Forum of Mathematics, Pi (2026), **Vol. 14**:e8 1–54
doi:10.1017/fmp.2026.10025

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Large sieve inequalities for exceptional Maass forms and the greatest prime factor of $n^2 + 1$

Alexandru Pascadi

Mathematical Institute, University of Oxford, United Kingdom; E-mail: alexpascadi@gmail.com.

Received: 9 January 2025; **Revised:** 11 September 2025; **Accepted:** 5 October 2025

2020 Mathematics Subject Classification: *Primary* – 11N75; *Secondary* – 11N32, 11L05

Abstract

We prove new large sieve inequalities for the Fourier coefficients $\rho_{j\mathfrak{a}}(n)$ of exceptional Maass forms of a given level, weighted by sequences (a_n) with sparse Fourier transforms – including two key types of sequences that arise in the dispersion method. These give the first savings in the exceptional spectrum for the critical case of sequences as long as the level, and lead to improved bounds for various multilinear forms of Kloosterman sums. As an application, we show that the greatest prime factor of $n^2 + 1$ is infinitely often greater than $n^{1.3}$, improving Merikoski's previous threshold of $n^{1.279}$. We also announce applications to the exponents of distribution of primes and smooth numbers in arithmetic progressions.

Contents

1. Introduction

Let $m, n, c \in \mathbb{Z}$ with $c \geq 1$, and consider the classical Kloosterman sums

$$S(m,n;c) := \sum_{x \in (\mathbb{Z}/c\mathbb{Z})^\times} e\left(\frac{mx + n\overline{x}}{c}\right), \tag{1.1}$$

where $e(\alpha) := \exp(2\pi i\alpha)$ and $x\overline{x} \equiv 1 \pmod{c}$. A great number of results in analytic number theory, particularly on the distribution of primes [2, 34, 35, 36, 9, 8, 37, 32] and properties of Dirichlet L-functions [11, 12, 46, 48, 15, 45], rely on bounding exponential sums of the form

$$\sum_{m \sim M} a_m \sum_{n \sim N} b_n \sum_{(c,r)=1} g\left(\frac{c}{C}\right) S(m\overline{r}, \pm n; sc), \tag{1.2}$$

where (a_m) and (b_n) are rough sequences of complex numbers, g is a compactly-supported smooth function, and r, s are coprime positive integers. One can often (but not always [37, 8, 34]) leverage some additional averaging over r and s, if one of the sequences (a_m), (b_n) is independent of r, s.

Estimates for sums like (1.2) are typically obtained via the spectral theory of automorphic forms [25, 24], following Deshouillers–Iwaniec [10]; this allows one to bound (1.2) by certain averages of the sequences (a_m), (b_n) with the Fourier coefficients of automorphic forms for $\Gamma_0(rs)$. Often in applications, the limitation in these bounds comes from our inability to rule out the existence of *exceptional Maass cusp forms*, corresponding to exceptional eigenvalues $\lambda \in (0, \frac{1}{4})$ of the hyperbolic Laplacian. This is measured by a parameter $\theta = \max_\lambda \max(0, \frac{1}{4} - \lambda)^{1/2}$; under *Selberg's eigenvalue conjecture* there would be no exceptional eigenvalues [41], so one could take $\theta = 0$. But unconditionally, the record is Kim–Sarnak's bound $\theta \leq \frac{7}{64}$, based on the automorphy of symmetric fourth power L-functions [28, Appendix 2].

This creates a power-saving gap between the best conditional and unconditional results in various arithmetic problems, for example, on the prime factors of quadratic polynomials [8, 37], the exponents of distribution of primes [32] and smooth numbers [39] in arithmetic progressions, and low-lying zeros of Dirichlet L-functions [15]. Improvements to the dependency on θ, which help narrow this gap, come from large sieve inequalities for the Fourier coefficients of exceptional Maass cusp forms (see [10, Theorems 5, 6, 7] and their optimizations in [14, 1, 32, 39]), which function as weak on-average substitutes for Selberg's eigenvalue conjecture. However, in the key setting of fixed r, s and sequences (a_n) of length $N \approx rs$, no such savings were previously available.

Luckily, for many of the most important applications, we don't need to handle (1.2) for completely arbitrary sequences, but only for those arising from variations of Linnik's dispersion method [33, 18, 2, 3, 4]; these often have the rough form

$$a_m = e(m\alpha) \qquad \text{and} \qquad b_n = \sum_{\substack{h_1, h_2 \sim H \\ h_1\ell_1 - h_2\ell_2 = n}} 1, \tag{1.3}$$

for $\alpha \in \mathbb{R}/\mathbb{Z}$ and $\ell_1 \asymp \ell_2 \gg H$ with $(\ell_1, \ell_2) = 1$. Our main results in this paper are new large sieve inequalities for such sequences, with Fourier transforms that obey strong concentration conditions. These are obtained by combining the framework of Deshouillers–Iwaniec with combinatorial ideas – specifically, with new estimates for bilinear sums of Kloosterman sums, stemming from a counting argument of Cilleruelo–Garaev [7]. The resulting improved bounds for (1.2) can then feed through to the strongest results on several well-studied arithmetic problems.

Figure 1 summarizes the results outlined above, which go from *counting problems* (on the top row), to *exponential sums* (middle row), to *automorphic forms* (bottom row), and then backwards. The transition between the first two rows is mostly elementary (using successive applications of Poisson summation, Cauchy–Schwarz, combinatorial decompositions, and/or sieve methods), while the transition between the last two rows uses the Kuznetsov trace formula [31, 10].

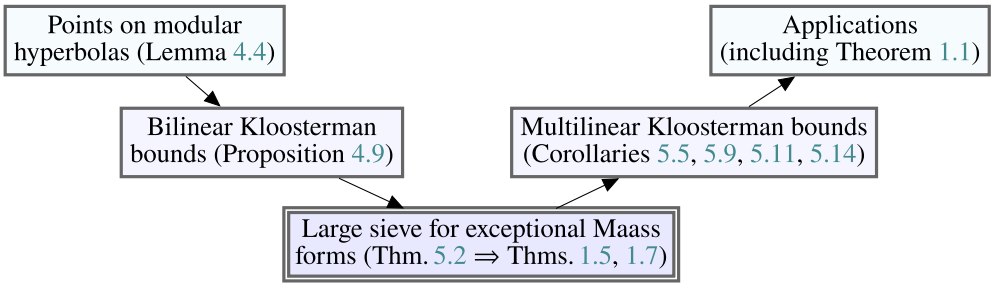

Figure 1. *Structure of paper (arrows signify logical implications).*

Before we dive into the large sieve inequalities, let us motivate our discussion with applications.

Theorem 1.1. *For infinitely many $n \in \mathbb{Z}_+$, the greatest prime factor of n^2+1 is larger than $n^{1.3}$.*

This result makes progress on a longstanding problem, approximating the famous conjecture that there exist infinitely many primes of the form n^2+1. Back in 1967, Hooley [23] proved the same result with an exponent of 1.1001, using the Weil bound for Kloosterman sums. In 1982, Deshouillers–Iwaniec [9] used their bounds on multilinear forms of Kloosterman sums [10] to improve this substantially, up to an exponent of 1.2024. More recently, using Kim–Sarnak's bound $\theta \le \frac{7}{64}$ [28, Appendix 2], de la Bretèche and Drappeau [8] optimized the exponent to 1.2182. Finally, Merikoski [37] proved a new bilinear estimate (still relying on the bounds of Deshouillers–Iwaniec [10]), and used Harman's sieve to reach the exponent 1.279; assuming Selberg's eigenvalue conjecture, Merikoski also reached the conditional exponent 1.312. With our new large sieve inequalities (Theorems 1.5 and 1.7), we can improve the arithmetic information due to both Merikoski [37] and de la Bretèche–Drappeau [8], leading to the unconditional result in Theorem 1.1. As in [37, 8], by adapting our proof, it should be possible to obtain similar results for other irreducible quadratic polynomials.

Additionally, we announce applications to the distribution of primes and smooth numbers in arithmetic progressions to large moduli. In [38], the author will show that the primes have *exponent of distribution* $5/8-\varepsilon$ using "triply-well-factorable" weights (λ_q) [35], in the sense that

$$\sum_{\substack{q\le x^{5/8-\varepsilon}\\(q,a)=1}} \lambda_q\left(\pi(x;q,a)-\frac{\pi(x)}{\varphi(q)}\right) \ll_{\varepsilon,A,a} \frac{x}{(\log x)^A},$$

where $\pi(x;q,a)$ denotes the number of primes up to x which are congruent to a mod q. A similar result, with the same exponent of $5/8-\varepsilon$, will be established for smooth numbers, using arbitrary 1-bounded weights (λ_q). These will improve results of Maynard [35] and Lichtman [32], respectively Drappeau [13] and the author [39]. Notably, our large sieve inequalities will suffice to completely eliminate the dependency on Selberg's eigenvalue conjecture in these cases.

We also note that an extension of our large sieve inequalities to Maass forms with a general nebentypus should have consequences to counting smooth values of irreducible quadratic polynomials [8, 21, 22] (by improving de la Bretèche–Drappeau's [8, Théorème 5.2]), and to enlarging the Fourier support in one-level density estimates for Dirichlet L-functions [15].

1.1. The large sieve inequalities

We now turn to our main technical results. The sums of Kloosterman sums from (1.2) are related to the Fourier coefficients of GL_2 automorphic forms of *level* $q=rs$ by the Kuznetsov trace formula [31, 10] for the congruence group $\Gamma_0(q)$.

More precisely, the spectral side of the Kuznetsov formula contains three terms, corresponding to the contribution of holomorphic forms, Maass forms, and Eisenstein series. The *exceptional* Maass forms are eigenfunctions of the hyperbolic Laplacian on $L^2(\Gamma_0(q)\backslash\mathbb{H})$ with eigenvalues $0 < \lambda < 1/4$; this (conjecturally empty) exceptional spectrum typically produces losses of the form $X^{\theta(q)}$, where X is a large parameter and $\theta(q) := \max_\lambda \sqrt{\max(0, \frac{1}{4} - \lambda)}$. The aforementioned large sieve inequalities for exceptional Maass forms can help alleviate this loss, by incorporating factors of X^θ. Below we state a known result for general sequences (a_n) (the values $X \in \{1, q/N\}$ corresponding to [10, Theorems 2 and 5]), which we aim to improve; we detail our notation in Section 3.

Theorem 1.2 (Large sieve with general sequences [10]). *Let $\varepsilon > 0$, $X > 0$, $N \geq 1/2$, and $(a_n)_{n\sim N}$ be a complex sequence. Let $q \in \mathbb{Z}_+$, $\mathfrak{a}$ be a cusp of $\Gamma_0(q)$ with $\mu(\mathfrak{a}) = q^{-1}$, and $\sigma_\mathfrak{a} \in \mathrm{PSL}_2(\mathbb{R})$ be a scaling matrix for $\mathfrak{a}$. Consider an orthonormal basis of Maass cusp forms for $\Gamma_0(q)$, with eigenvalues λ_j and Fourier coefficients $\rho_{j\mathfrak{a}}(n)$ around the cusp $\mathfrak{a}$ (via $\sigma_\mathfrak{a}$). Then with $\theta_j := \sqrt{\frac{1}{4} - \lambda_j}$, one has*

$$\sum_{\lambda_j<1/4} X^{2\theta_j} \left|\sum_{n\sim N} a_n \rho_{j\mathfrak{a}}(n)\right|^2 \ll_\varepsilon (qN)^\varepsilon \left(1 + \frac{N}{q}\right) \|a_n\|_2^2, \tag{1.4}$$

for any

$$X \ll \max\left(1, \frac{q}{N}, \frac{q^2}{N^3}\right). \tag{1.5}$$

Remark 1.3. As in [34, 39, 32], we use Deshouillers–Iwaniec's normalization [10] for the Fourier coefficients $\rho_{j\mathfrak{a}}(n)$ of Maass forms. In various other works [45, 14, 8, 37], $\rho_{j\mathfrak{a}}(n)$ are rescaled by $n^{-1/2}$.

Remark 1.4. An equivalent (and more common [10, 14]) way to phrase results like Theorem 1.2 is that

$$\sum_{\lambda_j<1/4} X^{2\theta_j} \left|\sum_{n\sim N} a_n \rho_{j\mathfrak{a}}(n)\right|^2 \ll_\varepsilon (qN)^\varepsilon \left(1 + \frac{X}{X_0}\right)^{2\theta(q)} \left(1 + \frac{N}{q}\right) \|a_n\|_2^2,$$

for *any* $X > 0$, and $X_0 = X_0(N, q)$ given by the right-hand side of (1.5). We prefer to state our large sieve inequalities in terms of the maximal value of X which does not produce any losses in the right-hand side, compared to the regular spectrum (i.e., $X \ll X_0$). We note that in applications, one usually has $\sqrt{q} \ll N \ll q$, and the best choice in (1.5) for this range is $X \asymp q/N$. But in the critical range $N \asymp q$, Theorem 1.2 is as good as the large sieve inequalities for the full spectrum [10, Theorem 2], since the limitation $X \ll 1$ forestalls any savings in the θ-aspect.

When some averaging over levels $q \sim Q$ is available, $\mathfrak{a} = \infty$, and (a_n), σ_∞ are independent of q, Deshouillers–Iwaniec [10, Theorem 6] improved the admissible range to $X \ll \max(1, (Q/N)^2)$; Lichtman [32] recently refined this to $X \ll \max(1, \min((Q/N)^{32/7}, Q^2/N))$, by making θ-dependencies explicit in [10, §8.2]. We note that these results are still limited at $X \ll 1$ when $N \asymp Q$.

Although it seems difficult to improve Theorem 1.2 in general (see Section 2.1), one can hope to do better for special sequences (a_n); for instance, the last term in (1.5) can be improved if the sequence (a_n) is sparse. In this paper, we consider the "dual" setting when (a_n) is sparse in frequency space, that is, when the Fourier transform $\widehat{a}(\xi) := \sum_n a_n e(-n\xi)$ is concentrated on a subset of $\mathbb{R}/\mathbb{Z}$. We give a general result of this sort in Theorem 5.2, which also depends on rational approximations to the support of $\widehat{a}$. Below we state the two main cases of interest, corresponding to the sequences in (1.3) (we also incorporate a scalar a in the Fourier coefficients, but on a first read one should take $a = 1$).

Theorem 1.5 (Large sieve with exponential phases). *Let $\varepsilon, X > 0$, $N \geq 1/2$, $\alpha \in \mathbb{R}/\mathbb{Z}$, and $q, a \in \mathbb{Z}_+$. Then with the notation of Theorem 1.2 and the choice of scaling matrix in (3.9), the bound*

$$\sum_{\lambda_j < 1/4} X^{2\theta_j} \left| \sum_{n \sim N} e(n\alpha)\, \rho_{j\mathfrak{a}}(an) \right|^2 \ll_\varepsilon (qaN)^\varepsilon \left(1 + \frac{aN}{q}\right) N \tag{1.6}$$

holds for all

$$X \ll \frac{\max\left(N, \frac{q}{a}\right)}{\min_{t \in \mathbb{Z}_+} (t + N\|t\alpha\|)}. \tag{1.7}$$

In particular, this implies the range $X \ll \max(\sqrt{N}, \frac{q}{a\sqrt{N}})$, uniformly in α and $\sigma_{\mathfrak{a}}$. The same result holds if $e(n\alpha)$ is multiplied by $\Phi(n/N)$, for any smooth function $\Phi : (0, 4) \to \mathbb{C}$ with $\Phi^{(j)} \ll_j 1$.

Here, $\|\alpha\|$ denotes the distance from α to 0 inside $\mathbb{R}/\mathbb{Z}$; the fact that the worst ("minor-arc") range covered by (1.7) is $X \ll \max(\sqrt{N}, \frac{q}{a\sqrt{N}})$ follows from a pigeonhole argument. The best range, $X \ll \max(N, \frac{q}{a})$, is achieved when α is $O(N^{-1})$ away from a rational number with bounded denominator. In particular, Theorem 1.5 obtains significant savings in the θ-aspect in the critical case $N \asymp q$, for an individual level q, which was previously impossible to the best of our knowledge.

Remark 1.6. As detailed in Section 3.2, altering the scaling matrix $\sigma_{\mathfrak{a}}$ in bounds like (1.6) is equivalent to altering the phase α; the canonical choice in (3.9) leads to several simplifications in practice.

When $a = 1$, $\mathfrak{a} = \infty$, and α is independent of q, Deshouillers–Iwaniec [10, Theorem 7] showed that the bound in (1.6) holds on average over levels $q \sim Q$ in the larger range $X \ll \max(N, Q^2/N)$. In this on-average setting, we also mention the large sieve inequality of Watt [46, Theorem 2], which saves roughly $X = Q^2/N^{3/2}$ when a_n is a smoothed divisor-type function.

For the second sequences mentioned in (1.3), we state a bound which also incorporates exponential phases $e(h_i\alpha_i)$. The reader should keep in mind the case of parameter sizes $N \asymp HL$, $H \asymp L$, and $\alpha_i = 0$, when the X-factor saved below can be as large as $\max(\sqrt{N}, \frac{q}{a\sqrt{N}})$.

Theorem 1.7 (Large sieve with dispersion coefficients). *Let $\varepsilon, X > 0$, $N \geq 1/2$, $L, H \gg 1$, $\alpha_1, \alpha_2 \in \mathbb{R}/\mathbb{Z}$, and $q, a, \ell_1, \ell_2 \in \mathbb{Z}_+$ satisfy $\ell_1, \ell_2 \asymp L$, $(\ell_1, \ell_2) = 1$. Consider the sequence $(a_n)_{n \sim N}$ given by*

$$a_n := \sum_{\substack{h_1, h_2 \in \mathbb{Z} \\ h_1\ell_1 - h_2\ell_2 = n}} \Phi_1\left(\frac{h_1}{H}\right) \Phi_2\left(\frac{h_2}{H}\right) e(h_1\alpha_1 + h_2\alpha_2),$$

where $\Phi_i : (-\infty, \infty) \to \mathbb{C}$ are smooth functions supported in $(-O(1), O(1))$, with $\Phi_i^{(j)} \ll_j 1$, $\forall j \geq 0$. Then with the notation of Theorem 1.2 and the choice of scaling matrix in (3.9), if $q \gg L^2$, one has

$$\sum_{\lambda_j < 1/4} X^{2\theta_j} \left| \sum_{n \sim N} a_n\, \rho_{j\mathfrak{a}}(an) \right|^2 \ll_\varepsilon (qaH)^\varepsilon \left(1 + \frac{aN}{q}\right) \left(\|a_n\|_2^2 + \gcd(a, q) N \left(\frac{H}{L} + \frac{H^2}{L^2} \right) \right), \tag{1.8}$$

whenever

$$X \ll \max\left(1, \frac{q}{aN}\right) \max\left(1, \frac{NH}{(H+L)LM}\right), \qquad M := \min_{\substack{t \in \mathbb{Z}_+ \\ i \in \{1,2\}}} \left(t + \frac{N}{L} \|t\alpha_i\| \right). \tag{1.9}$$

Remark 1.8. In Theorem 1.7, when $N \asymp HL$ and $\alpha_i = 0$, the norm $\|a_n\|_2^2$ is on the order of $N(\frac{H}{L} + \frac{H^2}{L^2})$. So in this setting, which is the limiting case for our applications, the right-hand side of (1.8) produces no important losses over the regular-spectrum bound of $(qN)^\varepsilon(1 + \frac{aN}{q})\, \|a_n\|_2^2$.

Remark 1.9. Some instances of the dispersion method [15, 14, 1] use coefficients roughly of the shape

$$b_n = \sum_{\substack{h \sim H \\ h(\ell_1 - \ell_2) = n}} 1, \tag{1.10}$$

where $\ell_1 \asymp \ell_2 \gg H$, $\ell_1 \neq \ell_2$, and the level is $q = \ell_1 \ell_2$. Although these resemble the second sequence from (1.3) (treated by Theorem 1.7), one should actually handle this case using Theorem 1.5, with $\alpha = 0$, $N = H$, and $a = |\ell_1 - \ell_2|$. In particular, for these ranges we have $aN = |\ell_1 - \ell_2| H \ll \ell_1 \ell_2 = q$, so the 1-term in the right-hand side of (1.6) is dominant, and the range in (1.7) becomes $X \ll \ell_1 \ell_2 / |\ell_1 - \ell_2|$.

