^{-\frac{3}{10}+\frac{9}{50}u+\frac{1}{5}(u+v^*)+\varepsilon} + q^{-\frac{21}{64}+\frac{5}{16}(u+v^*)+\varepsilon} \leq q^{-\eta+\varepsilon} \end{aligned}$$

for $v^* \leq \frac{2}{3} - 2\eta$ with $\eta = \frac{1}{216}$. The remaining case $v^* > \frac{2}{3} - 2\eta$ can be effectively handled using the estimate (1.4), which is deduced from the Pólya–Vinogradov method. $\square$

Remark 7.1. It remains to justify the claims

$$P(1), Q(1) = \exp(O_\varepsilon((\log \log q)^\varepsilon)), \quad P'(1)/P(1) = O_\varepsilon((\log \log q)^{1+\varepsilon})$$

for the finite Euler products defined in (1.5). Recall that

$$\lambda(p^2) = \lambda(p)^2 + 1, \quad \lambda(p) \ll p^\theta$$

for an absolute $\theta < 1/4$ (where $\theta = 7/64$ is admissible). Choosing any $A > 1/(1 - 2\theta)$, we have that $\lambda(p^2)/p \ll 1/\log q$ for all $p \geq \log^A q$, whence, keeping in mind the simple estimate $\omega(q) \ll \log q$,

$$\sum_{p|q} \frac{\lambda(p^2)}{p} \ll \sum_{p \leq \log^A q} \frac{\lambda(p)^2 + 1}{p}.$$

Taking a large integer r and rewriting every product $p_1 \dots p_r = nm$ with n square-free, m square-full, and $(n, m) = 1$, we have that

$$\begin{aligned} \left(\sum_{p \leq \log^A q} \frac{\lambda(p)^2 + 1}{p} \right)^r &= \sum_{p_1, \dots, p_r \leq \log^A q} \frac{(\lambda(p_1)^2 + 1) \cdots (\lambda(p_r)^2 + 1)}{p_1 \cdots p_r} \\ &\ll_r \sum_{n \leq \log^{Ar} q} \frac{1 * \lambda^2(n)}{n} \sum_{m \text{ square-full}} \frac{1}{m^{1-2\theta}} \\ &\ll_r \log \log q \sum_{n \leq \log^{Ar} q} \frac{\lambda(n)^2}{n} \ll_r (\log \log q)^2. \end{aligned}$$

Taking $r > 2/\varepsilon$ yields²

$$(7.20) \quad \sum_{p \leq \log^A q} \frac{\lambda(p)^2 + 1}{p} \ll_\varepsilon (\log \log q)^\varepsilon,$$

²We note that, in fact, $\theta < 1/2$ suffices to reach the conclusions that follow, by instead separating out s -power-full numbers with a fixed $s > 1/(1 - 2\theta)$ and then taking $r > 2s/\varepsilon$.

hence

$$P(1), Q(1) = \exp \left(O \left(\sum_{p|q} \frac{\lambda(p^2)}{p} \right) + O(1) \right) = \exp (O_\varepsilon((\log \log q)^\varepsilon)),$$

$$\frac{P'(1)}{P(1)} \ll \sum_{p|q} \frac{\lambda(p^2) \log p}{p} + O(1) \ll_\varepsilon (\log \log q)^{1+\varepsilon}.$$

REFERENCES

- [1] V. Blomer, Shifted convolution sums and subconvexity bounds for automorphic L -functions, *Int. Math. Res. Not* (2004), 3905–3926. [33](#)
- [2] V. Blomer, É. Fouvry, E. Kowalski, Ph. Michel, and D. Milićević, On moments of twisted L -functions, *Amer. J. Math.* 139 (2017), 707–768. [2](#), [15](#), [30](#), [33](#)
- [3] V. Blomer, É. Fouvry, E. Kowalski, Ph. Michel, D. Milićević, and W. Sawin, The Second Moment Theory of Families of L -Functions — The Case of Twisted Hecke L -Functions, *Mem. Amer. Math. Soc.* 282 (2023), no. 1394, v+148 pp. [2](#), [3](#)
- [4] V. Blomer and D. Milićević, The second moment of twisted modular L -functions, *Geom. Funct. Anal.* 25 (2015), 453–516. [2](#), [3](#), [6](#), [7](#), [11](#), [12](#), [30](#), [31](#), [32](#)
- [5] F. Brumley, Effective multiplicity one on GL_n and narrow zero-free regions for Rankin-Selberg L -functions, *Amer. J. Math.*, 128 (2006), 1455–1474. [4](#)
- [6] P. Deligne, La conjecture de Weil, I, *Inst. Hautes Études Sci. Publ. Math.* 43 (1974), 273–307. [30](#)
- [7] 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), 979–1014. [33](#)
- [8] É. Fouvry, E. Kowalski, and Ph. Michel, A study in sums of products, *Phil. Trans. R. Soc. A.* 373 (2015), no. 2040, 20140309, 26 pp. [1](#)
- [9] É. Fouvry, E. Kowalski, and Ph. Michel, Algebraic trace functions over the primes, *Duke Math. J.* 163 (2014), 1683–1736. [2](#)
- [10] É. Fouvry and Ph. Michel, Sur certaines sommes d’exponentielles sur les nombres premiers, *Ann. Sci. École Norm. Sup. (4)* 31 (1998), 93–130. [14](#), [15](#)
- [11] J. Hoffstein and P. Lockhart, Coefficients of Maaß forms and the Siegel zero, with an appendix by D. Goldfeld, J. Hoffstein, and D. Lieman, *Ann. of Math. (2)*, 140 (1994), 161–181. [4](#)
- [12] H. Iwaniec and E. Kowalski, *Analytic Number Theory*, Amer. Math. Soc. Colloq. Publ. 53, Amer. Math. Soc., Providence, RI, 2004. [9](#), [30](#)
- [13] H. H. Kim, Functoriality for the exterior square of GL_4 and the symmetric fourth of GL_2 , *J. Amer. Math. Soc.* 16 (2003), 139–183, with Appendix 1 by Dinakar Ramakrishnan and Appendix 2 by Kim and Peter Sarnak. [30](#)
- [14] E. Kowalski, Ph. Michel, and W. Sawin, Bilinear forms with Kloosterman sums and applications, *Ann. of Math. (2)* 186 (2017), 413–500. [1](#), [2](#), [3](#), [6](#), [16](#), [17](#)
- [15] A. Pascadi, Non-abelian amplification and bilinear forms with Kloosterman sums, arXiv e-prints, November 2025. [2](#), [3](#)
- [16] D. Ramakrishnan and S. Wang, On the exceptional zeros of Rankin-Selberg L -functions, *Compos. Math.*, 135 (2003), 211–244. [4](#)
- [17] I. D. Shkredov, On asymptotic formulae in some sum-product questions. *Trans. Moscow Math. Soc.* 79 (2018), 231–281. [3](#)
- [18] I. D. Shkredov, Modular hyperbolas and bilinear forms of Kloosterman sums, *J. Number Theory* 220 (2021), 182–211. [3](#)
- [19] M. P. Young, The fourth moment of Dirichlet L -functions, *Ann. of Math. (2)* 173 (2011), 1–50. [3](#)

D. MILIĆEVIĆ: BRYN MAWR COLLEGE, DEPARTMENT OF MATHEMATICS, 101 NORTH MERION AVENUE, BRYN MAWR, PA 19010, USA
Email address: `dmilicevic@brynmawr.edu`

X. QIN: SCHOOL OF MATHEMATICS, HEFEI UNIVERSITY OF TECHNOLOGY, HEFEI 230009, P.R. CHINA
Email address: `qinxh@mail.hfut.edu.cn`

X. WU: SCHOOL OF MATHEMATICS, HEFEI UNIVERSITY OF TECHNOLOGY, HEFEI 230009, P.R. CHINA
Email address: `xswu@amss.ac.cn`