Remark 1.10. For simplicity, we state and prove our results in the setting of arbitrary bases of classical Maass forms, following the original work of Deshouillers–Iwaniec [10]. However, our work should admit two independent extensions, which are relevant for some applications. The first is handling Maass forms with a nontrivial nebentypus, following Drappeau [14]; this leads to bounds for sums like (1.2) with c restricted to an arithmetic progression. The second is considering Hecke–Maass forms which are exceptional with respect to the Ramanujan–Petersson conjecture at finite places, the non-Archimedean analogue of Selberg's conjecture; this should improve the dependency on the scalar a when $aN > q$. One could either follow Assing–Blomer–Li [1] to 'factor out' a from $\rho_{j\mathfrak{a}}(an)$ (and apply Kim–Sarnak's bound [28] at places dividing a before using our large sieve inequalities), or treat the exceptional forms at places dividing a similarly to the Archimedean case, to match the regular-spectrum bound whenever aX is at most a function of q and N (this option should work better when a is well-factorable).

2. Informal overview

Let us summarize the key ideas behind our work, ignoring a handful of technical details such as smooth weights, GCD constraints, or keeping track of $x^{o(1)}$ factors.

2.1. Large sieve with general sequences

Let $q \in \mathbb{Z}_+$ and consider the simplified version

$$\sum_{\lambda_j < 1/4} X^{2\theta_j} \left| \sum_{n \sim N} a_n \rho_{j\infty}(n) \right|^2 \lesssim \left(1 + \frac{N}{q}\right) \|a_n\|_2^2 \tag{2.1}$$

of the large sieve inequality from Theorem 1.2, for $\mathfrak{a} = \infty$, ignoring $(qN)^{o(1)}$ factors. Here (a_n) are arbitrary complex coefficients, and the reader may pretend that $|a_n| \approx 1$ for each n, so that $\|a_n\|_2^2 \approx N$. Such an inequality follows from [10, Theorem 2] when $X = 1$, but we need larger values of X to temper the contribution of exceptional eigenvalues. The Kuznetsov trace formula [31] in Proposition 3.5, combined with large sieve inequalities for the regular spectrum [10, Theorem 2], essentially reduces the problem to bounding (a smoothed variant of) the sum

$$\sum_{\substack{c \sim NX \\ c \equiv 0 \pmod q}} \frac{1}{c} \sum_{m \sim N} \overline{a_m} \sum_{n \sim N} a_n S(m, n; c) \tag{2.2}$$

by the same amount as in the right-hand side of (2.1) – see Corollary 3.10 for a formal statement in this direction. The left-hand side vanishes for $X < q/(2N)$, so we immediately obtain (2.1) for $X \ll q/N$, which is the content of [10, Theorem 5]. Alternatively, we can plug in the pointwise Weil bound for $S(m, n; c)$ and apply Cauchy–Schwarz, to obtain an upper bound of roughly

$$\frac{NX}{q}\frac{1}{NX}N\|a_n\|_2^2\sqrt{NX} = \frac{N^{3/2}X^{1/2}}{q}\|a_n\|_2^2. \tag{2.3}$$

This is acceptable in (2.1) provided that $X \le q^2/N^3$, which completes the range from Theorem 1.2.

Improving the range $X \le \max(1, q/N, q^2/N^3)$ turns out to be quite difficult. Indeed, it is not clear how to exploit the averaging over c without the Kuznetsov formula, so any savings are more likely to come from bounding bilinear forms of Kloosterman sums $\sum_{m\sim N} a_m \sum_{n\sim N} b_n S(m,n;c)$; this is a notoriously hard problem for general sequences (a_m), (b_n) [29, 30, 27, 47]. For example, an extension of the work of Kowalski–Michel–Sawin [29] to general moduli should improve Theorem 1.2 in the critical range $q \approx N^2$, but even then the final numerical savings would be relatively small.

The other critical case encountered in applications is $q \approx N$, where Theorem 1.2 gives no nontrivial savings in the θ-aspect (i.e., $X \ll 1$), and where such savings should in fact be impossible for general sequences (a_n). Indeed, we expect $|\rho_j(n)|$ to typically be of size $\approx q^{-1/2}$, so by picking $a_n = q\overline{\rho_1(n)}$, the left-hand side of (2.1) is at least $X^{2\theta(q)}N^2$, while the right-hand side is $(1+\frac{N}{q})qN$; this limits the most optimistic savings for general sequences at $X = (1+\frac{q}{N})^{1/(2\theta(q))}$.

The key idea in our work is to make use of the special structure of the sequences (a_n) which show up in variations of the dispersion method [33]. Often, such sequences have sparse Fourier transforms, and using Fourier analysis on the corresponding exponential sums leads to a combinatorial problem.

2.2. *Exponential phases and a counting problem*

Let us focus on the case $a_n = e(n\alpha)$, for some $\alpha \in [0,1)$. Expanding the Kloosterman sums from (2.2) and Fourier-completing in m, n leads to a variant of the identity

$$\sum_{m\sim N} e(-m\alpha)\sum_{n\sim N} e(n\alpha)\,S(m,n;c) \approx N^2 \sum_{\substack{|x-c\alpha|\le c/N\\ |y+c\alpha|\le c/N}} e\left(\frac{N(x+y)}{c}\right)\mathbb{1}_{xy\equiv 1 \pmod c}. \tag{2.4}$$

Taking absolute values and ignoring the outer averaging over c, we are left with the task of bounding

$$\sum_{\substack{|x-c\alpha|\le X\\ |y+c\alpha|\le X}} \mathbb{1}_{xy\equiv 1 \pmod c}, \tag{2.5}$$

for $c \sim NX$, which is just a count of points on a modular hyperbola in short intervals (as considered in [7]). When $\alpha = 0$, one can directly use the divisor bound to write

$$\sum_{|x|,|y|\le X} \mathbb{1}_{xy\equiv 1 \pmod c} = \sum_{|z|\le \frac{X^2}{c}} \sum_{|x|,|y|\le X} \mathbb{1}_{xy=cz+1} \lesssim \frac{X^2}{c} + 1,$$

up to a factor of $X^{o(1)}$, which leads to a variant of

$$\sum_{m,n\sim N} S(m,n;c) \lesssim c + N^2 = NX + N^2.$$

(This type of bound was also observed by Shparlinski and Zhang [43].) Overall, we roughly obtain

$$\sum_{\substack{c\sim NX\\ c\equiv 0 \pmod q}} \frac{1}{c}\sum_{m,n\sim N} S(m,n;c) \lesssim \frac{NX+N^2}{q}, \tag{2.6}$$

which is at most $(1+\frac{N}{q})N$, as required in (2.1), provided that

$$X \leq \max(N, q).$$

This gives the best-case range from (1.7) (when $a = 1$). The analogue of this argument for other values of $\alpha \in \mathbb{R}/\mathbb{Z}$ depends on the quality of the best rational approximations to α, due to a rescaling trick of Cilleruelo–Garaev [7]. For an arbitrary value of α, a pigeonhole argument (Dirichlet approximation) leads to a bound of the shape

$$\sum_{\substack{c\sim NX\\ c\equiv 0 \pmod q}} \frac{1}{c}\sum_{m\sim N} e(-m\alpha)\sum_{n\sim N} e(n\alpha)\, S(m,n;c) \lesssim \frac{N^{3/2}X+N^2}{q}, \tag{2.7}$$

and ultimately to the range $X \leq \max(\sqrt{N}, q/\sqrt{N})$, which is the worst (and average) case in (1.7) when $a = 1$. Incorporating a scalar a inside $\rho_{j\infty}(an)$ is not too difficult, since a similar argument handles the analogous bilinear sums of $S(am, an; c)$, up to a loss of $\gcd(a, c)$.

Remark 2.1. A consequence of not leveraging the exponential phases in the right-hand side of (2.4) is that the same argument extends to sums over $|m|, |n| \leq N$. In particular, the term $m = n = 0$ already gives a contribution of about $c \asymp NX$, which produces a term of NX/q in (2.6) with a linear growth in X (as opposed to the square-root growth from (2.3), coming from the Weil bound).

2.3. Sequences with frequency concentration

It will probably not come as a surprise that one can extend the preceding discussion by Fourier-expanding other sequences (a_n), given a strong-enough concentration condition for their Fourier transforms, but there are some subtleties in how to do this optimally. If $a_n = \check{\mu}(n) = \int_{\mathbb{R}/\mathbb{Z}} e(n\alpha)\, d\mu(\alpha)$ for all $n \sim N$ and some bounded-variation complex measure μ, then there are at least two ways to proceed – depending on whether the integral over α is kept inside or outside of the square.

Indeed, by applying Cauchy–Schwarz in α and our Theorem 1.5 for exponential phases as a black-box, one can directly obtain a bound like

$$\sum_{\lambda_j<1/4} X^{2\theta_j}\left|\sum_{n\sim N} a_n\, \rho_{j\mathfrak{a}}(n)\right|^2 \lesssim \left(1+\frac{N}{q}\right) N|\mu|(\mathbb{R}/\mathbb{Z})^2, \tag{2.8}$$

for all $X \leq \max(\sqrt{N}, q/\sqrt{N})$ (and this range can be slightly improved given more information about the support of μ near rational numbers of small denominators). Unfortunately, this replaces the norm $\|a_n\|_2$ from Theorem 1.2 with $\sqrt{N}|\mu|(\mathbb{R}/\mathbb{Z})$, which produces a significant loss unless μ is very highly concentrated – and it is difficult to make up for this loss through gains of $X^{2\theta}$.

The alternative approach is to expand the square in the left-hand side of (2.8), pass to a sum of Kloosterman sums as in (2.2) by Kuznetsov, and only then Fourier-expand (two instances of) the sequence (a_n). Using similar combinatorial ideas as for (2.7), we can then essentially bound

$$\sum_{\substack{c\sim NX\\ c\equiv 0 \pmod q}} \frac{1}{c}\sum_{m\sim N} e(m\alpha)\sum_{n\sim N} e(n\beta)\, S(m,n;c) \lesssim \frac{N^{5/3}X+N^2}{q}, \tag{2.9}$$

for arbitrary values of $\alpha, \beta \in \mathbb{R}/\mathbb{Z}$. With no further information about the support of μ, this ultimately gives a bound like

$$\sum_{\lambda_j<1/4} X^{2\theta_j}\left|\sum_{n\sim N} a_n\,\rho_{j\mathfrak{a}}(n)\right|^2 \lesssim \left(1+\frac{N}{q}\right)\|a_n\|_2^2 + \frac{N^{5/3}X+N^2}{q}|\mu|(\mathbb{R}/\mathbb{Z})^2,$$

which is acceptable in (2.1), in particular, whenever $X < N^{1/3}$ and $\sqrt{N}|\mu|(\mathbb{R}/\mathbb{Z}) \le \sqrt{q/N}\|a_n\|_2$. Compared to the first approach, this generally gains less in the X-aspect, but it relaxes the concentration condition on μ if $N < q$. This second approach turns out to be better for our applications; the resulting large sieve inequality is Theorem 5.2, which particularizes to Theorems 1.5 and 1.7.

What is perhaps more surprising, though, is that strong-enough frequency concentration (i.e., $\sqrt{N}|\mu|(\mathbb{R}/\mathbb{Z})^2 \le \sqrt{q/N}\|a_n\|_2$) arises in applications, beyond the case of exponential sequences. A key observation is that the aforementioned dispersion coefficients

$$a_n = \sum_{\substack{h_1,h_2\sim H\\ h_1\ell_1-h_2\ell_2=n}} 1, \tag{2.10}$$

with $\ell_1 \asymp \ell_2 \asymp L$, come from a convolution of two "arithmetic progressions" $\mathbb{1}_{n\equiv 0 \pmod{\ell_i}}\mathbb{1}_{n\sim H\ell_i}$. The Fourier transform of each of these two sequences has ℓ_i periodic peaks of height H and width $(H\ell_i)^{-1}$, supported around multiples of $1/\ell_i$. When $(\ell_1,\ell_2)=1$, multiplying these two Fourier transforms results in cancellation everywhere away from a small number ($\le 1+\frac{L}{H}$) of rational points (and thus, in frequency concentration on a set of size $\frac{1}{HL}+\frac{1}{H^2}$); see Lemma 4.11.

2.4. Multilinear forms of Kloosterman sums

Consider once again the sums (1.2), in the ranges

$$M,N \le rs, \qquad X := \frac{s\sqrt{r}C}{\sqrt{MN}} \ge 1,$$

which are relevant for most applications. An additional use of the Kuznetsov formula, for the level $q=rs$ and the cusps $\infty, 1/s$ (with suitable scaling matrices), gives a variant of the bound

$$\begin{aligned}&\sum_{m\sim M} a_m \sum_{n\sim N} b_n \sum_{(c,r)=1} g\left(\frac{c}{C}\right) S(m\bar{r},\pm n; sc)\\ &\lesssim s\sqrt{r}C \sum_{\lambda_j<1/4} X^{2\theta_j}\left|\sum_{m\sim M} a_m\,\rho_{j\infty}(m)\right|\left|\sum_{n\sim N} b_n\,\rho_{j\,1/s}(n)\right| + \dots\end{aligned}$$

Here we omitted the contribution of the regular Maass forms, Eisenstein series and holomorphic forms (which will not be dominant). A priori, this arrangement introduces a factor of $X^{2\theta(q)}$ in our bounds, recalling that $\theta(q) = \max_{\lambda_j(q)<1/4}\theta_j(q)$ (if the maximum is nonempty, and $\theta(q)=0$ otherwise). However, the value of X in this loss can be decreased through the large sieve inequalities for exceptional Maass forms. Indeed, after splitting $X = X_0\sqrt{X_1X_2}$, taking out a factor of only $(1+X_0)^{2\theta(q)}$, and applying Cauchy–Schwarz, we reach

$$s\sqrt{r}C\,(1+X_0)^{2\theta(q)}\left(\sum_{\lambda_j<1/4} X_1^{2\theta_j}\left|\sum_{m\sim M} a_m\,\rho_{j\infty}(m)\right|^2\right)^{1/2}\left(\sum_{\lambda_j<1/4} X_2^{2\theta_j}\left|\sum_{n\sim N} b_n\,\rho_{j\,1/s}(n)\right|^2\right)^{1/2}.$$

Above, we can choose X_1 and X_2 as the maximal values that can be fully incorporated in large sieve inequalities like (2.1) without producing losses in the right-hand side, for the specific sequences (a_m)

and (b_n). In this case, we roughly obtain a final bound of

$$s\sqrt{r}C\left(1+\frac{s\sqrt{r}C}{\sqrt{MNX_1X_2}}\right)^{2\theta(q)}\|a_m\|_2\,\|b_n\|_2.$$

For example, if $a_m = e(m\alpha_{r,s})$ for some $\alpha_{r,s} \in \mathbb{R}/\mathbb{Z}$, then we may take $X_1 = \max(\sqrt{N}, q/\sqrt{N})$ by Theorem 1.5, which ultimately saves a factor of $N^{\theta/2}$. Similarly, if (b_n) are of the form in (2.10), where $H \asymp L \asymp \sqrt{N}$, then by Theorem 1.7 we may also take $X_2 = \max(\sqrt{N}, q/\sqrt{N})$.

If some averaging over $r \sim R, s \sim S$ is available and the sequence (a_m) does not depend on r, s, then larger values of X_1 are available due to Deshouillers–Iwaniec [10, Theorems 6, 7]. In this setting, if $a_m = e(m\omega)$ for a fixed $\omega \in \mathbb{R}/\mathbb{Z}$, one can combine the essentially-optimal value $X_1 = Q^2/N$ (see Theorem 3.11 below) with our savings in the X_2-aspect. Following [10, Theorem 12], similar estimates can be deduced for multilinear forms of incomplete Kloosterman sums, simply by Fourier-completing them and appealing to the estimates for complete sums; see our Corollary 5.14. Such bounds feed directly into the dispersion method and its applications, as we shall see in Section 6.

2.5. *Layout of paper*

In Section 3, we cover notation and preliminary results, including several key lemmas from the spectral theory of automorphic forms. Section 4 only contains elementary arguments, from counting points on modular hyperbolas in Lemma 4.4 (following Cilleruelo–Garaev [7]), to the bilinear Kloosterman bounds in Proposition 4.9 (which may be of independent interest to the reader). In Section 5.1, we combine these combinatorial inputs with the Deshouillers–Iwaniec setup [10] to prove a general large sieve inequality in Theorem 5.2, which can be viewed as our main technical result; we then deduce Theorems 1.5 and 1.7 from it. Section 5.3 contains the corollaries of these large sieve inequalities: various bounds for multilinear forms of Kloosterman sums, with improved dependencies on the θ parameter. Finally, in Section 6 we will use these bounds to prove Theorem 1.1, building on the work of Merikoski [37] and de la Bretèche–Drappeau [8].

3. Notation and preliminaries

3.1. *Standard analytic notation*

We write $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{H}$ for the sets of integers, rational numbers, real numbers, complex numbers, respectively complex numbers with positive imaginary part. We may scale these sets by constants, and may add the subscript + to restrict to positive numbers; so for example $2\mathbb{Z}_+$ denotes the set of even positive integers, while $i\mathbb{R}$ is the imaginary line. For $\alpha \in \mathbb{R}$ (or $\mathbb{R}/\mathbb{Z}$), we denote $e(\alpha) := \exp(2\pi i\alpha)$, and set $\|\alpha\| := \min_{n\in\mathbb{Z}} |\alpha - n|$, which induces a metric on $\mathbb{R}/\mathbb{Z}$. We write $\mathbb{Z}/c\mathbb{Z}$ for the ring of residue classes modulo a positive integer c, $(\mathbb{Z}/c\mathbb{Z})^\times$ for its multiplicative group of units, and $\overline{x}$ for the inverse of $x \in (\mathbb{Z}/c\mathbb{Z})^\times$. We may use the latter notation inside congruences, with $x \equiv y\overline{z} \pmod c$ meaning that $xz \equiv y \pmod c$ (for $\gcd(z, c) = 1$). We may also use the notation (a, b) for $\gcd(a, b)$, and $[a, b]$ for $\operatorname{lcm}(a, b)$, when it is clear from context to not interpret these as pairs or intervals. We write $\mathbb{1}_S$ for the indicator function of a set S (or for the truth value of a statement S), $n \sim N$ for the statement that $N < n \le 2N$ (so, e.g., $\mathbb{1}_{n\sim N} = \mathbb{1}_{n\le 2N} - \mathbb{1}_{n\le N}$), and interpret sums like $\sum_{n\sim N}$, $\sum_{n\equiv 0 \pmod q}$, or $\sum_{d|n}$ with the implied restrictions that $n, d \in \mathbb{Z}_+$. For $n \in \mathbb{Z}_+$, we define the divisor-counting function by $\tau(n) := \sum_{d|n} 1$, and Euler's totient function by $\varphi(n) := \sum_{m=1}^{n} \mathbb{1}_{(m,n)=1}$. We say that a complex sequence (a_n) is *divisor-bounded* iff $|a_n| \ll \tau(n)^{O(1)}$. We also write $P^+(n)$ and $P^-(n)$ for the largest and smallest prime factors of a positive integer n, and recall that n is called y-smooth iff $P^+(n) \le y$.

We use the standard asymptotic notation $f \ll g$, $f \asymp g$, $f = O(g)$, $f = o_{x\to\infty}(g)$ from analytic number theory, and indicate that the implicit constants depend on some parameter ε through subscripts (e.g., $f \ll_\varepsilon g$, $f = O_\varepsilon(g)$). In particular, one should read bounds like $f(x) \ll x^{o(1)}$ as $\forall \varepsilon > 0$, $f(x) \ll_\varepsilon x^\varepsilon$. Given $\ell \in \mathbb{Z}_+$, we write $f^{(\ell)}$ for the ℓth derivative of a function $f : \mathbb{R} \to \mathbb{C}$, and $f^{(0)} = f$.

For $q \in [1,\infty]$, we denote by $\|f\|_{L^q}$ the L^q-norm of a function $f : \mathbb{R} \to \mathbb{C}$ (or $f : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$), and by $\|a\|_q$ (or $\|a_n\|_q$) the ℓ^q norm of a sequence (a_n).

We require multiple notations for the Fourier transforms of L^1 functions $f, \Phi : \mathbb{R} \to \mathbb{C}$, $\varphi : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$, and $a : \mathbb{Z} \to \mathbb{C}$ (the latter could be, e.g., a finite sequence $(a_n)_{n\sim N}$ extended with zeroes elsewhere). These are given by

$$\begin{aligned}
f : \mathbb{R} \to \mathbb{C} \quad &\rightsquigarrow \quad \widehat{f} : \mathbb{C} \to \mathbb{C}, \qquad \widehat{f}(\xi) := \int_{\mathbb{R}} f(t)\, e(-\xi t)\, dt, \\
\Phi : \mathbb{R} \to \mathbb{C} \quad &\rightsquigarrow \quad \check{\Phi} : \mathbb{C} \to \mathbb{C}, \qquad \check{\Phi}(t) := \int_{\mathbb{R}} \Phi(\xi)\, e(\xi t)\, d\xi, \\
a : \mathbb{Z} \to \mathbb{C} \quad &\rightsquigarrow \quad \widehat{a} : \mathbb{R}/\mathbb{Z} \to \mathbb{C}, \qquad \widehat{a}(\alpha) := \sum_{n\in\mathbb{Z}} a_n\, e(-n\alpha), \\
\varphi : \mathbb{R}/\mathbb{Z} \to \mathbb{C} \quad &\rightsquigarrow \quad \check{\varphi} : \mathbb{Z} \to \mathbb{C}, \qquad \check{\varphi}(n) := \int_{\mathbb{R}/\mathbb{Z}} \varphi(\alpha)\, e(n\alpha)\, d\alpha.
\end{aligned} \tag{3.1}$$

Note that the first two and the last two of these transforms are inverse operations under suitable conditions; in particular, if Φ is Schwarz, a is L^1, and φ is smooth (so $\check{\varphi}(n)$ decays rapidly as $|n| \to \infty$), one has

$$\check{\widehat{\Phi}}\Big|_{\mathbb{R}} = \widehat{\check{\Phi}}\Big|_{\mathbb{R}} = \Phi, \qquad \check{\widehat{a}} = a, \qquad \widehat{\check{\varphi}} = \varphi. \tag{3.2}$$

We also denote the Fourier transform of a bounded-variation complex Borel measure μ on $\mathbb{R}/\mathbb{Z}$ by $\check{\mu}(n) := \int_{\mathbb{R}/\mathbb{Z}} e(n\alpha)\, d\mu(\alpha)$. For instance, one has $\check{\lambda}(n) = \mathbb{1}_{n=0}$ for the Lebesgue measure λ, and $\check{\delta}_A(n) = \sum_{\alpha\in A} e(n\alpha)$ for the Dirac delta measure on a finite set $A \subset \mathbb{R}/\mathbb{Z}$. Moreover, if $d\mu = \varphi\, d\lambda$ for some L^1 function $\varphi : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$, then $\check{\mu} = \check{\varphi}$. Finally, with our notation, the Parseval–Plancherel identity reads $\|a_n\|_2^2 = \|\widehat{a}\|_{L^2}^2$ (and $\|f\|_{L^2} = \|\widehat{f}\|_{L^2}$), while Poisson summation states that for any Schwarz function f,

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \widehat{f}(n) = \sum_{n\in\mathbb{Z}} \check{f}(n). \tag{3.3}$$

In practice it will be useful to truncate the Poisson summation formula; we combine this with a smooth dyadic partition of unity and a separation of variables, in the following lemma.

Lemma 3.1 (Truncated Poisson with extra steps). *Let $x, N, Q \gg 1$ with $N, Q \ll x^{O(1)}$, $q \asymp Q$ be a positive integer, $a \in \mathbb{Z}$ (or $\mathbb{Z}/q\mathbb{Z}$), and $\Phi : (0,\infty) \to \mathbb{C}$ be a smooth function with $\Phi(t)$ supported in $t \asymp 1$ and $\Phi^{(j)} \ll_j 1$ for $j \geq 0$. Then for any $A, \delta > 0$ and $H := x^\delta N^{-1} Q$, one has*

$$\begin{aligned}
\sum_{n\equiv a \pmod q} \Phi\left(\frac{n}{N}\right) = \frac{N}{q}\widehat{\Phi}(0) + O_{A,\delta}\left(x^{-A}\right) \\
+ \frac{N}{Q}\int \sum_{\substack{H_j = 2^j \\ 1\leq H_j \leq H}} \sum_{\frac{1}{2}H_j \leq |h| \leq 2H_j} c_{j,u}(h)\, \Phi\left(\frac{uq}{Q}\right) e\left(\frac{ah}{q}\right) du,
\end{aligned}$$

where the support of the integral in u is bounded, and

$$c_{j,u}(h) := \Psi_j\left(\frac{|h|}{H_j}\right) e\left(-h\frac{uN}{Q}\right), \tag{3.4}$$

for some compactly-supported smooth functions $\Psi_j : (\frac{1}{2}, 2) \to \mathbb{C}$ with $\Psi_j^{(k)} \ll_k 1$ for $k \geq 0$.

Proof. The Poisson identity (3.3) with a change of variables yields

$$\sum_{n\equiv a \pmod q} \Phi\left(\frac{n}{N}\right) = \frac{N}{q}\sum_{h\in\mathbb{Z}} \widehat{\Phi}\left(\frac{hN}{q}\right) e\left(\frac{ah}{q}\right).$$

We take out the main term at $h = 0$, put $|h| \geq 1$ in dyadic ranges via a smooth partition of unity

$$\mathbb{1}_{\mathbb{Z}_+}(|h|) = \mathbb{1}_{\mathbb{Z}_+}(|h|) \sum_{H_j = 2^j \geq 1} \Psi_j\left(\frac{|h|}{H_j}\right),$$

and bound the contribution of $H_j > H = x^{\delta} N^{-1} Q$ by $O_{A,\delta}(x^{-A})$ using the Schwarz decay of Φ. In the remaining sum

$$\frac{N}{q} \sum_{\substack{H_j = 2^j \\ 1 \leq H_j \leq H}} \sum_{\frac{1}{2}H_j \leq |h| \leq 2H_j} \Psi_j\left(\frac{|h|}{H_j}\right) \widehat{\Phi}\left(\frac{hN}{q}\right) e\left(\frac{ah}{q}\right),$$

we separate the h, q variables via the Fourier integral

$$\widehat{\Phi}\left(\frac{hN}{q}\right) = \int \Phi(t)\, e\left(-h\frac{tN}{q}\right) dt = \frac{q}{Q}\int \Phi\left(\frac{uq}{Q}\right) e\left(-h\frac{uN}{Q}\right) du,$$

where we let $t = uq/Q$. Swapping the (finite) sums with the integral completes our proof. □

We also highlight the nonstandard Notation 4.1, pertaining to rational approximations. Further analytic notation specific to each section is described therein (see, for example, Notations 5.1, 6.2 and 6.7). For the rest of this section, we recount the main concepts relevant to bounding sums of Kloosterman sums via the Kuznetsov trace formula, mostly to clarify our notation (in particular, to point out small changes to the notation in [10]), and to explicitate a few useful lemmas.

3.2. Cusps, automorphic forms, Kloosterman sums

Recall that $\mathrm{PSL}_2(\mathbb{R}) := \mathrm{SL}_2(\mathbb{R})/\{\pm 1\}$ acts naturally on $\mathbb{C}\cup\{\infty\}$ by $\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right) z := \frac{az+b}{cz+d}$. For $q \in \mathbb{Z}_+$, we denote by $\Gamma_0(q)$ the modular subgroup of level q, consisting of those matrices $\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right) \in \mathrm{PSL}_2(\mathbb{Z})$ with $c \equiv 0 \pmod q$. A number $\mathfrak{a} \in \mathbb{C}\cup\{\infty\}$ is called a *cusp* of $\Gamma_0(q)$ iff it is the unique fixed point of some $\sigma \in \Gamma_0(q)$; we write $\Gamma_{\mathfrak{a}} := \{\sigma \in \Gamma_0(q) : \sigma\mathfrak{a} = \mathfrak{a}\}$ for the stabilizer of $\mathfrak{a}$ inside $\Gamma_0(q)$. Two cusps are *equivalent* iff they lie in the same orbit of $\Gamma_0(q)$; the corresponding stabilizers are then conjugate inside $\Gamma_0(q)$. By [10, Lemma 2.3], the fractions

$$\left\{\frac{u}{w} \ : \ u, w \in \mathbb{Z}_+,\ (u,w) = 1,\ w \mid q,\ u \leq \gcd\left(w, \frac{q}{w}\right)\right\} \tag{3.5}$$

form a maximal set of inequivalent cusps of $\Gamma_0(q)$. Following [10, (1.1)], given a cusp $\mathfrak{a}$ of $\Gamma_0(q)$ and its equivalent representative u/w from (3.5), we denote

$$\mu(\mathfrak{a}) := \frac{\gcd\left(w, \frac{q}{w}\right)}{q}, \tag{3.6}$$

(Like most of our notation involving cusps, this implicitly depends on the level q as well.) In particular, the cusp at ∞ of $\Gamma_0(q)$ is equivalent to the fraction $1/q$, so we have $\mu(\infty) = q^{-1}$. More generally, we have $\mu(1/s) = q^{-1}$ whenever $q = rs$ with $\gcd(r,s) = 1$, and it is these cusps which account for most applications to sums of Kloosterman sums; thus for simplicity, we restrict all of our main results to

cusps with $\mu(\mathfrak{a}) = q^{-1}$. Following [10, (1.2)], a *scaling matrix* $\sigma_{\mathfrak{a}}$ for a cusp $\mathfrak{a}$ is an element of $\mathrm{PSL}_2(\mathbb{R})$ such that

$$\sigma_{\mathfrak{a}}\infty = \mathfrak{a} \qquad \text{and} \qquad \sigma_{\mathfrak{a}}^{-1}\Gamma_{\mathfrak{a}}\sigma_{\mathfrak{a}} = \Gamma_\infty = \left\{ \left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right) : n \in \mathbb{Z} \right\}. \tag{3.7}$$

Scaling matrices will allow us to expand $\Gamma_{\mathfrak{a}}$-invariant functions $f : \mathbb{H} \to \mathbb{C}$ as Fourier series around the cusp $\mathfrak{a}$, via the change of coordinates $z \leftarrow \sigma_{\mathfrak{a}} z$ (note that if f is $\Gamma_{\mathfrak{a}}$-invariant, then $z \mapsto f(\sigma_{\mathfrak{a}} z)$ is Γ_∞-invariant). For a given cusp $\mathfrak{a}$, the choice of $\sigma_{\mathfrak{a}}$ can only vary by simple changes of coordinates

$$\widetilde{\sigma}_{\mathfrak{a}} = \sigma_{\mathfrak{a}} \left(\begin{smallmatrix} 1 & \alpha \\ 0 & 1 \end{smallmatrix}\right), \tag{3.8}$$

for $\alpha \in \mathbb{R}$ (which result in multiplying the Fourier coefficients by exponential phases $e(n\alpha)$). When $\mu(\mathfrak{a}) = q^{-1}$, we must have $\mathfrak{a} = \tau(1/s)$ for some $\tau \in \Gamma_0(q)$ and $rs = q$ with $(r, s) = 1$; in this case, inspired by Watt [46, p. 195], we will use the canonical choice of scaling matrix

$$\sigma_{\mathfrak{a}} = \tau \cdot \begin{pmatrix} \sqrt{r} & -\overline{s}/\sqrt{r} \\ s\sqrt{r} & \overline{r}\sqrt{r} \end{pmatrix}, \tag{3.9}$$

where $\overline{r}, \overline{s}$ are integers such that $r\overline{r} + s\overline{s} = 1$ (for definiteness, let us say we pick $\overline{s} \geq 0$ to be minimal). This is different from the choice in [10, (2.3)], and leads to the simplification of certain extraneous exponential phases. For the cusp $\mathfrak{a} = \infty = \left(\begin{smallmatrix} 1 & 0 \\ -q & 1 \end{smallmatrix}\right)(1/q)$, (3.9) reduces back to the identity matrix.

We refer the reader to the aforementioned work of Deshouillers–Iwaniec [10] for a brief introduction to the classical spectral theory of GL_2 automorphic forms, to [25, 24, 26] for a deeper dive into this topic, to [14, 45, 15, 8, 32, 39] for follow-up works and optimizations, and to [6] for the modern viewpoint of automorphic representations. For our purposes, an *automorphic form* of level q and integer weight $k \geq 0$ will be a smooth function $f : \mathbb{H} \to \mathbb{C}$ satisfying the transformation law

$$f(\sigma z) = j(\sigma, z)^k f(z) \qquad \forall \sigma \in \Gamma_0(q), \qquad \text{where} \qquad j\left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), z\right) := cz + d.$$

as well as moderate (at-most-polynomial) growth conditions near every cusp. We say that f is *square-integrable* iff $\langle f, f\rangle_k < \infty$, where $\langle f, g\rangle_k := \iint_{\Gamma_0(q)\backslash\mathbb{H}} f(x+iy)\,\overline{g(x+iy)}\,y^{k-2}\,dx\,dy$ is the Petersson inner product. We denote by $L^2(\Gamma_0(q)\backslash\mathbb{H}, k)$ the space of square-integrable automorphic forms of level q and weight k; when we drop the dependency on k, it should be understood that $k = 0$. Finally, we call f a *cusp form* iff it is square-integrable and vanishes at all cusps.

Kloosterman sums show up in the Fourier coefficients of *Poincaré series*, which are useful in detecting the Fourier coefficients of other automorphic forms via inner products (see [10, (1.8), (1.18)]). In fact, by Fourier expanding a Poincaré series corresponding to a cusp $\mathfrak{a}$ around another cusp $\mathfrak{b}$, one is led to a more general family of Kloosterman-type sums, depending on both $\mathfrak{a}$ and $\mathfrak{b}$.

More specifically (following [10, (1.3)], [14, §4.1.1], [24]), given two cusps $\mathfrak{a}, \mathfrak{b}$ of $\Gamma_0(q)$, we first let

$$\mathcal{C}_{\mathfrak{a}\mathfrak{b}} := \left\{ c \in \mathbb{R}_+ : \exists\, a, b, d \in \mathbb{R}, \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \sigma_{\mathfrak{a}}^{-1}\Gamma_0(q)\,\sigma_{\mathfrak{b}} \right\}.$$

Here $\sigma_{\mathfrak{a}}$ and $\sigma_{\mathfrak{b}}$ are arbitrary scaling matrices for $\mathfrak{a}$ and $\mathfrak{b}$, but the set $\mathcal{C}_{\mathfrak{a}\mathfrak{b}}$ actually depends only on $\mathfrak{a}$ and $\mathfrak{b}$ (since multiplication by matrices $\left(\begin{smallmatrix} 1 & \alpha \\ 0 & 1 \end{smallmatrix}\right)$ does not affect the bottom-left entry). Then we let

$$\mathcal{D}_{\mathfrak{a}\mathfrak{b}}(c) := \left\{ \widetilde{d} \in \mathbb{R}/c\mathbb{Z} : \exists\, a, b \in \mathbb{R},\ d \in \widetilde{d}, \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \sigma_{\mathfrak{a}}^{-1}\Gamma_0(q)\,\sigma_{\mathfrak{b}} \right\},$$

for any $c \in \mathbb{R}_+$ (although this is only nonempty when $c \in \mathcal{C}_{\mathfrak{a}\mathfrak{b}}$). By this definition, the set $\mathcal{D}_{\mathfrak{a}\mathfrak{b}}(c)$ is finite, does not depend on $\sigma_{\mathfrak{a}}$, and only depends on $\sigma_{\mathfrak{b}}$ up to translations. It turns out that a given $\widetilde{d} \in \mathcal{D}_{\mathfrak{a}\mathfrak{b}}(c)$ uniquely determines the value of $\widetilde{a} \in \mathbb{R}/c\mathbb{Z}$ such that $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \sigma_{\mathfrak{a}}^{-1}\Gamma_0(q)\,\sigma_{\mathfrak{b}}$ for some $a \in \widetilde{a}$, $d \in \widetilde{d}$ (see [10, p. 239]). Symmetrically, this $\widetilde{a}$ does not depend on $\sigma_{\mathfrak{b}}$, and only depends on $\sigma_{\mathfrak{a}}$ up to translations.

Thus given $c \in \mathbb{R}_+$ and $m, n \in \mathbb{Z}$, it makes sense to define

$$S_{\mathfrak{a}\mathfrak{b}}(m,n;c) := \sum_{\widetilde{d} \in \mathcal{D}_{\mathfrak{a}\mathfrak{b}}(c)} e\left(\frac{m\widetilde{a} + n\widetilde{d}}{c}\right), \tag{3.10}$$

where $\widetilde{a}$ and $\widetilde{d}$ are corresponding values mod c; note that this vanishes unless $c \in \mathcal{C}_{\mathfrak{a}\mathfrak{b}}$. Since varying the choices of $\sigma_{\mathfrak{a}}$ and $\sigma_{\mathfrak{b}}$ has the effect of uniformly translating $\widetilde{a}$, respectively $\widetilde{d}$, it follows that $S_{\mathfrak{a}\mathfrak{b}}(m,n;c)$ only depends on $\sigma_{\mathfrak{a}}, \sigma_{\mathfrak{b}}$ up to multiplication by exponential phases $e(m\alpha), e(n\beta)$. In fact, the same holds true when varying $\mathfrak{a}$ and $\mathfrak{b}$ in equivalence classes of cusps [10, p. 239]. We also note the symmetries

$$S_{\mathfrak{a}\mathfrak{b}}(m,-n;c) = \overline{S_{\mathfrak{a}\mathfrak{b}}(-m,n;c)}, \qquad S_{\mathfrak{a}\mathfrak{b}}(m,n;c) = \overline{S_{\mathfrak{b}\mathfrak{a}}(n,m;c)}, \tag{3.11}$$

the second one following from the fact that

$$\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \sigma_{\mathfrak{a}}^{-1}\Gamma_0(q)\,\sigma_{\mathfrak{b}} \qquad \Longleftrightarrow \qquad \left(\begin{smallmatrix} -d & b \\ c & -a \end{smallmatrix}\right) = -\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)^{-1} \in \sigma_{\mathfrak{b}}^{-1}\Gamma_0(q)\,\sigma_{\mathfrak{a}}.$$

Let us now relate these sums to the classical Kloosterman sums from (1.1).

Lemma 3.2 (Explicit Kloosterman sums [46]). *Let $q = rs$ with $r, s \in \mathbb{Z}_+$, $\gcd(r,s) = 1$. Then for any $c \in \mathbb{R}_+$ and $m, n \in \mathbb{Z}$, with the choice of scaling matrices from (3.9), one has*

$$S_{\infty\, 1/s}\left(m,n;s\sqrt{r}c\right) = \mathbb{1}_{c \in \mathbb{Z}_+,\,(c,r)=1}\, S(m\overline{r},n;sc). \tag{3.12}$$

Moreover, let $\mathfrak{a}$ be any cusp of $\Gamma_0(q)$ with $\mu(\mathfrak{a}) = q^{-1}$, and $\sigma_{\mathfrak{a}}$ be as in (3.9). Then one has

$$S_{\mathfrak{a}\mathfrak{a}}(m,n;c) = \mathbb{1}_{c \in q\mathbb{Z}_+}\, S(m,n;c). \tag{3.13}$$

Varying the choice of scaling matrix as in (3.8) would result in an additional factor of $e((n-m)\alpha)$.

Proof. These identities are precisely [46, (3.5) and (3.4)], at least when $\mathfrak{a} = 1/s$ for some $rs = q$, with $(r,s) = 1$. For a general cusp with $\mu(\mathfrak{a}) = q^{-1}$, we have $\mathfrak{a} = \tau(1/s)$ for some $\tau \in \Gamma_0(q)$, but the presence of τ in the scaling matrix from (3.9) does not affect the set $\sigma_{\mathfrak{a}}^{-1}\Gamma_0(q)\,\sigma_{\mathfrak{a}}$, nor the generalized Kloosterman sum $S_{\mathfrak{a}\mathfrak{a}}(m,n;c)$. For explicit computations of this type, see [10, §2]. □

3.3. The Kuznetsov formula and exceptional eigenvalues

We now recognize some important classes of GL_2 automorphic forms of level q:

(1). *Classical modular forms*, which are holomorphic with removable singularities at all cusps, and can only have even weights $k \in 2\mathbb{Z}_+$ (except for the zero form). A *holomorphic cusp form* f additionally vanishes at all cusps; such forms have Fourier expansions

$$j(\sigma_{\mathfrak{a}},z)^{-k} f(\sigma_{\mathfrak{a}} z) = \sum_{n=1}^{\infty} \psi_{\mathfrak{a}}(n)\, e(nz) \tag{3.14}$$

around each cusp $\mathfrak{a}$ of $\Gamma_0(q)$ (see [10, (1.7)]). We mention that the space of holomorphic cusp forms of weight k is finite-dimensional, and denote its dimension by $h_k = h_k(q)$.

(2). *Maass forms* (of weight 0), which are invariant under the action of $\Gamma_0(q)$, and are eigenfunctions of the hyperbolic Laplacian $\Delta = -y^2\left(\partial_x^2 + \partial_y^2\right)$. These include:

   (a). Maass *cusp* forms, which additionally vanish at all cusps and are square-integrable. These (plus the constant functions) correspond to the discrete spectrum of the hyperbolic Laplacian on $L^2\left(\Gamma_0(q)\backslash\mathbb{H}\right)$, consisting of eigenvalues $0 = \lambda_0 < \lambda_1 \le \lambda_2 \le \lambda_3 \le \dots$ with

no limit point. Around a given cusp $\mathfrak{a}$, Maass cusp forms have Fourier expansions (see [10, (1.15)])

$$u(\sigma_{\mathfrak{a}} z) = y^{1/2} \sum_{n \neq 0} \rho_{\mathfrak{a}}(n) K_{i\kappa}(2\pi |n| y)\, e(mx), \tag{3.15}$$

where $z = x + iy$ and K is the Whittaker function normalized as in [10, p. 264].

(b). *Eisenstein series*, explicitly defined by $E_{\mathfrak{a}}(z; s) := \sum_{\tau \in \Gamma_{\mathfrak{a}} \backslash \Gamma_0(q)} \mathrm{Im}^s \left(\sigma_{\mathfrak{a}}^{-1} \tau z\right)$ for $\mathrm{Re}\, s > 1$, and meromorphically continued to $s \in \mathbb{C}$. Although not square-integrable themselves, "incomplete" versions of Eisenstein series with $s = \frac{1}{2} + ir$ (and $r \in \mathbb{R}$) can be used to describe the orthogonal complement in $L^2(\Gamma_0(q) \backslash \mathbb{H})$ of the space of Maass cusp forms, corresponding to the continuous spectrum of the hyperbolic Laplacian. Sharing similarities with both Maass cusp forms and Poincaré series, the Eisenstein series $E_{\mathfrak{a}}$ have Fourier expansions [10, (1.17)] around any cusp $\mathfrak{b}$, involving the Whittaker function and the Kloosterman-resembling coefficients (for $n \in \mathbb{Z}$, $n \neq 0$)

$$\varphi_{\mathfrak{a}\mathfrak{b}}(n; s) := \sum_{c \in \mathcal{C}_{\mathfrak{a}\mathfrak{b}}} c^{-2s} \sum_{\widetilde{d} \in \mathcal{D}_{\mathfrak{a}\mathfrak{b}}(c)} e\left(\frac{n\widetilde{d}}{c}\right). \tag{3.16}$$

We are particularly interested in the *exceptional* Maass cusp forms, which have eigenvalues $\lambda_j \in (0, 1/4)$; there can only be finitely many such forms of each level q, and Selberg conjectured [41] that there are none. With implicit dependencies on q, we denote

$$\kappa_j^2 := \lambda_j - \frac{1}{4} \qquad \text{and} \qquad \theta_j := i\kappa_j, \tag{3.17}$$

where κ_j is chosen such that $i\kappa_j > 0$ or $\kappa_j \geq 0$; thus exceptional forms correspond to imaginary values of κ_j and positive values of θ_j. Letting

$$\theta(q) := \sqrt{\max\left(0, \tfrac{1}{4} - \lambda_1(q)\right)} = \begin{cases} \theta_1(q), & \theta_1(q) > 0 \\ 0, & \text{otherwise.} \end{cases}, \qquad \theta_{\max} := \sup_{q \geq 1} \theta(q),$$

Selberg's eigenvalue conjecture asserts that $\theta_{\max} = 0$, and the best result towards it is due to Kim–Sarnak [28, Appendix 2]. This deep unconditional result requires the theory of GL_n automorphic representations [6], but it is a very useful black-box input to spectral methods, where various bounds have exponential dependencies on θ.

Theorem 3.3 (Kim–Sarnak's eigenvalue bound [28]). *One has* $\theta_{\max} \leq \frac{7}{64}$.

Based on earlier work of Kuznetsov [31], Deshouillers–Iwaniec [10] developed a trace formula relating weighted sums over c of the generalized Kloosterman sums from (3.10) to (sums of products of) the Fourier coefficients of holomorphic cusp forms, Maass cusp forms, and Eisenstein series, around any two cusps $\mathfrak{a}, \mathfrak{b}$ of $\Gamma_0(q)$. Roughly speaking, this follows by summing two applications of Parseval's identity for the aforementioned Poincaré series: one in the space of holomorphic cusp forms (summing over all weights $k \in 2\mathbb{Z}_+$), and one in the space $L^2(\Gamma_0(q)/\mathbb{H})$ of square-integrable automorphic forms of weight 0, via the spectral decomposition of the hyperbolic Laplacian (leading to the terms from Maass cusp forms and Eisenstein series).

One can arrange the resulting *Kuznetsov trace formula* so that the Kloosterman sums in the left-hand side are weighted by an arbitrary compactly-supported smooth function φ; in the right-hand side, the Fourier coefficients of automorphic forms are consequently weighted by *Bessel transforms* of φ, defined

for $r \in \mathbb{R} \setminus \{0\}$ by

$$\begin{aligned}\widetilde{\mathcal{B}}_\varphi(r) &:= \int_0^\infty J_r(y)\,\varphi(y)\frac{dy}{y},\\ \widehat{\mathcal{B}}_\varphi(r) &:= \frac{\pi}{\sinh(\pi r)}\int_0^\infty \frac{J_{2ir}(x)-J_{-2ir}(x)}{2i}\varphi(x)\frac{dx}{x},\\ \check{\mathcal{B}}_\varphi(r) &:= \frac{4}{\pi}\cosh(\pi r)\int_0^\infty K_{2ir}(x)\,\varphi(x)\frac{dx}{x},\end{aligned} \tag{3.18}$$

where K_{it} is the aforementioned Whittaker function, and the Bessel functions J_ℓ, J_{it} are defined as in [10, p. 264–265] (above we slightly departed from the notation in [10, 14], to avoid confusion with Fourier transforms). All we will need to know about these transforms are the following bounds.

Lemma 3.4 (Bessel transform bounds [10]). *Let $Y > 0$ and $\varphi : \mathbb{R} \to \mathbb{C}$ be a smooth function with compact support in $y \asymp Y$, satisfying $\varphi^{(j)}(y) \ll_j Y^{-j}$ for $j \geq 0$. Then one has*

$$\widehat{\mathcal{B}}_\varphi(ir),\ \check{\mathcal{B}}_\varphi(ir) \ll \frac{1+Y^{-2r}}{1+Y}, \qquad \text{for } 0 < r < \frac{1}{2}, \tag{3.19}$$

$$\widetilde{\mathcal{B}}_\varphi(r),\ \widehat{\mathcal{B}}_\varphi(r),\ \check{\mathcal{B}}_\varphi(r) \ll \frac{1+|\log Y|}{1+Y}, \qquad \text{for } r \in \mathbb{R}\setminus\{0\}, \tag{3.20}$$

$$\widetilde{\mathcal{B}}_\varphi(r),\ \widehat{\mathcal{B}}_\varphi(r),\ \check{\mathcal{B}}_\varphi(r) \ll |r|^{-5/2} + |r|^{-3}Y, \qquad \text{for } r \in \mathbb{R},\ |r| \geq 1, \tag{3.21}$$

Moreover, if φ is nonnegative with $\int \varphi(y)\,dy \gg Y$, and $Y < c$ for some constant $c \ll 1$ (depending on the implied constants so far), then one has

$$\widehat{\mathcal{B}}_\varphi(\kappa) \ll (\kappa^2+1)^{-1}, \qquad \text{for } \kappa \in \mathbb{R}\setminus\{0\}, \tag{3.22}$$

$$\widehat{\mathcal{B}}_\varphi(\kappa) \asymp Y^{-2i\kappa}, \qquad \text{for } 0 < i\kappa < \frac{1}{2}. \tag{3.23}$$

Proof. The bounds in (3.19) to (3.21) constitute [10, Lemma 7.1] (note that φ satisfies the requirements in [10, (1.43) and (1.44)] for $(Y, 1)$ in place of (X, Y)). Similarly, (3.22) and the lower bound in (3.23) are [10, (8.2) and (8.3), following from (8.1)], using an appropriate choice of the constants η_1, η_2. The upper bound in (3.23) also follows from [10, (8.1)], but is in fact already covered by (3.19) (using $r = -i\kappa$ and the fact that $\widehat{\mathcal{B}}_\varphi$ is even). □

Finally, let us state the Kuznetsov trace formula, following the notation of Deshouillers–Iwaniec [10].

Proposition 3.5 (Kuznetsov trace formula [10, 31]). *Let $\varphi : \mathbb{R} \to \mathbb{C}$ be a compactly-supported smooth function, $q \in \mathbb{Z}_+$, and $\mathfrak{a}, \mathfrak{b}$ be cusps of $\Gamma_0(q)$. Then for any positive integers m, n and $\mathrm{sgn} \in \{1, -1\}$, one has*

$$\sum_{c \in \mathcal{C}_{\mathfrak{a}\mathfrak{b}}} \frac{S_{\mathfrak{a}\mathfrak{b}}(m, \mathrm{sgn}\cdot n; c)}{c}\varphi\left(\frac{4\pi\sqrt{mn}}{c}\right) = \begin{cases}\mathcal{H}+\mathcal{M}+\mathcal{E}, & \mathrm{sgn} = 1,\\ \mathcal{M}'+\mathcal{E}', & \mathrm{sgn} = -1,\end{cases} \tag{3.24}$$

with the following notations. Firstly, the holomorphic contribution is

$$\mathcal{H} = \frac{1}{2\pi}\sum_{k \in 2\mathbb{Z}_+} \widetilde{\mathcal{B}}_\varphi(k-1)\frac{i^k(k-1)!}{(4\pi\sqrt{mn})^{k-1}}\sum_{j=1}^{h_k(q)} \overline{\psi_{jk\mathfrak{a}}(m)}\,\psi_{jk\mathfrak{b}}(n), \tag{3.25}$$

for any orthonormal bases of level-q holomorphic cusp forms $(f_{jk})_j$ of weight $k \in 2\mathbb{Z}_+$, with Fourier coefficients $\psi_{jk\mathfrak{a}}(n)$ as in (3.14). Secondly, the Maass contributions are

$$\mathcal{M} = \sum_{j=1}^{\infty} \frac{\widehat{\mathcal{B}}_\varphi(\kappa_j)}{\cosh(\pi\kappa_j)} \overline{\rho_{j\mathfrak{a}}(m)}\, \rho_{j\mathfrak{b}}(n), \qquad \mathcal{M}' = \sum_{j=1}^{\infty} \frac{\check{\mathcal{B}}_\varphi(\kappa_j)}{\cosh(\pi\kappa_j)} \rho_{j\mathfrak{a}}(m)\, \rho_{j\mathfrak{b}}(n), \tag{3.26}$$

for any orthonormal basis $(u_j)_j$ of level-q Maass cusp forms, with eigenvalues λ_j (and κ_j, θ_j as in (3.17)), and Fourier coefficients $\rho_{j\mathfrak{a}}(n)$ as in (3.15). Thirdly, the Eisenstein contributions are

$$\begin{aligned} \mathcal{E} &= \frac{1}{\pi} \sum_{\mathfrak{c}} \int_{-\infty}^{\infty} \widehat{\mathcal{B}}_\varphi(r) \left(\frac{m}{n}\right)^{-ir} \overline{\varphi_{\mathfrak{c}\mathfrak{a}}\left(m; \frac{1}{2} + ir\right)} \varphi_{\mathfrak{c}\mathfrak{b}}\left(n; \frac{1}{2} + ir\right) dr, \\ \mathcal{E}' &= \frac{1}{\pi} \sum_{\mathfrak{c}} \int_{-\infty}^{\infty} \check{\mathcal{B}}_\varphi(r)\, (mn)^{ir}\, \varphi_{\mathfrak{c}\mathfrak{a}}\left(m; \frac{1}{2} + ir\right) \varphi_{\mathfrak{c}\mathfrak{b}}\left(n; \frac{1}{2} + ir\right) dr, \end{aligned} \tag{3.27}$$

where the Fourier coefficients $\varphi_{\mathfrak{a}\mathfrak{b}}(n; s)$ are as in (3.16), and $\mathfrak{c}$ varies over the cusps of $\Gamma_0(q)$.

Proof. This is [10, Theorem 2]. □

Remark 3.6. Upon inspecting the Maass contribution (3.26) in light of the bounds (3.19) and (3.23), the losses due to the exceptional spectrum are apparent. Indeed, if $\varphi(y)$ is supported in $y \asymp Y \asymp \frac{\sqrt{mn}}{C}$ for some $C > 0$ (indicating the size of c), then the Bessel transforms bounds for exceptional eigenvalues are (a priori) worse by a factor of

$$\max\left(1, Y^{-2\theta(q)}\right) \asymp \left(1 + \frac{C}{\sqrt{mn}}\right)^{2\theta(q)},$$

compared to the regular (nonexceptional) spectrum.

3.4. Bounds for Fourier coefficients

If one is interested in a particular holomorphic or Maass cusp form (ideally, a Hecke eigenform), then various bounds for its Fourier coefficients follow from the theory of automorphic representations and their L-functions [10, 40, 28, 6, 19, 20]. Here we are interested in bounding averages over bases of automorphic forms, resembling those that show up in (3.25) to (3.27); naturally, these would be useful in combination with the Kuznetsov formula.

Remarkably, such bounds are often derived using the Kuznetsov formula once again (with different parameters, including the support range of the smooth function φ), together with various bounds for sums of Kloosterman sums, such as the Weil bound below.

Lemma 3.7 (Weil–Ramanujan bound). *For any $c \in \mathbb{Z}_+$ and $m, n \in \mathbb{Z}$, one has*

$$S(m, n; c) \ll \tau(c)\, (m, n, c)^{1/2} c^{1/2}.$$

Also, for $m = 0$, one has $|S(0, n; c)| \le (n, c)$.

Proof. See [26, Corollary 11.12] for the first bound; the second bound, concerning Ramanujan sums, is classical and follows by Möbius inversion. □

The first results that we mention keep the index n of the Fourier coefficients fixed, while varying the automorphic form.

Lemma 3.8 (Fourier coefficient bounds with fixed n). *Let $K \gg 1$ and $\varepsilon > 0$. With the notation of Proposition 3.5, each of the three expressions*

$$\sum_{\substack{k \in 2\mathbb{Z}_+ \\ k \le K}} \frac{(k-1)!}{(4\pi n)^{k-1}} \sum_{j=1}^{h_k(q)} |\psi_{jk\mathfrak{a}}(n)|^2, \qquad \sum_{|\kappa_j| \le K} \frac{|\rho_{j\mathfrak{a}}(n)|^2}{\cosh(\pi\kappa_j)}, \qquad \sum_{\mathfrak{c}} \int_{-K}^{K} \left|\varphi_{\mathfrak{c}\mathfrak{a}}\left(n; \frac{1}{2} + ir\right)\right|^2 dr$$

is bounded up to a constant depending on ε by

$$K^2 + (qnK)^{\varepsilon}\,(q,n)^{1/2}\,\mu(\mathfrak{a})\,n^{1/2}.$$

Moreover, for the exceptional spectrum we have

$$\sum_{\lambda_j<1/4} X^{2\theta_j}\,|\rho_{j\mathfrak{a}}(n)|^2 \ll_{\varepsilon} (qN)^{\varepsilon}\left(1+(q,n)^{1/2}\,\mu(\mathfrak{a})\,n^{1/2}\right), \tag{3.28}$$

for any $X \ll \max\left(1, \left((q,n)\,\mu(\mathfrak{a})^2\,n\right)^{-1}\right)$.

Proof. These bounds roughly follow by combining Lemma 3.7 with trace formulas like Proposition 3.5, for $m = n$ and suitable choices of φ. See for example [44, Lemmas 2.7 and 2.9] with $q_0 = 1$ and $X = X_0$, noting the different normalizations of the Fourier coefficients. □

One of the key insights of Deshouillers–Iwaniec [10] was that the bounds in Lemma 3.8 can be improved when averaging over $n \sim N$, by exploiting the bilinear structure in m, n of the spectral side of the Kuznetsov formula (3.24). This leads to the so-called *weighted large sieve inequalities* for the Fourier coefficients of automorphic forms, involving arbitrary sequences (a_n); for 1-bounded sequences, the result below saves a factor of roughly N over the pointwise bounds in Lemma 3.8.

Lemma 3.9 (Deshouillers–Iwaniec large sieve for the regular spectrum [10]). *Let $K, N \geq 1/2$, $\varepsilon > 0$, and (a_n) be a sequence of complex numbers. With the notation of Proposition 3.5, each of the three expressions*

$$\sum_{\substack{k\in 2\mathbb{Z}_+\\ k\leq K}} \frac{(k-1)!}{(4\pi)^{k-1}} \sum_{j=1}^{h_k(q)} \left|\sum_{n\sim N} a_n n^{-(k-1)/2}\,\psi_{jk\mathfrak{a}}(n)\right|^2, \qquad \sum_{|\kappa_j|\leq K} \frac{1}{\cosh(\pi\kappa_j)} \left|\sum_{n\sim N} a_n\,\rho_{j\mathfrak{a}}(n)\right|^2,$$

$$\sum_{\mathfrak{c}} \int_{-K}^{K} \left|\sum_{n\sim N} a_n\,n^{ir}\,\varphi_{\mathfrak{c}\mathfrak{a}}\left(n;\frac{1}{2}+ir\right)\right|^2 dr$$

is bounded up to a constant depending on ε by

$$\left(K^2 + \mu(\mathfrak{a})N^{1+\varepsilon}\right)\|a_n\|_2^2.$$

Proof. This is [10, Theorem 2]. □

Lemma (3.9) includes the contribution of the exceptional Maass cusp forms, but is not the optimal result for handling it. Indeed, to temper the growth of the Bessel functions weighing the exceptional Fourier coefficients in (3.26), one needs to incorporate factors of $X^{2\theta_j}$ into the sum over Maass forms, as in (3.28). The following is a preliminary result toward such bounds.

Corollary 3.10 (Preliminary bound for exceptional forms [10]). *Let $X, N \geq 1/2$, $\varepsilon > 0$, $(a_n)_{n\sim N}$ be a complex sequence. Let $\Phi(t)$ be a nonnegative smooth function supported in $t \asymp 1$, with $\Phi^{(j)}(t) \ll_j 1$ for $j \geq 0$, and $\int \Phi(t)\,dt \gg 1$. Then with the notation of Proposition 3.5, one has*

$$\begin{aligned}\sum_{\lambda_j<1/4} X^{2\theta_j}\left|\sum_{n\sim N} a_n\,\rho_{j\mathfrak{a}}(n)\right|^2 \ll \left|\sum_{c\in\mathcal{C}_{\mathfrak{a}\mathfrak{a}}} \frac{1}{c} \sum_{m,n\sim N} \overline{a_m}\,a_n\,S_{\mathfrak{a}\mathfrak{a}}(m,n;c)\,\Phi\left(\frac{\sqrt{mn}}{c}X\right)\right| \\ + O_{\varepsilon}\left(1+\mu(\mathfrak{a})N^{1+\varepsilon}\right)\|a_n\|_2^2.\end{aligned} \tag{3.29}$$

Proof. This is essentially present in [10] (see [10, first display on p. 271], and [10, (8.7)] for the case $\mathfrak{a} = \infty$), but let us give a short proof for completion. If $X \ll 1$, the result follows immediately from

Lemma (3.9) with $K = 1/4$, and the bound $\cosh(\pi\kappa) \asymp 1$ for $i\kappa \in [0, 1/4]$ (recall that $i\kappa_j = \theta_j \le \frac{7}{64}$ by Theorem 3.3, but the weaker Selberg bound $\theta_j \le \frac{1}{4}$ suffices here).

Otherwise, let $\varphi(y) := \Phi(yX(4\pi)^{-1})$, which satisfies all the assumptions in Lemma 3.4 for $Y = 4\pi X^{-1}$; in particular, we have

$$\max(\widehat{\mathcal{B}}_\varphi(r), \widetilde{\mathcal{B}}_\varphi(r)) \ll |r|^{-5/2}, \qquad \text{for } |r| \ge 1, \tag{3.30}$$

$$\widehat{\mathcal{B}}_\varphi(\kappa) \ll \left(\kappa^2 + 1\right)^{-1}, \qquad \text{for } \kappa \in \mathbb{R} \setminus \{0\}, \tag{3.31}$$

$$\widehat{\mathcal{B}}_\varphi(\kappa) \gg X^{2i\kappa}, \qquad \text{for } 0 < i\kappa < 1/2. \tag{3.32}$$

Now apply Proposition 3.5 with this choice of φ and $\mathfrak{a} = \mathfrak{b}$, multiply both sides by $\overline{a_m}\, a_n$, and sum over $m, n \sim N$, to obtain

$$\begin{aligned}\sum_{c\in C_{\mathfrak{a}\mathfrak{a}}} \frac{1}{c} \sum_{m,n\sim N} \overline{a_m}\, a_n\, S_{\mathfrak{a}\mathfrak{a}}(m,n;c)\, \varphi\left(\frac{4\pi\sqrt{mn}}{c}\right) &= \sum_{j\ge 1} \frac{\widehat{\mathcal{B}}_\varphi(\kappa_j)}{\cosh(\pi\kappa_j)} \left|\sum_{n\sim N} a_n\, \rho_{j\mathfrak{a}}(n)\right|^2 \\ &+ \frac{1}{\pi} \sum_{\mathfrak{c}} \int_{-\infty}^{\infty} \widehat{\mathcal{B}}_\varphi(r) \left|\sum_{n\sim N} a_n\, n^{ir}\, \varphi_{\mathfrak{c}\mathfrak{a}}\left(n; \frac{1}{2} + ir\right)\right|^2 dr \\ &+ \frac{1}{2\pi} \sum_{k\in 2\mathbb{Z}_+} \widetilde{\mathcal{B}}_\varphi(k-1) \frac{(k-1)!}{(4\pi)^{k-1}} \sum_{1\le j\le h_k(q)} \left|\sum_{n\sim N} a_n\, n^{-\frac{k-1}{2}} \psi_{jk\mathfrak{a}}(n)\right|^2.\end{aligned}$$

Bounding the contribution of nonexceptional Maass cusp forms, holomorphic cusp forms, and Eisenstein series via (3.30), (3.31), and Lemma (3.9) (in dyadic ranges $K = 2^p$), this reduces to

$$\begin{aligned}\sum_{c\in C_{\mathfrak{a}\mathfrak{a}}} \frac{1}{c} \sum_{m,n\sim N} \overline{a_m}\, a_n\, S_{\mathfrak{a}\mathfrak{a}}(m,n;c)\, \Phi\left(\frac{\sqrt{mn}}{c} X\right) &= \sum_{\lambda_j < 1/4} \frac{\widehat{\mathcal{B}}_\varphi(\kappa_j)}{\cosh(\pi\kappa_j)} \left|\sum_{n\sim N} a_n\, \rho_{j\mathfrak{a}}(n)\right|^2 \\ &+ O_\varepsilon\left(1 + \mu(\mathfrak{a}) N^{1+\varepsilon}\right) \|a_n\|_2^2.\end{aligned} \tag{3.33}$$

Combining this with the lower bound $\widehat{\mathcal{B}}_\varphi(\kappa_j) \gg X^{2\theta_j}$ (due to (3.32)), we recover the desired bound in (3.29). □

Finally, for the results with averaging over the level q, we will also need the following result of Deshouillers–Iwaniec [10].

Theorem 3.11 (Deshouillers–Iwaniec's large sieve with level averaging [10]). *Let $\varepsilon > 0$, $X > 0$, $N, Q \ge 1/2$, and $\omega \in \mathbb{R}/\mathbb{Z}$. Let $q \in \mathbb{Z}_+$ and ∞_q denote the cusp at ∞ of $\Gamma_0(q)$, with the choice of scaling matrix $\sigma_{\infty_q} = \mathrm{Id}$. Then with the notation of Proposition 3.5, one has*

$$\sum_{q\sim Q} \sum_{\lambda_j(q)<1/4} X^{2\theta_j(q)} \left|\sum_{n\sim N} e(n\omega)\, \rho_{j\infty_q}(n)\right|^2 \ll_\varepsilon (QN)^\varepsilon\, (Q+N)\, N, \tag{3.34}$$

for any

$$X \ll \max\left(N, \frac{Q^2}{N}\right). \tag{3.35}$$

Proof. This follows immediately from [10, Theorem 1.7] with $X \leftarrow X^{1/2}$. As noted in previous works [39, 3, 8], although [10, Theorem 7] was only stated for $\alpha = 0$, the same proof holds uniformly in $\alpha \in \mathbb{R}/\mathbb{Z}$. □

4. Combinatorial bounds

In this section, we obtain bounds for bilinear sums of the form $\sum_m a_m \sum_n b_n S(m,n;c)$ (say, in the range $c^{1/4} \ll N \ll c$), saving over the Pólya–Vinogradov and Weil bounds if the Fourier transforms $\widehat{a}$ and $\widehat{b}$ are concentrated enough. Our computations here are elementary (not requiring the spectral theory of automorphic forms yet, nor any other prerequisites beyond Section 3.1), and use a combinatorial argument inspired by [7]; the latter was also used, for example, in [27].

We highlight the following nonstandard notation.

Notation 4.1 (Rational approximation). Given $M, N > 0$, let $T_{M,N} : (\mathbb{R}/\mathbb{Z})^2 \to \mathbb{R}$ denote the function

$$\begin{gathered} T_{M,N}(\alpha,\beta) := \min_{t \in \mathbb{Z}_+} (t + M\|\alpha t\| + N\|\beta t\|) \\ (\text{abbreviating } T_N := T_{N,N},\ T_N(\alpha) := T_N(\alpha,\alpha)). \end{gathered} \tag{4.1}$$

This measures how well α and β can be simultaneously approximated by rational numbers with small denominators t, in terms of the balancing parameters M, N. The inverses of these parameters indicate the scales at which $T_{M,N}(\alpha,\beta)$ has roughly constant size, due to the following lemma.

Lemma 4.2 (Basic properties of $T_{M,N}$). *Let $M, N > 0$ and $\alpha, \beta, \gamma, \delta \in \mathbb{R}/\mathbb{Z}$. One has $T_{N,M}(\beta,\alpha) = T_{M,N}(\alpha,\beta) = T_{M,N}(\pm\alpha, \pm\beta)$ and*

$$T_N(\alpha, \beta \pm \alpha) \asymp T_N(\alpha,\beta). \tag{4.2}$$

Moreover,

$$T_{M,N}(\alpha+\gamma,\beta) \le (1 + M\|\gamma\|)\, T_{M,N}(\alpha,\beta). \tag{4.3}$$

In particular, if $\|\gamma\| \ll M^{-1}$ and $\|\delta\| \ll N^{-1}$, then

$$T_{M,N}(\alpha+\gamma,\beta+\delta) \asymp T_{M,N}(\alpha,\beta). \tag{4.4}$$

Proof. The first equalities are obvious, and (4.2) follows from the triangle inequalities

$$\|(\beta \pm \alpha)t\| \le \|\alpha t\| + \|\beta t\|, \qquad \|\beta t\| \le \|\alpha t\| + \|(\beta \pm \alpha)t\|.$$

For (4.3), we note that

$$\begin{aligned} t + M\|(\alpha+\gamma)t\| + N\|\beta t\| &\le t + M\|\gamma t\| + M\|\alpha t\| + N\|\beta t\| \\ &\le t(1 + M\|\gamma\|) + M\|\alpha t\| + N\|\beta t\| \\ &\le (1 + M\|\gamma\|)\,(t + M\|\alpha t\| + N\|\beta t\|), \end{aligned}$$

and take a minimum of both sides over $t \in \mathbb{Z}_+$. Finally, (4.4) follows immediately from (4.3). □

Lemma 4.3 (Dirichlet-style approximation). *Let $\alpha, \beta \in \mathbb{R}/\mathbb{Z}$. Given any parameters $A, B \gg 1$, there exists a positive integer t such that*

$$t \ll AB, \qquad \|\alpha t\| \ll \frac{1}{A}, \qquad \|\beta t\| \ll \frac{1}{B}.$$

In particular, for $N \geq 1/2$, one has

$$T_N(\alpha,\beta) \ll \min\left(\sqrt{N(1+\|\alpha-\beta\|N)}, N^{2/3}\right). \tag{4.5}$$

Proof. Consider the sequence of points $\{(t\alpha, t\beta)\}_{t \leq \lceil A\rceil\lceil B\rceil+2}$ in $(\mathbb{R}/\mathbb{Z})^2$; by the pigeonhole principle, at least two of these must lie in a box of dimensions $A^{-1} \times B^{-1}$, say $(t_i\alpha, t_i\beta)$ for $i \in \{1,2\}$. Then we can pick $t := |t_1 - t_2|$ to establish the first claim.

Using $A = B = N^{1/3}$, we find that

$$T_N(\alpha,\beta) \ll N^{2/3},$$

uniformly in $\alpha, \beta \in \mathbb{R}/\mathbb{Z}$. Using $A = \sqrt{N/(1+\|\beta\|N)}$ and $B = 1$, we also have

$$\begin{aligned} T_N(\alpha,\beta) \leq \min_{t\in\mathbb{Z}_+}(t + N(\|\alpha t\| + \|\beta\| t)) &\ll A + \frac{N}{A} + N\|\beta\|A \\ &\ll \sqrt{N(1+\|\beta\|N)}, \end{aligned}$$

and thus

$$T_N(\alpha,\beta) \ll T_N(\alpha,\alpha-\beta) \ll \sqrt{N(1+\|\alpha-\beta\|N)}.$$

This proves (4.5). □

Lemma 4.4 (Concentration of points on modular hyperbolas, following Cilleruelo–Garaev [7]). *Let $c \in \mathbb{Z}_+$, $a, b, \lambda \in \mathbb{Z}/c\mathbb{Z}$, $0 < X, Y \ll c$, and $I, J \subset \mathbb{R}$ be intervals of lengths $|I| = X$, $|J| = Y$. Then for any $\varepsilon > 0$ and any $(c\alpha, c\beta) \in I \times J$, one has*

$$\#\{(x,y) \in (I\cap\mathbb{Z}) \times (J\cap\mathbb{Z}) : xy \equiv \lambda \ (mod\ c)\} \ll_\varepsilon c^\varepsilon \left(\frac{XY}{c} T_{\frac{c}{X},\frac{c}{Y}}(\alpha,\beta) + \gcd(\lambda, c)\right), \tag{4.6}$$

with $T_{M,N}(\alpha,\beta)$ as in Notation 4.1.

Remark 4.5. Lemma 4.4 counts solutions to the congruence $xy \equiv \lambda \pmod c$ in short intervals. On average over intervals of length $X, Y \gg \sqrt{c}$, one should expect around XY/c solutions; (4.6) essentially recovers this average bound when α and β can be simultaneously approximated by rational numbers with a bounded denominator.

Remark 4.6. One can also interpret Lemma 4.4 in terms of sum-product phenomena over $\mathbb{Z}/c\mathbb{Z}$. Indeed, the intervals $a + [-X, X]$ and $b + [-Y, Y]$ have many "additive collisions" of the form $x_1 + y_1 \equiv x_2 + y_2 \pmod c$ (with $x_1, x_2 \in a + [-X, X]$ and $y_1, y_2 \in b + [-Y, Y]$), so they should have few "multiplicative collisions" of the form $x_1 y_1 \equiv \lambda \equiv x_2 y_2 \pmod c$.

Proof. If $I \cap \mathbb{Z} = \emptyset$ or $J \cap \mathbb{Z} = \emptyset$, the claim is trivial. So let $a \in I \cap \mathbb{Z}$ and $b \in J \cap \mathbb{Z}$; by a change of variables, we have

$$\#\{(x,y) \in I \times J : xy \equiv \lambda \pmod c\} \leq \#S(a,b),$$

where

$$S(a,b) := \{(x,y) \in ([-X,X]\cap\mathbb{Z}) \times ([-Y,Y]\cap\mathbb{Z}) : (x+a)(y+b) \equiv \lambda \pmod c\}.$$

The key idea, borrowed from [7, Theorem 1] (and also used, for example, in [27, Lemma 5.3]), is to effectively reduce the size of a and b by appropriately scaling the congruence $(x+a)(y+b) \equiv \lambda \pmod c$,

and then to pass to an equation in the integers. Indeed, let $t \in \mathbb{Z}_+$ be a scalar, and let a', b' be the integers with minimal absolute values such that

$$at \equiv a' \pmod{c} \qquad \text{and} \qquad bt \equiv b' \pmod{c}. \tag{4.7}$$

Then any given pair $(x, y) \in S(a, b)$ also satisfies the scaled congruence

$$t(x+a)(y+b) \equiv t\lambda \pmod{c} \qquad \Longleftrightarrow \qquad txy + b'x + a'y \equiv t(\lambda - ab) \pmod{c}.$$

Denoting by $r \in \{0, 1, \dots, c-1\}$ the residue of $t(\lambda - ab) \pmod{c}$, and

$$z = z(x, y) := \frac{txy + b'x + a'y - r}{c},$$

it follows that (x, y, z) is an integer solution to the equation

$$txy + b'x + a'y = cz + r \qquad \Leftrightarrow \qquad (tx + a')(ty + b') = t(cz + r) + a'b'.$$

Note that

$$\begin{aligned} z &\ll \frac{tXY + |b'|X + |a'|Y + c}{c} \\ &\ll \frac{t}{c}XY + \left\|\frac{bt}{c}\right\| X + \left\|\frac{at}{c}\right\| Y + 1 =: Z(t). \end{aligned}$$

Now let $n(z) := t(cz + r) + a'b'$. The number of pairs $(x, y) \in S(a, b)$ with $n(z) \neq 0$ is at most

$$\sum_{\substack{z \ll Z(t) \\ n(z) \neq 0}} \sum_{\substack{x, y \in \mathbb{Z} \\ (tx+a')(ty+b') = n(z) \\ (x+a)(y+b) \equiv \lambda \pmod{c}}} 1 \ll_\varepsilon (ct)^\varepsilon Z(t),$$

by the divisor bound. On the other hand, if $(x, y) \in S(a, b)$ satisfies $n(z) = (tx + a')(ty + b') = 0$, this forces $tx = -a'$ or $ty = -b'$, determining one of x and y uniquely. Suppose x is determined; the condition $c \mid (x + a)(y + b) - \lambda$ implies $d := \gcd(x + a, c) \mid \gcd(\lambda, c)$, so

$$\frac{c}{d} \,\Big|\, \frac{x+a}{d}(y+b) - \frac{\lambda}{d}.$$

Since $\gcd(c/d, (x + a)/d) = 1$, this uniquely determines the value of $y \pmod{c/d}$, leading to a total contribution of $1 + Yd/c$. Putting things together, we conclude that

$$\begin{aligned} \#S(a, b) &\ll_\varepsilon c^\varepsilon \min_{t \in \mathbb{Z}_+} (t^\varepsilon Z(t)) + 1 + \frac{X+Y}{c} \gcd(\lambda, c) \\ &\ll c^\varepsilon \left(\frac{XY}{c} \min_{t \in \mathbb{Z}_+} t^\varepsilon \left(t + \frac{c}{X} \left\|\frac{at}{c}\right\| + \frac{c}{Y} \left\|\frac{bt}{c}\right\| \right) + 1 + \frac{X+Y}{c} \gcd(\lambda, c) \right) \\ &\ll c^{2\varepsilon} \left(\frac{XY}{c} T_{\frac{c}{X}, \frac{c}{Y}} \left(\frac{a}{c}, \frac{b}{c} \right) + \gcd(\lambda, c) \right), \end{aligned}$$

where we used that $X, Y \ll c$ in the last line (and implicitly that the minimum of $t + \frac{c}{X}\|at/c\| + \frac{c}{Y}\|bt/c\|$ is attained for $t \ll c$). Now if $\alpha, \beta \in \mathbb{R}$ satisfy $(c\alpha, c\beta) \in I \times J$, then we have $|a - c\alpha| \leq X$ and $|b - c\beta| \leq Y$, that is,

$$\left| \frac{a}{c} - \alpha \right| \leq \frac{X}{c}, \qquad \left| \frac{b}{c} - \beta \right| \leq \frac{Y}{c}.$$

So by (4.4), we have

$$T_{\frac{c}{X},\frac{c}{Y}}\left(\frac{a}{c},\frac{b}{c}\right) \asymp T_{\frac{c}{X},\frac{c}{Y}}(\alpha,\beta).$$

We thus obtain the desired bound, up to a rescaling of ε. □

We now work towards our bilinear Kloosterman bound for sequences with sparse Fourier transforms, reminding the reader of the Fourier-analytic notation in Section 3.1. The connection to counting solutions to congruences of the form $xy \equiv 1 \pmod c$ comes from the identity

$$\sum_m a_m \sum_n b_n\, S(m,n;c) = \sum_{x,y \pmod c} \widehat{a}\left(\frac{x}{c}\right)\widehat{b}\left(\frac{y}{c}\right)\mathbb{1}_{xy\equiv 1 \pmod c}, \tag{4.8}$$

obtained by expanding $S(m,n;c)$ and swapping sums. One can interpret this as a Parseval–Plancherel identity, the Kloosterman sum $S(m,n;c)$ being dual to the function $\mathbb{1}_{xy\equiv 1 \pmod c}$; this duality is often exploited in the converse direction (see, e.g., [35, Chapter 6] and [17]), but it turns out to also be a useful input for methods from the spectral theory of automorphic forms.

Proposition 4.7 (Bilinear Kloosterman bound with exponential phases). *Let $c, a \in \mathbb{Z}_+$, $\alpha,\beta \in \mathbb{R}/\mathbb{Z}$, $1 \ll M, N \ll c$, and $I, J \subset \mathbb{Z}$ be nonempty discrete intervals of lengths $|I| = M$, $|J| = N$. Then for any $\varepsilon > 0$, one has*

$$\sum_{m\in I} e(m\alpha) \sum_{n\in J} e(n\beta)\, S(am,an;c) \ll_\varepsilon c^\varepsilon \left(c\, T_{M,N}(\alpha,\beta) + \gcd(a,c)MN\right).$$

Remark 4.8. When $\alpha = \beta = 0$, this recovers a result of Shparlinski and Zhang [43]. A similar argument produces the more general bound

$$\sum_{m\in I} e(m\alpha) \sum_{n\in J} e(n\beta)\, S(am+r,bn+s;c) \ll_\varepsilon c^\varepsilon \left(c\, T_{M,N}(\alpha,\beta) + \gcd\left(\frac{ab}{\gcd(a,b,c)},c\right)MN\right),$$

for entries of Kloosterman sums in arithmetic progressions, where $a, b \in \mathbb{Z}\setminus\{0\}$, $r, s \in \mathbb{Z}$.

Proof. We first note that if $\gcd(a,c) > 1$ and $a' := a/\gcd(a,c)$, $c' := c/\gcd(a,c)$, the Chinese remainder theorem yields $S(am,an;c) = \frac{\varphi(c)}{\varphi(c')} S(a'm,a'n;c')$. Since $\frac{\varphi(c)}{\varphi(c')} \le \frac{c}{c'} = \gcd(a,c)$, one can deduce the desired bound from the same bound for $(a,c) \mapsto (a',c')$. This allows us to assume without loss of generality that $\gcd(a,c) = 1$.

Let $\mathcal{S}$ denote the sum in Proposition 4.7; as in (4.8), we expand $S(am,an;c)$ and swap sums to obtain

$$\mathcal{S} = \sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times} \sum_{m\in I} e\left(m\alpha + m\frac{ax}{c}\right) \sum_{n\in J} e\left(n\beta + n\frac{a\bar{x}}{c}\right).$$

We note that

$$\sum_{m\in I} e\left(m\alpha + m\frac{ax}{c}\right) \ll \min\left(M, \left\|\alpha + \frac{ax}{c}\right\|^{-1}\right),$$

and put $M\|\alpha + \frac{ax}{c}\|$ into dyadic ranges

$$A_0 := [0,2], \qquad A_j := \left(2^j, 2^{j+1}\right].$$

Proceeding similarly for the sum over n, we obtain

$$\begin{aligned}
\mathcal{S} &= \sum_{\substack{0\le j\le \log_2 M\\ 0\le k\le \log_2 N}} \sum_{\substack{x\in(\mathbb{Z}/c\mathbb{Z})^\times\\ M\|\alpha+\frac{a\overline{x}}{c}\|\in A_j\\ N\|\beta+\frac{a\overline{x}}{c}\|\in A_k}} \sum_{m\in I} e\left(m\alpha+m\frac{ax}{c}\right)\sum_{n\in J} e\left(n\beta+n\frac{a\overline{x}}{c}\right)\\
&\ll \sum_{\substack{0\le j\le \log_2 M\\ 0\le k\le \log_2 N}} \sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times} \mathbb{1}_{M\|\alpha+\frac{ax}{c}\|\in A_j}\mathbb{1}_{N\|\beta+\frac{a\overline{x}}{c}\|\in A_k}\frac{MN}{2^{j+k}}\\
&\le \sum_{\substack{0\le j\le \log_2 M\\ 0\le k\le \log_2 N}} \frac{MN}{2^{j+k}} \sum_{\substack{x,y\in\mathbb{Z}/c\mathbb{Z}\\ xy\equiv a^2 \pmod c}} \mathbb{1}_{M\|\alpha+\frac{x}{c}\|\in A_j}\,\mathbb{1}_{N\|\beta+\frac{y}{c}\|\in A_k}.\\
&\le \sum_{\substack{0\le j\le \log_2 M\\ 0\le k\le \log_2 N}} \frac{MN}{2^{j+k}} \sum_{\substack{x,y\in\mathbb{Z}\\ xy\equiv a^2 \pmod c}} \mathbb{1}_{|x+c\alpha|\le c\frac{2^{j+1}}{M}}\,\mathbb{1}_{|y+c\beta|\le c\frac{2^{k+1}}{N}},
\end{aligned}$$

where we noted that for any $x_0, y_0 \in \mathbb{Z}/c\mathbb{Z}$, there exist $x, y \in \mathbb{Z}$ with $x \equiv x_0 \pmod c$, $y \equiv y_0 \pmod c$, and $\|\alpha + \frac{x_0}{c}\| = |\alpha + \frac{x}{c}|$, $\|\beta + \frac{y_0}{c}\| = |\beta + \frac{y}{c}|$.

We can bound the inner sum using Lemma 4.4 with $X = c2^{j+2}M^{-1}$, $Y = c2^{k+2}N^{-1}$, and $\lambda = a^2$. Since the function $T_{M,N}$ is nondecreasing in M, N, this gives

$$\begin{aligned}
\mathcal{S} &\ll_\varepsilon c^\varepsilon \sum_{\substack{0\le j\le \log_2 M\\ 0\le k\le \log_2 N}} \frac{MN}{2^{j+k}}\left(\frac{(c2^jM^{-1})(c2^kN^{-1})}{c}T_{\frac{M}{2^{j+1}},\frac{N}{2^{k+1}}}(-\alpha,-\beta)+\gcd(a^2,c)\right)\\
&\ll_\varepsilon c^{2\varepsilon}\left(cT_{M,N}(\alpha,\beta)+MN\right).
\end{aligned}$$

This yields the desired bound up to a rescaling of ε. □

Proposition 4.9 (Bilinear Kloosterman bound with frequency concentration). *Let $c, a \in \mathbb{Z}_+$, $1 \ll M, N \ll c$, and $I, J \subset \mathbb{Z}$ be nonempty discrete intervals of lengths $|I| = M$, $|J| = N$. Let $(a_m)_{m\in I}, (b_n)_{n\in J}$ be complex sequences, and μ, ν be bounded-variation complex Borel measures on $\mathbb{R}/\mathbb{Z}$, such that $\check{\mu}(m) = a_m$ for $m \in I$ and $\check{\nu}(n) = b_n$ for $n \in J$. Then for any $\varepsilon > 0$, one has*

$$\sum_{m\in I} a_m \sum_{n\in J} b_n\, S(am, an; c) \ll_\varepsilon c^\varepsilon \iint_{(\mathbb{R}/\mathbb{Z})^2} \left(c\,T_{M,N}(\alpha,\beta)+\gcd(a,c)MN\right) d|\mu|(\alpha)\, d|\nu|(\beta), \quad (4.9)$$

By (4.5), when $M = N$, this bound is $\ll c^\varepsilon(cN^{2/3} + \gcd(a, c)N^2)\,|\mu|(\mathbb{R}/\mathbb{Z})\,|\nu|(\mathbb{R}/\mathbb{Z})$.

Proof. By Fourier inversion, expand

$$a_m = \int_{\mathbb{R}/\mathbb{Z}} e(m\alpha)\, d\mu(\alpha), \qquad b_n = \int_{\mathbb{R}/\mathbb{Z}} e(n\beta)\, d\mu(\beta),$$

then swap sums and integrals, and apply Proposition 4.7. □

Remark 4.10. Suppose $M = N$ and $a = 1$. By comparison, the pointwise Weil bound would yield a right-hand side in (4.9) of roughly $N\sqrt{c}\,\|a_m\|_2\|b_n\|_2$, while applying Cauchy–Schwarz after (4.8) gives the bound $c\|a_m\|_2\|b_n\|_2$ (these essentially lead to the ranges in Theorem 1.2). It is a very difficult problem [29, 27] to improve these bounds for general sequences $(a_m), (b_n)$, but it becomes easier given suitable information in the frequency space. Indeed, with the natural choice of measures $d\mu = \widehat{a}\, d\lambda$, $d\nu = \widehat{b}\, d\lambda$ (where λ is the Lebesgue measure), Proposition 4.9 saves over the relevant bound

$c\|a_m\|_2\|b_n\|_2 = c\|\widehat{a}\|_{L^2}\|\widehat{b}\|_{L^2}$ whenever $\widehat{a}, \widehat{b}$ satisfy the concentration inequality

$$\frac{\|\widehat{a}\|_{L^1}}{\|\widehat{a}\|_{L^2}} \cdot \frac{\|\widehat{b}\|_{L^1}}{\|\widehat{b}\|_{L^2}} = o\left(\frac{1}{N^{2/3} + N^2 c^{-1}}\right).$$

For reference, the left-hand side is always $\gg N^{-1}$. One may do better by treating the integral in (4.9) more carefully, or by including the contribution of other frequencies into μ and ν (this liberty is due to the handling of sharp cutoffs in Proposition 4.7). For instance, one could extend the sequences $(a_m), (b_n)$ with a smooth decay beyond I and J before taking their Fourier transforms, or one could construct μ, ν out of Dirac delta measures (in particular, one recovers Proposition 4.7 this way).

We will ultimately use Proposition 4.9 for sequences (a_n) of the shape in (1.3), so it is necessary to understand their Fourier transforms. The case of exponential phases $a_n = e(n\alpha)$ is trivial, but the dispersion coefficients from Theorem 1.7 are more interesting, warranting a separate lemma.

Lemma 4.11 (Fourier transform of dispersion coefficients). *Let $\varepsilon > 0$ and $H, L \gg 1$. For $i \in \{1, 2\}$, let $\ell_i \in \mathbb{Z}_+$ with $\ell_i \asymp L$ and $(\ell_1, \ell_2) = 1$, $\alpha_i \in \mathbb{R}/\mathbb{Z}$, and $\Phi_i : (-\infty, \infty) \to \mathbb{C}$ be smooth functions, with $\Phi_i(t)$ supported in $t \ll 1$ and $\Phi_i^{(j)} \ll_j 1$ for all $j \geq 0$. Then for any $\varepsilon > 0$, the sequence*

$$a_n := \sum_{\substack{h_1, h_2 \in \mathbb{Z} \\ h_1\ell_1 + h_2\ell_2 = n}} \Phi_1\left(\frac{h_1}{H}\right)\Phi_2\left(\frac{h_2}{H}\right) e(h_1\alpha_1 + h_2\alpha_2),$$

supported in $n \ll HL$, has Fourier transform bounds

$$\widehat{a} \ll H^2, \qquad \widehat{a}(\alpha) \ll_\varepsilon H^{-100} \text{ unless } \|\ell_i\alpha - \alpha_i\| \leq H^{\varepsilon - 1}\ \forall i \in \{1, 2\}. \tag{4.10}$$

In consequence,

$$\|\widehat{a}\|_{L^1} \ll_\varepsilon H^\varepsilon\left(1 + \frac{H}{L}\right), \qquad \|\widehat{a}\|_{L^2} \ll_\varepsilon H^\varepsilon\left(H + \frac{H^{3/2}}{L^{1/2}}\right).$$

Proof of Lemma 4.11. We take $\varepsilon \in (0, 1)$ without loss of generality. The sequence (a_n) can be expressed as a discrete convolution,

$$a_n = a(n) = \sum_{m \in \mathbb{Z}} b_1(m)\, b_2(n - m) \qquad \Rightarrow \qquad \widehat{a}(\alpha) = \widehat{b_1}(\alpha) \cdot \widehat{b_2}(\alpha), \tag{4.11}$$

where for $i \in \{1, 2\}$,

$$b_i(n) := \mathbb{1}_{n \equiv 0 \pmod{\ell_i}}\, \Phi_i\left(\frac{n}{H\ell_i}\right) e\left(\frac{n}{\ell_i}\alpha_i\right).$$

But we further have

$$\widehat{b_i}(\alpha) = \widehat{c_i}(\ell_i\alpha - \alpha_i), \tag{4.12}$$

where $c_i(h) := \Phi_i(h/H)$. By Poisson summation and the Schwarz decay of $\widehat{\Phi}_i$, identifying $\alpha \in \mathbb{R}/\mathbb{Z}$ with $\alpha \in (-1/2, 1/2]$, we have

$$\begin{aligned}\widehat{c_i}(\alpha) = \sum_{h \in \mathbb{Z}} \Phi_i\left(\frac{h}{H}\right) e(-h\alpha) &= \sum_{n \in \mathbb{Z}} H\widehat{\Phi}_i\left(H(n + \alpha)\right) \\ &= H\widehat{\Phi}_i(H\alpha) + O\left(H^{-200}\right).\end{aligned}$$

In fact, we also have $H\widehat{\Phi_i}(H\alpha) = O_\varepsilon(H^{-200})$ when $|H\alpha| > H^\varepsilon$. So overall,

$$\begin{aligned} \widehat{c_i}(\alpha) &\ll H, && \forall\, \alpha \in \mathbb{R}/\mathbb{Z}, \\ \widehat{c_i}(\alpha) &\ll O_\varepsilon(H^{-200}), && \text{if } \|\alpha\| > H^{\varepsilon-1}. \end{aligned}$$

Thus by (4.11) and (4.12), we obtain

$$\widehat{a}(\alpha) \ll \begin{cases} H^2, & \max\left(\|\ell_1\alpha - \alpha_1\|, \|\ell_2\alpha - \alpha_2\|\right) \le H^{\varepsilon-1}, \\ O_\varepsilon(H^{-100}), & \max\left(\|\ell_1\alpha - \alpha_1\|, \|\ell_2\alpha - \alpha_2\|\right) > H^{\varepsilon-1}, \end{cases} \tag{4.13}$$

which proves (4.10). Now suppose that $\max\left(\|\ell_1\alpha - \alpha_1\|, \|\ell_2\alpha - \alpha_2\|\right) \le H^{\varepsilon-1}$; we would like to estimate how often this happens. Identifying $\alpha, \alpha_i \in \mathbb{R}/\mathbb{Z}$ with $\alpha, \alpha_i \in (-1/2, 1/2]$, there must exist integers $m_i(\alpha) \ll L$ such that

$$\ell_1\alpha - \alpha_1 = m_1 + O\left(H^{\varepsilon-1}\right), \qquad \ell_2\alpha - \alpha_2 = m_2 + O\left(H^{\varepsilon-1}\right),$$

so in particular,

$$\ell_1 m_2 - \ell_2 m_1 = \ell_2\alpha_1 - \ell_1\alpha_2 + O\left(H^{\varepsilon-1}L\right). \tag{4.14}$$

Since $\gcd(\ell_1, \ell_2) = 1$, as $m_1, m_2 \ll L$ vary, the difference $\ell_1 m_2 - \ell_2 m_1$ can only cover any given integer $O(1)$ times; thus there are a total of $O(1 + H^{\varepsilon-1}L)$ pairs $(m_1, m_2) \in \mathbb{Z}^2$ satisfying (4.14). Moreover, to each such pair (m_1, m_2) there can correspond an interval of α's of length at most $O(H^{\varepsilon-1}L^{-1})$, since

$$\alpha = \frac{m_1(\alpha) + \alpha_1}{\ell_1} + O\left(H^{\varepsilon-1}L^{-1}\right).$$

Overall, we obtain that the set

$$\left\{\alpha \in \mathbb{R}/\mathbb{Z} : \max\left(\|\ell_1\alpha - \alpha_1\|, \|\ell_2\alpha - \alpha_2\|\right) \le H^{\varepsilon-1}\right\}$$

has Lebesgue measure at most

$$O\left(\left(1 + H^{\varepsilon-1}L\right) \cdot H^{\varepsilon-1}L^{-1}\right) = O\left(H^{\varepsilon-1}L^{-1} + H^{2\varepsilon-2}\right).$$

By (4.13), we conclude that for any $p \ge 1$,

$$\begin{aligned} \|\widehat{a}\|_{L^p} &\ll_\varepsilon H^{O(\varepsilon)}\left(H^{2p}\left(H^{-1}L^{-1} + H^{-2}\right) + 1\right)^{\frac{1}{p}} \\ &\ll_p H^{O(\varepsilon)}\left(H^{2-\frac{2}{p}} + \frac{H^{2-\frac{1}{p}}}{L^{\frac{1}{p}}}\right), \end{aligned}$$

which completes our proof up to a rescaling of ε. □

Remark 4.12. As in [42], the arguments in this subsection extend immediately to sums of weighted Kloosterman sums

$$S_w(m, n; c) := \sum_{x \in (\mathbb{Z}/c\mathbb{Z})^\times} w(x)\, e\left(\frac{mx + n\overline{x}}{c}\right),$$

for arbitrary 1-bounded coefficients $w(x)$. In particular, choosing $w(x)$ in terms of a Dirichlet character χ mod q_0, where $q_0 \mid q \mid c$, should ultimately extend our large sieve inequalities to the exceptional Maass forms of level q associated to a general nebentypus χ, rather than the trivial one.

5. Spectral bounds

We now combine the combinatorial arguments from the previous section with techniques from the spectral theory of automorphic forms (inspired by [10]), to prove new large sieve inequalities for exceptional Maass cusp forms, and then to deduce bounds for multilinear forms of Kloosterman sums. The reader should be familiar with the prerequisites in all of Section 3, especially Section 3.4.

5.1. A general large sieve for exceptional Maass forms

Our common generalization of Theorems 1.5 and 1.7 requires the following notation, applied to the Fourier transform of a sequence (a_n).

Notation 5.1 (Rational-approximation integrals). Given $N \geq 1/2$ and a bounded-variation complex Borel measure μ on $\mathbb{R}/\mathbb{Z}$, we denote

$$\mathcal{I}_N(\mu) := \iint_{(\mathbb{R}/\mathbb{Z})^2} T_N(\alpha,\beta)\, d|\mu|(\alpha)\, d|\mu|(\beta),$$

recalling the definition of $T_N(\alpha,\beta)$ from Notation 4.1.

In general, the bound in (4.5) ensures that

$$\mathcal{I}_N(\mu) \ll \iint_{(\mathbb{R}/\mathbb{Z})^2} \min\left(\sqrt{N(1+\|\alpha-\beta\|N)}, N^{2/3}\right) d|\mu|(\alpha)\, d|\mu|(\beta), \tag{5.1}$$

which is invariant under translations of μ. Noting the trivial lower bound $T_N(\alpha,\beta) \geq 1$, this implies

$$|\mu|(\mathbb{R}/\mathbb{Z})^2 \ll \mathcal{I}_N(\mu) \ll N^{2/3}|\mu|(\mathbb{R}/\mathbb{Z})^2. \tag{5.2}$$

Theorem 5.2 (Large sieve with frequency concentration). *Let $\varepsilon > 0$, $X, A > 0$, $N \geq 1/2$, $q, a \in \mathbb{Z}_+$, and $(a_n)_{n\sim N}$ be a complex sequence. Let $f : (0,4) \to \mathbb{C}$ be a smooth function with $f^{(j)} \ll_j 1$ for $j \geq 0$, and μ be a bounded-variation complex Borel measure on $\mathbb{R}/\mathbb{Z}$, such that*[1]

$$a_n = f\left(\frac{n}{N}\right)\check{\mu}(n),$$

for all $n \sim N$ (in particular, one can take $f \equiv 1$, $d\mu = \widehat{a}\, d\lambda$). Let $\mathfrak{a}, \rho_{j\mathfrak{a}}(n), \lambda_j, \theta_j$ be as in Theorem 1.2, with $\mu(\mathfrak{a}) = q^{-1}$ and the choice of scaling matrix $\sigma_{\mathfrak{a}}$ in (3.9). Then one has

$$\sum_{\lambda_j<1/4} X^{2\theta_j}\left|\sum_{n\sim N} a_n \rho_{j\mathfrak{a}}(an)\right|^2 \ll_\varepsilon (qaNX)^{2\varepsilon}\left(1+\frac{aN}{q}\right)A^2, \tag{5.3}$$

provided that both of the following hold:

$$A \gg \|a_n\|_2 + \frac{\sqrt{\gcd(a,q)}N}{\sqrt{q+aN}}|\mu|(\mathbb{R}/\mathbb{Z}), \tag{5.4}$$

$$X \ll \max\left(1, \frac{q}{aN}\right)\max\left(1, \frac{A^2}{\mathcal{I}_N(\mu)}\right). \tag{5.5}$$

[1] We slightly abuse notation in this section: the measure μ should not be confused with the cusp parameter $\mu(\mathfrak{a}) = q^{-1}$, and the scalar a should not be confused with the sequence (a_n).

Remark 5.3. Theorem 5.2 obtains a saving over Theorem 1.2 whenever we can take $A \ll (qN)^{o(1)}\|a_n\|_2$ and $X > \max(1, \frac{q}{aN})$. To satisfy (5.4) and (5.5) in this context, assuming $\gcd(a,q) = 1$, we need

$$|\mu|(\mathbb{R}/\mathbb{Z}) \ll (qN)^{o(1)}\frac{\sqrt{q+aN}}{N}\|a_n\|_2 \qquad \text{and} \qquad \mathcal{I}_N(\mu) = o\left(\|a_n\|_2^2\right). \tag{5.6}$$

These should be compared with the lower bound

$$|\mu|(\mathbb{R}/\mathbb{Z}) \gg N^{-1/2}\|a_n\|_2, \tag{5.7}$$

which always holds, by Fourier expansion and Cauchy–Schwarz. This has the following implications:

(1). From (5.7) and the lower bound in (5.2), we have $\mathcal{I}_N(\mu) \gg N^{-1/2}\|a_n\|_2$. With $A \ll (qN)^{o(1)}\|a_n\|_2$, this limits the range of X in (5.5) to the best-case scenario $X \ll \max(N, \frac{q}{a})$. This is indeed achieved by Theorem 1.5 when $\alpha = 0$.

(2). When $a \ll 1$ and $q \approx N$, (5.6) requires nearly-optimal concentration for μ, in the sense that $|\mu|(\mathbb{R}/\mathbb{Z})$ is almost as small as possible; this happens to hold for the sequences in (2.10).

(3). Using the upper bound $\mathcal{I}_N(\mu) \ll N^{2/3}|\mu|(\mathbb{R}/\mathbb{Z})^2$ from (5.2) and choosing $f \equiv 1$, $d\mu = \widehat{a}\,d\lambda$ (so that $|\mu|(\mathbb{R}/\mathbb{Z}) = \|\widehat{a}\|_{L^1}$ and $\|a_n\|_2 = \|\widehat{a}\|_{L^2}$), we see that (5.6) holds in particular when

$$\frac{\|\widehat{a}\|_{L^1}}{\|\widehat{a}\|_{L^2}} = o\left(\frac{\min(q^{1/2+o(1)}, (aN)^{1/2+o(1)}, N^{2/3})}{N}\right),$$

which gives a more palpable concentration condition on the Fourier transform $\widehat{a}$. The weights of $T_N(\alpha,\beta)$ inside $\mathcal{I}_N(\mu)$, combined with the liberty to choose other measures μ and functions f, allow for additional flexibility when more information about the sequence (a_n) is available.

Proof of Theorem 5.2. We assume without loss of generality that $\varepsilon < 1$, and that f is supported in $[0.5, 3]$ (otherwise, multiply f by a fixed smooth function supported in $[0.5,3]$ and equal to 1 on $[1,2]$; then the identity $a_n = f(n/N)\,\check{\mu}(n)$ remains true for $n \sim N$).

In light of Lemma (3.9), we are immediately done if $X \le 1$, so assume $X > 1$. Let Φ be a fixed nonnegative smooth function supported in $[2,4]$, with positive integral. Then by Corollary 3.10 and the fact that $A \gg \|a_n\|_2$ (from (5.4)), it suffices to show that

$$\mathcal{S} := \sum_{c\in\mathcal{C}_{\mathfrak{a}\mathfrak{a}}}\frac{1}{c}\sum_{m,n\sim N}\overline{a_m}\,a_n\,S_{\mathfrak{a}\mathfrak{a}}(am,an;c)\,\Phi\left(\frac{a\sqrt{mn}}{c}X\right) \ll_\varepsilon (qaNX)^{2\varepsilon}\left(1+\frac{aN}{q}\right)A^2, \tag{5.8}$$

subject to (5.4) and (5.5). Since $\mu(\mathfrak{a}) = q^{-1}$, Lemma 3.2 implies that

$$\mathcal{S} = \sum_{\substack{c\in(aNX/4,aNX)\\ c\equiv 0 \pmod q}}\frac{\mathcal{S}(c)}{c}, \tag{5.9}$$

where

$$\mathcal{S}(c) := \sum_{m,n\sim N}\overline{a_m}\,a_n\,S(am,an;c)\,\Phi\left(\frac{a\sqrt{mn}}{c}X\right).$$

If $aNX \le q$, the sum over c is void; so we may assume that $X > \max\left(1, \frac{q}{aN}\right)$, which by (5.5) implies

$$\mathcal{I}_N(\mu) \ll A^2. \tag{5.10}$$

We aim to bound each of the $\asymp aNX/q$ inner sums $\mathcal{S}(c)$ separately, using Proposition 4.9. To this end, we need to separate the variables m, n, c; we can rewrite

$$\mathcal{S}(c) = \sum_{m,n\sim N} \overline{\breve{\mu}(m)}\, \breve{\mu}(n)\, S(am, an; c)\, \Psi_c\left(\frac{m}{N}, \frac{n}{N}\right), \tag{5.11}$$

where

$$\Psi_c(x_1, x_2) := \overline{f(x_1)} f(x_2)\, \Phi\left(\sqrt{x_1 x_2}\frac{aNX}{c}\right)$$

is a compactly-supported smooth function with bounded derivatives (since $c \asymp aNX$ and we assumed WLOG that f is supported in $[0.5, 3]$). By two-dimensional Fourier inversion, we have

$$\Psi_c(x_1, x_2) = \iint_{\mathbb{R}^2} \widehat{\Psi_c}(t_1, t_2)\, e(t_1 x_1 + t_2 x_2)\, dt_1\, dt_2,$$

where

$$\widehat{\Psi_c}(t_1, t_2) = \iint_{(0,\infty)^2} \Psi_c(x_1, x_2)\, e(-t_1 x_1 - t_2 x_2)\, dx_1\, dx_2.$$

Since $\Psi_c(x_1, x_2)$ is Schwarz, so is $\widehat{\Psi}(t_1, t_2)$; in particular, we have $\widehat{\Psi_c}(t_1, t_2) \ll (1 + t_1^4)^{-1}(1 + t_2^4)^{-1}$ with an absolute implied constant. Plugging the inversion formula into (5.11) and swapping sums and integrals, we obtain

$$\mathcal{S}(c) = \iint_{\mathbb{R}^2} \widehat{\Psi_c}(t_1, t_2)\, \mathcal{S}(c, t_1, t_2)\, dt_1\, dt_2, \tag{5.12}$$

where

$$\mathcal{S}(c, t_1, t_2) := \sum_{m,n\sim N} \overline{\breve{\mu}(m)\, e\left(\frac{-mt_1}{N}\right)}\, \breve{\mu}(n)\, e\left(\frac{nt_2}{N}\right) S(am, an; c).$$

Note that translating μ corresponds to multiplying $\breve{\mu}(n)$ by exponential factors $e(n\alpha)$, so Proposition 4.9 and a change of variables yield

$$\begin{aligned}
&\mathcal{S}(c, t_1, t_2)\\
&\ll_\varepsilon c^\varepsilon \iint_{(\mathbb{R}/\mathbb{Z})^2} \left(cT_N(\alpha, \beta) + \gcd(a, c)N^2\right) d|\mu|\left(-\alpha + \frac{t_1}{N}\right) d|\mu|\left(\beta - \frac{t_2}{N}\right)\\
&= c^\varepsilon \iint_{(\mathbb{R}/\mathbb{Z})^2} \left(cT_N\left(-\alpha + \frac{t_1}{N}, \beta + \frac{t_2}{N}\right) + \gcd(a, c)N^2\right) d|\mu|(\alpha)\, d|\mu|(\beta).
\end{aligned}$$

Recalling that $T_N(\alpha, \beta) = T_N(-\alpha, \beta)$ and the bound (4.3), we have

$$T_N\left(-\alpha + \frac{t_1}{N}, \beta + \frac{t_2}{N}\right) \ll (1 + |t_1|)(1 + |t_2|)\, T_N(\alpha, \beta),$$

so that

$$\mathcal{S}(c, t_1, t_2) \ll_\varepsilon (1 + |t_1|)(1 + |t_2|)\, c^\varepsilon \left(c\, \mathcal{I}_N(\mu) + \gcd(a, c)N^2 |\mu|(\mathbb{R}/\mathbb{Z})^2\right).$$

Together with (5.12) and the bound $\widehat{\Psi_c}(t_1,t_2) \ll (1+t_1^4)^{-1}(1+t_2^4)^{-1}$, we obtain

$$\mathcal{S}(c) \ll_\varepsilon c^\varepsilon \left(c\,\mathcal{I}_N(\mu) + \gcd(a,c)N^2|\mu|(\mathbb{R}/\mathbb{Z})^2 \right),$$

and by (5.9) we conclude that

$$\mathcal{S} \ll_\varepsilon (aNX)^{2\varepsilon} \left(\frac{aNX}{q}\mathcal{I}_N(\mu) + \frac{\gcd(a,q)N^2}{q}|\mu|(\mathbb{R}/\mathbb{Z})^2 \right). \tag{5.13}$$

By the assumed lower bound for A in (5.4), the contribution of the second term is

$$\ll_\varepsilon (aNX)^{2\varepsilon} \left(1 + \frac{aN}{q}\right) A^2,$$

which is acceptable in (5.8). Similarly, the first term in (5.13) is acceptable provided that

$$\frac{aNX}{q}\mathcal{I}_N(\mu) \ll \left(1 + \frac{aN}{q}\right) A^2,$$

that is,

$$X \ll \max\left(1, \frac{q}{aN}\right) \frac{A^2}{\mathcal{I}_N(\mu)},$$

which follows from (5.5) and (5.10). □

5.2. *Proofs of Theorems 1.5 and 1.7*

We can now deduce the large sieve inequalities promised in Section 1.1, starting from Theorem 5.2.

Proof of Theorem 1.5. Consider the sequence $a_n := \Phi(n/N)\,e(n\alpha)$ for $n \sim N$ and some $\alpha \in \mathbb{R}/\mathbb{Z}$, which has $\|a_n\|_2 \asymp \sqrt{N} =: A$. Choosing $\mu := \delta_{\{\alpha\}}$, we have $a_n = \Phi(n/N)\,\check{\mu}(n)$ for $n \sim N$, and $|\mu|(\mathbb{R}/\mathbb{Z}) = 1$. In particular, the lower bound for A in (5.4) holds for any values of q and a, since

$$|\mu|(\mathbb{R}/\mathbb{Z}) = 1 \ll N^{-1/2}\,\|a_n\|_2.$$

Finally, we have

$$\mathcal{I}_N(\mu) = T_N(\alpha,\alpha) \asymp \min_{t\in\mathbb{Z}_+}(t + N\|t\alpha\|),$$

so Theorem 5.2 (i.e., (5.5)) recovers the large sieve range

$$X \ll \max\left(1, \frac{q}{aN}\right) \frac{N}{\min_{t\in\mathbb{Z}_+}(t + N\|t\alpha\|)}$$

from (1.7). In particular, we can recall from (4.5) that $T_N(\alpha,\alpha) \ll \sqrt{N}$, so this includes the range $X \ll (\sqrt{N}, \frac{q}{a\sqrt{N}})$ uniformly in α. Since varying the choice of scaling matrix $\sigma_{\mathfrak{a}}$ is equivalent to varying α, we can use the same range $X \ll (\sqrt{N}, \frac{q}{a\sqrt{N}})$ for an arbitrary scaling matrix. □

Proof of Theorem 1.7. Assume without loss of generality that $\varepsilon \in (0,1)$. By changing $h_2 \leftrightarrow -h_2$, $\Phi_2(t) \leftrightarrow \Phi_2(-t)$ and $\alpha_2 \leftrightarrow -\alpha_2$, we can equivalently consider the sequence $(a_n)_{n\sim N}$ given by

$$a_n = \sum_{\substack{h_1,h_2\in\mathbb{Z}\\ h_1\ell_1+h_2\ell_2=n}} \Phi_1\left(\frac{h_1}{H}\right)\Phi_2\left(\frac{h_2}{H}\right) e(h_1\alpha_1 + h_2\alpha_2).$$

We may of course assume that $N \ll HL$, since otherwise $(a_n)_{n\sim N}$ vanishes. Note that the extension $(a_n)_{n\in\mathbb{Z}}$ is exactly the sequence considered in Lemma 4.11. Thus letting $\varphi : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$ be the Fourier transform of $(a_n)_{n\in\mathbb{Z}}$, and $d\mu := \varphi\, d\lambda$ (where λ is the Lebesgue measure on $\mathbb{R}/\mathbb{Z}$), we have

$$\check{\mu}(n) = \check{\varphi}(n) = a_n, \qquad \forall n \sim N.$$

Moreover, Lemma 4.11 implies that

$$\varphi \ll H^2, \qquad \varphi(\alpha) \ll_\varepsilon H^{-100} \text{ unless } \|\ell_i\alpha - \alpha_i\| \le H^{\varepsilon-1}\ \forall i \in \{1,2\}, \tag{5.14}$$

and

$$|\mu|(\mathbb{R}/\mathbb{Z}) = \|\varphi\|_{L^1} \ll_\varepsilon H^\varepsilon \left(1 + \frac{H}{L}\right). \tag{5.15}$$

To compute the integral

$$\mathcal{I}_N(\mu) = \iint_{(\mathbb{R}/\mathbb{Z})^2} T_N(\alpha,\beta)\, \varphi(\alpha)\, \varphi(\beta)\, d\alpha\, d\beta,$$

we first consider the contribution of α, β which have $\|\ell_i\alpha - \alpha_i\| > H^{\varepsilon-1}$ or $\|\ell_i\beta - \alpha_i\| > H^{\varepsilon-1}$ for some $i \in \{1,2\}$. By (5.14), either $\varphi(\alpha)$ or $\varphi(\beta)$ is $\ll_\varepsilon H^{-100}$ in this case, so the total contribution to $\mathcal{I}_N(\mu)$ is

$$\ll_\varepsilon N^{2/3} H^{-100} H^2 \ll LH^{-90}.$$

On the other hand, when $\max_{i\in\{1,2\}} \max(\|\ell_i\alpha - \alpha_i\|, \|\ell_i\beta - \alpha_i\|) \le H^{\varepsilon-1}$, we have by definition (Notation 4.1) that for any $t \in \mathbb{Z}_+$,

$$\begin{aligned} T_N(\alpha,\beta) &\le t\ell_i + N\|t\ell_i\alpha\| + N\|t\ell_i\beta\| \\ &\ll tL + N\|t(\ell_i\alpha - \alpha_i)\| + N\|t(\ell_i\beta - \alpha_i)\| + N\|t\alpha_i\| \\ &\ll tL + NtH^{\varepsilon-1} + N\|t\alpha_i\| \\ &\ll H^\varepsilon tL + N\|t\alpha_i\| \\ &\ll H^\varepsilon L\left(t + \frac{N}{L}\|t\alpha_i\|\right). \end{aligned}$$

Taking a minimum over $t \in \mathbb{Z}_+$ and $i \in \{1,2\}$, we obtain

$$T_N(\alpha,\beta) \ll H^\varepsilon LM, \qquad M := \min_{i\in\{1,2\}} T_{N/L}(\alpha_i).$$

Using (5.15), we conclude that

$$\begin{aligned} \mathcal{I}_N(\mu) = \iint_{(\mathbb{R}/\mathbb{Z})^2} T_N(\alpha,\beta)\, d|\mu|(\alpha)\, d|\mu|(\beta) &\ll_\varepsilon LH^{-90} + H^\varepsilon LM\, |\mu|(\mathbb{R}/\mathbb{Z})^2 \\ &\ll_\varepsilon H^{2\varepsilon} LM \left(1 + \frac{H}{L}\right)^2. \end{aligned} \tag{5.16}$$

We are now in a position to apply Theorem 5.2, with

$$\frac{A}{C_\varepsilon H^\varepsilon} := \|a_n\|_2 + \sqrt{\gcd(a,q)N}\left(\sqrt{\frac{H}{L}} + \frac{H}{L}\right),$$

where C_ε is a sufficiently large constant. Note that by (5.15), the assumption $q \gg L^2$, and the fact that $N \ll HL$, we have

$$\begin{aligned}|\mu|(\mathbb{R}/\mathbb{Z}) &\ll C_\varepsilon H^\varepsilon \left(1 + \frac{H}{L}\right) \\ &\ll C_\varepsilon H^\varepsilon \frac{L + \sqrt{N}}{\sqrt{N}} \left(\sqrt{\frac{H}{L}} + \frac{H}{L}\right) \ll \frac{\sqrt{q + aN}}{\sqrt{\gcd(a,q)N}} A,\end{aligned}$$

so the lower bound for A in (5.4) holds (above we used that $\frac{L}{\sqrt{N}}\sqrt{\frac{H}{L}} = \sqrt{\frac{HL}{N}} \gg 1$). It follows that the large sieve bound (5.3) holds for all

$$X \ll \max\left(1, \frac{q}{aN}\right) \max\left(1, \frac{A^2}{\mathcal{I}_N(\mu)}\right),$$

where by (5.16),

$$\frac{A^2}{\mathcal{I}_N(\mu)} \gg \frac{H^{2\varepsilon} N \left(\frac{H}{L} + \frac{H^2}{L^2}\right)}{H^{2\varepsilon} LM \left(1 + \frac{H}{L}\right)^2} = \frac{NH}{(H+L)LM}.$$

This proves (1.8). □

5.3. *Multilinear Kloosterman bounds*

In contrast to the "vertical" bilinear averages of Kloosterman sums $S(m, n; c)$ over m, n from Section 4 (or from [29, 27]), the bounds in this subsection also require "horizontal" averaging over the modulus c – crucially, with a smooth weight in this variable. Generally, it is such horizontal averages that make use of the Kuznetsov trace formula for $\Gamma_0(q)$, leading to dependencies on the spectral parameter $\theta(q) = \sqrt{\max(0, \frac{1}{4} - \lambda_1(q))} \leq \frac{7}{64}$; we recall that the purpose of large sieve inequalities for the exceptional spectrum, like Theorem 5.2, is to improve the dependency on $\theta(q)$.

Throughout this subsection, we will work with sequences obeying the following condition.

Assumption 5.4 (Large sieve for the tuple $(q, N, Z, (a_n)_{n\sim N}, A_N, Y_N)$). This applies to complex sequences $(a_n)_{n\sim N}$ and parameters $q \in \mathbb{Z}_+$, $N \geq 1/2$, $Z \gg 1$, $A_N \gg \|a_n\|_2$, $Y_N > 0$. For any $\varepsilon > 0$, $\xi \in \mathbb{R}$, any cusp $\mathfrak{a}$ of $\Gamma_0(q)$ with $\mu(\mathfrak{a}) = q^{-1}$ and $\sigma_\mathfrak{a}$ chosen as in (3.9), and any orthonormal basis of Maass cusp forms for $\Gamma_0(q)$, with eigenvalues λ_j and Fourier coefficients $\rho_{j\mathfrak{a}}(n)$, one has

$$\sum_{\lambda_j < 1/4} X^{2\theta_j} \left| \sum_{n\sim N} e\left(\frac{n}{N}\xi\right) a_n \rho_{j\mathfrak{a}}(n) \right|^2 \ll_\varepsilon (qNZ)^\varepsilon \left(1 + \frac{N}{q}\right) A_N^2, \tag{5.17}$$

for all $X \ll \max\left(1, \frac{q}{N}\right) \frac{Y_N}{1+|\xi|^2}$.

For example, Theorem 1.2 shows that the tuple $(q, N, 1, (a_n)_{n\sim N}, \|a_n\|_2^2, 1)$ satisfies Assumption 5.4 for any $q \in \mathbb{Z}_+$, $N \geq 1/2$ and any complex sequence $(a_n)_{n\sim N}$; attaining higher values of Y_N requires more information about (a_n). Theorem 1.5 implies that another suitable choice of parameters is

$$a_n := e(n\alpha), \qquad Y_N := \frac{N}{T_N(\alpha)} \gg \sqrt{N}, \qquad A_N := \sqrt{N}, \tag{5.18}$$

for any $\alpha \in \mathbb{R}/\mathbb{Z}$ and $q \in \mathbb{Z}_+$, $N \geq 1/2$, $Z = 1$; note that the phase ξ/N can be incorporated into α, and we implicitly used that $T_N(\alpha + \xi/N) \ll (1 + |\xi|^2) T_N(\alpha)$ by (4.3). Likewise, incorporating $\ell_i \xi/N$ into

α_i, Theorem 1.7 shows that we can choose

$$\begin{aligned} a_n &:= \sum_{\substack{h_1,h_2\in\mathbb{Z}\\ h_1\ell_1-h_2\ell_2=n}} \Phi_1\left(\frac{h_1}{H}\right)\Phi_2\left(\frac{h_2}{H}\right) e(h_1\alpha_1+h_2\alpha_2),\\ Y_N &:= \max\left(1, \frac{NH}{(H+L)L\min_i T_H(\alpha_i)}\right), \qquad A_N := \|a_n\|_2 + \sqrt{N}\sqrt{\frac{H}{L}+\frac{H^2}{L^2}}, \end{aligned} \tag{5.19}$$

where $1 \ll L^2 \ll q$, $1 \ll H \ll Z$, $\alpha_i \in \mathbb{R}/\mathbb{Z}$, $\ell_i \asymp L$, $(\ell_1,\ell_2)=1$, and $\Phi_i(t)$ are smooth functions supported in $t \ll 1$ with $\Phi_i^{(j)} \ll_j 1$. Other than the input from Assumption 5.4 (and implicitly Theorems 1.5 and 1.7), all arguments in this subsection are fairly standard [10, 14, 8].

Corollary 5.5 (Kloosterman bounds with averaging over n, c). *Let $(q, N, Z, (a_n)_{n\sim N}, A_N, Y_N)$ satisfy Assumption 5.4. Let $\varepsilon > 0$, $C \gg 1$, $m \in \mathbb{Z}_+$, and $\mathfrak{a}, \mathfrak{b}$ be cusps of $\Gamma_0(q)$, with $\mu(\mathfrak{a}) = \mu(\mathfrak{b}) = q^{-1}$ and $\sigma_{\mathfrak{b}}$ as in (3.9). Let $\Phi : (0,\infty)^2 \to \mathbb{C}$ be a smooth function, with $\Phi(x,y)$ supported in $x, y \asymp 1$, and $\partial_x^j\partial_y^k\Phi(x,y) \ll_{j,k,\varepsilon} Z^{j\varepsilon}$ for $j,k \geq 0$. Then with a consistent choice of the $\pm$ sign, one has*

$$\begin{aligned} \sum_{n\sim N} a_n \sum_{c\in\mathcal{C}_{\mathfrak{a}\mathfrak{b}}} \Phi\left(\frac{n}{N},\frac{c}{C}\right) S_{\mathfrak{a}\mathfrak{b}}(m,\pm n;c) &\ll_\varepsilon (qmNCZ)^{O(\varepsilon)}(1+T)^{2\theta(q)}\frac{C^2A_N}{C+\sqrt{mN}}\\ &\times\left(1+\frac{mN}{C^2}+\frac{\sqrt{(q,m)m}}{q}\right)^{1/2}\left(1+\frac{mN}{C^2}+\frac{N}{q}\right)^{1/2}, \end{aligned} \tag{5.20}$$

for

$$T = \frac{T_0}{\sqrt{Y_N}}, \qquad T_0 := \frac{C}{\max\left(m, q^2(q,m)^{-1}\right)^{1/2}\max(N,q)^{1/2}} \leq \frac{C}{q^{3/2}(q,m)^{-1/2}}.$$

Remark 5.6. The parameter T_0 indicates the best known dependency on $\theta = \theta(q)$ that one could achieve without our large sieve inequalities; for example, when $a_n = e(n\alpha)$ and $Y_N = \sqrt{N}$, Corollary 5.5 saves a total factor of $N^{\theta/2}$ over previous bounds (and up to N^θ if α is close to a rational number of small denominator). We note that in practice, the second term in each maximum from T_0 is usually dominant, and the factors in the second line of (5.20) are typically $\asymp 1$.

Remark 5.7. While the smooth weight in the c variable is necessary here (stemming from Proposition 3.5), the smooth weight in n only confers additional flexibility. Indeed, one can take $\Phi(x,y) = f(x)g(y)$ for compactly-supported functions $f, g : (0,\infty) \to \mathbb{C}$, where $f \equiv 1$ on $(1,2)$; this effectively replaces $\Phi(n/N, c/C)$ with $g(c/C)$ in (5.20). The same remark applies to the next results.

Proof of Corollary 5.5. Denote by $\mathcal{S}$ the sum in (5.20). Letting $\Psi(x;y) := \sqrt{x}\,\Phi(x,\sqrt{x}/y)$, we can Fourier expand

$$\sqrt{x}\,\Phi\left(x,\frac{\sqrt{x}}{y}\right) = \int_{\mathbb{R}} \widehat{\Psi}(\xi;y)\,e(x\xi)\,d\xi,$$

where the Fourier transform is taken in the first variable. Integrating by parts in x, we note that for $k \geq 0$,

$$\partial_y^k\,\widehat{\Psi}(\xi;y) \ll_{j,\varepsilon} \frac{Z^{O(\varepsilon)}}{1+\xi^4},$$

where the implied constant in $O(\varepsilon)$ (say, $K > 0$) does not depend on k. Then we can let

$$\varphi_\xi(y) := Z^{-K\varepsilon}\left(1+\xi^4\right)\widehat{\Psi}\left(\xi; y\frac{C}{4\pi\sqrt{mN}}\right)\frac{4\pi\sqrt{mN}}{Cy},$$

which is supported in $y \asymp X^{-1}$ and satisfies $\varphi_\xi^{(k)} \ll_{k,\varepsilon} X^k$, for

$$X := \frac{C}{\sqrt{mN}}. \tag{5.21}$$

This way, we can rewrite

$$\begin{aligned}\Phi\left(\frac{n}{N}, \frac{c}{C}\right) &= \int_{\mathbb{R}} \sqrt{\frac{N}{n}} \widehat{\Psi}\left(\xi; \frac{C}{c}\sqrt{\frac{n}{N}}\right) e\left(\frac{n}{N}\xi\right) d\xi \\ &= Z^{K\varepsilon} \frac{C}{c} \int_{\mathbb{R}} \frac{1}{1+\xi^4} e\left(\frac{n}{N}\xi\right) \varphi_\xi\left(\frac{4\pi\sqrt{mn}}{c}\right) d\xi,\end{aligned}$$

and thus

$$\mathcal{S} \ll_\varepsilon Z^{O(\varepsilon)} C \int_{\mathbb{R}} \frac{|S(\xi)|\, d\xi}{1+\xi^4}, \tag{5.22}$$

where

$$S(\xi) := \sum_{n\sim N} e\left(\frac{n}{N}\xi\right) a_n \sum_{c \in \mathcal{C}_{\mathfrak{a}\mathfrak{b}}} \frac{\mathcal{S}_{\mathfrak{a}\mathfrak{b}}(m, \pm n; c)}{c} \varphi_\xi\left(\frac{4\pi\sqrt{mn}}{c}\right).$$

The inner sum is in a suitable form to apply the Kuznetsov trace formula from Proposition 3.5. We only show the case when the choice of the $\pm$ sign is positive; the negative case is analogous (and in fact simpler due to the lack of holomorphic cusp forms). The resulting contribution of the Maass cusp forms to $\mathcal{S}(\xi)$ is

$$\mathcal{S}_{\mathcal{M}}(\xi) \ll \sum_{j=1}^{\infty} \frac{|\widehat{\mathcal{B}}_{\varphi_\xi}(\kappa_j)|}{\cosh(\pi\kappa_j)} |\rho_{j\mathfrak{a}}(m)| \left|\sum_{n\sim N} e\left(\frac{n}{N}\xi\right) a_n \rho_{j\mathfrak{b}}(n)\right| =: \mathcal{S}_{\mathcal{M},\mathrm{exc}}(\xi) + \mathcal{S}_{\mathcal{M},\mathrm{reg}}(\xi),$$

where $\mathcal{S}_{\mathcal{M},\mathrm{exc}}$ contains the terms with $\lambda_j < 1/4$ and $\mathcal{S}_{\mathcal{M},\mathrm{reg}}$ contains the rest. We first bound $\mathcal{S}_{\mathcal{M},\mathrm{reg}}$; the contribution of the holomorphic cusp forms and Eisenstein series is bounded analogously. For the Bessel transforms, we apply (3.20) if $|r| \le R$ and (3.21) otherwise, where $R \ge 1$ will be chosen shortly. Together with Cauchy–Schwarz and the bounds in Lemma 3.8 (in m) and Lemma (3.9) (in $n \sim N$), this yields

$$\begin{aligned}\mathcal{S}_{\mathcal{M},\mathrm{reg}}(\xi) \ll_\varepsilon (qmNR)^\varepsilon &\left(\frac{1+|\log X|}{1+X^{-1}} + R^{-5/2} + R^{-3}X^{-1}\right) \\ &\times \left(R^2 + \frac{\sqrt{(q,m)m}}{q}\right)^{1/2} \left(R^2 + \frac{N}{q}\right)^{1/2} \|a_n\|_2.\end{aligned}$$

Picking $R := 1 + X^{-1}$, we get

$$\mathcal{S}_{\mathcal{M},\mathrm{reg}}(\xi) \ll_\varepsilon (qmNC)^{O(\varepsilon)} \frac{1}{1+X^{-1}} \left(1 + X^{-2} + \frac{\sqrt{(q,m)m}}{q}\right)^{1/2} \left(1 + X^{-2} + \frac{N}{q}\right)^{1/2} \|a_n\|_2. \tag{5.23}$$

For the exceptional spectrum, we let $X = X_0\sqrt{X_1X_2}$ for $X_1, X_2 \gg 1$ to be chosen shortly, and note the bound

$$1 + X^{2\theta_j} \ll (1+X_0)^{2\theta_j} X_1^{\theta_j} X_2^{\theta_j} \ll (1+X_0)^{2\theta(q)} X_1^{\theta_j} X_2^{\theta_j}.$$

Then by (3.19) and Cauchy–Schwarz, we obtain

$$\begin{aligned}\mathcal{S}_{\mathcal{M},\mathrm{exc}}(\xi) &\ll \frac{1}{1+X^{-1}} \sum_{\lambda_j<1/4} \frac{1+X^{2\theta_j}}{\cosh(\pi\kappa_j)} |\rho_{j\mathfrak{a}}(m)| \left| \sum_{n\sim N} e\left(\frac{n}{N}\xi\right) a_n \rho_{j\mathfrak{b}}(n) \right| \\ &\ll \frac{(1+X_0)^{2\theta(q)}}{1+X^{-1}} \left(\sum_{\lambda_j<1/4} X_1^{2\theta_j} |\rho_{j\mathfrak{a}}(m)|^2 \right)^{1/2} \left(\sum_{\lambda_j<1/4} X_2^{2\theta_j} \left| \sum_{n\sim N} e\left(\frac{n}{N}\xi\right) a_n \rho_{j\mathfrak{b}}(n) \right|^2 \right)^{1/2}. \end{aligned} \tag{5.24}$$

We pick X_1 and X_2 as large as (3.28) and Assumption 5.4 allow, specifically

$$X_1 := \max\left(1, \frac{q^2}{(q,m)m}\right), \qquad X_2(\xi) := \max\left(1, \frac{q}{N}\right) \frac{Y_N}{1+|\xi|^2}. \tag{5.25}$$

Then by Lemma 3.8 and Assumption 5.4, we obtain

$$\mathcal{S}_{\mathcal{M},\mathrm{exc}}(\xi) \ll_\varepsilon (qmNC)^{O(\varepsilon)} \left(1 + \frac{X}{\sqrt{X_1 X_2(\xi)}}\right)^{2\theta(q)} \frac{1}{1+X^{-1}} \left(1 + \frac{\sqrt{(q,m)m}}{q}\right)^{1/2} \left(1+\frac{N}{q}\right)^{1/2} A_N. \tag{5.26}$$

Putting together (5.23) (and the identical bounds for Eisenstein series and holomorphic cusp forms) with (5.26) and (5.22), while noting that $\|a_n\|_2 \ll A_N$ by Assumption 5.4, we conclude that

$$\begin{aligned}\mathcal{S} \ll_\varepsilon (qmNCZ)^{O(\varepsilon)} &\left(1 + \frac{X}{\sqrt{X_1 X_2(0)}}\right)^{2\theta(q)} \frac{C}{1+X^{-1}} \\ &\times \left(1 + X^{-2} + \frac{\sqrt{(q,m)m}}{q}\right)^{1/2} \left(1 + X^{-2} + \frac{N}{q}\right)^{1/2} A_N, \end{aligned} \tag{5.27}$$

where the factor of $1+|\xi|^2$ inside $X_2(\xi)$ disappeared in the integral over ξ with a greater decay. This recovers the desired bound after plugging in the values of X, X_1, X_2 from (5.21) and (5.25). □

Remark 5.8. In treating the regular spectrum, we picked a slightly suboptimal value of R (following [10, p. 268]), to simplify the final bounds; in practice, this does not usually matter since one has $X \gg 1$.

Corollary 5.9 (Kloosterman bounds with averaging over m, n, c). *Let $(q, M, Z, (a_m)_{m\sim M}, A_M, Y_M)$ and $(q, N, Z, (b_n)_{n\sim N}, A_N, Y_N)$ satisfy Assumption 5.4. Let $\varepsilon > 0$, $C \gg 1$, $m \in \mathbb{Z}_+$, and $\mathfrak{a}, \mathfrak{b}$ be cusps of $\Gamma_0(q)$, with $\mu(\mathfrak{a}) = \mu(\mathfrak{b}) = q^{-1}$ and $\sigma_\mathfrak{a}, \sigma_\mathfrak{b}$ as in (3.9). Let $\Phi : (0,\infty)^3 \to \mathbb{C}$ be a smooth function, with $\Phi(x,y,z)$ supported in $x, y, z \asymp 1$, and $\partial_x^j \partial_y^k \partial_z^\ell \Phi(x,y,z) \ll_{j,k,\ell,\varepsilon} Z^{(j+k)\varepsilon}$ for $j, k, \ell \geq 0$. Then with a consistent choice of the $\pm$ sign, one has*

$$\begin{aligned}\sum_{m\sim M} a_m \sum_{n\sim N} b_n \sum_{c\in\mathcal{C}_{\mathfrak{a}\mathfrak{b}}} \Phi\left(\frac{m}{M}, \frac{n}{N}, \frac{c}{C}\right) S_{\mathfrak{a}\mathfrak{b}}(m, \pm n; c) &\ll_\varepsilon (qMNCZ)^{O(\varepsilon)} (1+T)^{2\theta(q)} \\ &\times \frac{C^2 A_M A_N}{C + \sqrt{MN}} \left(1 + \frac{MN}{C^2} + \frac{M}{q}\right)^{1/2} \left(1 + \frac{MN}{C^2} + \frac{N}{q}\right)^{1/2}, \end{aligned} \tag{5.28}$$

for

$$T = \frac{T_0}{\sqrt{Y_M Y_N}}, \qquad T_0 := \frac{C}{\max(M,q)^{1/2} \max(N,q)^{1/2}} \leq \frac{C}{q}.$$

In particular, for relatively prime positive integers r, s with $rs = q$, one has

$$\sum_{m\sim M} a_m \sum_{n\sim N} b_n \sum_{(c,r)=1} \Phi\Big(\frac{m}{M}, \frac{n}{N}, \frac{c}{C}\Big) S(m\bar{r}, \pm n; sc) \ll_\varepsilon (rsMNCZ)^{O(\varepsilon)} \Big(1 + \frac{C}{\sqrt{rY_M Y_N}}\Big)^{2\theta(q)} \times A_M A_N \frac{\Big(s\sqrt{r}C + \sqrt{MN} + \sqrt{sM}C\Big)\Big(s\sqrt{r}C + \sqrt{MN} + \sqrt{sN}C\Big)}{s\sqrt{r}C + \sqrt{MN}}. \tag{5.29}$$

Remark 5.10. Once again, T_0 represents the smallest value of T that one could use prior to this work; see [10, Theorem 9]. When $a_m = e(m\alpha)$ and $b_n = e(n\beta)$, Corollary 5.9 saves a factor of $(MN)^{\theta/2}$ over previous bounds (and up to $(MN)^\theta$ if α, β are close to rational numbers with small denominators).

Proof of Corollary 5.9. We only mention what changes from the proof of Corollary 5.5. We expand the sum $\mathcal{S}$ in the left-hand side of (5.28) as a double integral in ζ, ξ, using the Fourier inversion formula

$$\sqrt{xy}\,\Phi\Big(x, y, \frac{\sqrt{xy}}{z}\Big) = \iint_{\mathbb{R}^2} \widehat{\Psi}(\zeta, \xi; z)\, e(x\zeta + y\xi)\, d\zeta\, d\xi,$$

for $\Psi(x, y; z) := \sqrt{xy}\,\Phi(x, y, \sqrt{xy}/z)$, where the Fourier transform is taken in the first two variables. This yields

$$\mathcal{S} \ll_\varepsilon Z^{O(\varepsilon)} C \iint_{\mathbb{R}^2} \frac{|S(\zeta, \xi)|\, d\zeta\, d\xi}{(1+\zeta^4)(1+\xi^4)},$$

where

$$\mathcal{S}(\zeta, \xi) := \sum_{m\sim M} a_m\, e\Big(m\frac{\zeta}{M}\Big) \sum_{n\sim N} b_n\, e\Big(n\frac{\xi}{N}\Big) \sum_{c\in\mathcal{C}_{\mathfrak{a}\mathfrak{b}}} \frac{\mathcal{S}_{\mathfrak{a}\mathfrak{b}}(m, \pm n; c)}{c} \varphi_{\zeta,\xi}\Big(\frac{4\pi\sqrt{mn}}{c}\Big),$$

and $\varphi_{\zeta,\xi}(z)$ is a smooth function supported in $z \asymp X^{-1}$, satisfying $\varphi^{(\ell)}_{\zeta,\xi} \ll_\ell X^\ell$ for

$$X := \frac{C}{\sqrt{MN}}.$$

We proceed as before, applying the Kuznetsov formula from Proposition 3.5 to the inner sum, then using the Bessel transform bounds from Lemma 3.4. When applying Cauchy–Schwarz we keep the variable m inside (as for n), and in consequence we use large sieve inequalities for the sequence (a_m) (i.e., Lemma (3.9) and Assumption 5.4). The resulting bounds are symmetric in M, N, with

$$X_1(\zeta) := \max\Big(1, \frac{q}{M}\Big)\frac{Y_M}{1+|\zeta|^2} \qquad \text{and} \qquad X_2(\xi) := \max\Big(1, \frac{q}{N}\Big)\frac{Y_N}{1+|\xi|^2}.$$

Instead of (5.27), we thus obtain

$$\mathcal{S} \ll_\varepsilon (qMNCZ)^{O(\varepsilon)} \Big(1 + \frac{X}{\sqrt{X_1(0)X_2(0)}}\Big)^{2\theta(q)} \frac{C}{1+X^{-1}} \times \Big(1 + X^{-2} + \frac{M}{q}\Big)^{1/2} \Big(1 + X^{-2} + \frac{N}{q}\Big)^{1/2} \|a_m\|_2\, \|b_n\|_2, \tag{5.30}$$

which recovers (5.28) after plugging in the values of X, X_1, X_2.

Finally, to prove (5.29) for $q = rs$, we pick $\mathfrak{a} = \infty$, and $\mathfrak{b} = 1/s$, keeping the scaling matrices in (3.9), and use (3.12) to rewrite $S(m\bar{r}, n; sc)$ as $S_{\infty\, 1/s}(m, \pm n; s\sqrt{r}c)$ when $(c, r) = 1$. After substituting

$C \leftarrow s\sqrt{r}C$, the value of T inside the θ-factor becomes

$$\frac{s\sqrt{r}C}{\max(M,q)^{1/2}\max(N,q)^{1/2}\sqrt{Y_M Y_N}} \leq \frac{s\sqrt{r}C}{rs\sqrt{Y_M Y_N}} = \frac{C}{\sqrt{rY_M Y_N}},$$

and so (5.28) recovers (5.29) up to minor rearrangements. □

Corollary 5.11 (Kloosterman bounds with averaging over q, m, n, c). *Let $Q, M, N \geq 1/2$, $C, Z \gg 1$, $Y_N > 0$, $\varepsilon > 0$, and $\omega \in \mathbb{R}/\mathbb{Z}$. For each $q \sim Q$, let $(q, N, Z, (a_{n,q})_{n\sim N}, A_{N,q}, Y_N)$ satisfy Assumption 5.4, $w_q \in \mathbb{C}$, $\mathfrak{b}_q$ be a cusp of $\Gamma_0(q)$, and $\Phi_q : (0,\infty)^3 \to \mathbb{C}$ be a smooth function, with $\Phi_q(x,y,z)$ supported in $x, y, z \asymp 1$, and $\partial_x^j \partial_y^k \partial_z^\ell \Phi_q(x,y,z) \ll_{j,k,\ell,\varepsilon} Z^{(j+k)\varepsilon}$ for $j, k, \ell \geq 0$. Then with the choice of scaling matrices in (3.9) and a consistent choice of the $\pm$ sign, one has*

$$\sum_{q\sim Q} w_q \sum_{m\sim M} e(m\omega) \sum_{n\sim N} a_{n,q} \sum_{c\in\mathcal{C}_{\infty\mathfrak{b}_q}} \Phi_q\left(\frac{m}{M}, \frac{n}{N}, \frac{c}{C}\right) S_{\infty\mathfrak{b}_q}(m, \pm n; c) \ll_\varepsilon (QMNCZ)^{O(\varepsilon)}$$
$$\times (1+T)^{2\theta_{\max}} \frac{\sqrt{QM}\|w_q A_{N,q}\|_2 C^2}{C+\sqrt{MN}} \left(1 + \frac{MN}{C^2} + \frac{M}{Q}\right)^{1/2} \left(1 + \frac{MN}{C^2} + \frac{N}{Q}\right)^{1/2}, \tag{5.31}$$

for

$$T = \frac{T_0}{\sqrt{Y_N}}, \qquad T_0 := \frac{C}{\max(M,Q)\max(N,Q)^{1/2}} \leq \frac{C}{Q^{3/2}}.$$

In particular, let $R, S \geq 1/2$; for every $r \sim R$, $s \sim S$ with $(r,s) = 1$, let $w_{r,s} \in \mathbb{C}$, $\Phi_{r,s}$ be as above, and $(rs, N, Z, (a_{n,r,s})_{n\sim N}, A_{N,r,s}, Y_N)$ satisfy Assumption 5.4. Then one has

$$\sum_{\substack{r\sim R\\ s\sim S\\ (r,s)=1}} w_{r,s} \sum_{m\sim M} e(m\omega) \sum_{n\sim N} a_{n,r,s} \sum_{(c,r)=1} \Phi_{r,s}\left(\frac{m}{M}, \frac{n}{N}, \frac{c}{C}\right) S(m\overline{r}, \pm n; sc) \ll_\varepsilon (RSMNCZ)^{O(\varepsilon)}$$
$$\times \left(1 + \frac{C}{R\sqrt{SY_N}}\right)^{2\theta_{\max}} \sqrt{RSM}\|w_{r,s}A_{N,r,s}\|_2$$
$$\times \frac{\left(S\sqrt{RC} + \sqrt{MN} + \sqrt{SMC}\right)\left(S\sqrt{RC} + \sqrt{MN} + \sqrt{SNC}\right)}{S\sqrt{RC} + \sqrt{MN}}. \tag{5.32}$$

Remark 5.12. The norms $\|w_q A_{N,q}\|_2$ and $\|w_{r,s}A_{N,r,s}\|_2$ refer to sequences indexed by $q \sim Q$, respectively $r \sim R$, $s \sim S$ (but not $n \sim N$). In practice, it is often helpful to follow (5.32) with the bound

$$\frac{\left(S\sqrt{RC} + \sqrt{MN} + \sqrt{SMC}\right)\left(S\sqrt{RC} + \sqrt{MN} + \sqrt{SNC}\right)}{S\sqrt{RC} + \sqrt{MN}}$$
$$\ll S\sqrt{RC} + \sqrt{MN} + \sqrt{SMC} + \sqrt{SNC} + \frac{\sqrt{SMC}\sqrt{SNC}}{S\sqrt{RC}}$$
$$\ll \left(\frac{C^2}{R}(M+RS)(N+RS) + MN\right)^{1/2}. \tag{5.33}$$

Remark 5.13. Corollary 5.11 should be compared with [10, Theorem 11], the relevant saving being $Y_N^{\theta_{\max}}$. One can state a similar result, to be compared with [10, Theorem 10], using a general sequence $(b_m)_{m\sim M}$ instead of $b_m = e(m\omega)$; one would need to replace a factor of $\sqrt{M}$ with $\|b_m\|_2$, and adjust the value of T_0 using [10, Theorem 6] (or rather, its optimization in [32]) instead of [10, Theorem 7].

Proof of Corollary 5.11. We proceed as in the proof of Corollary 5.9, swapping the sum over q with the integral to bound the sum $\mathcal{S}$ in the left-hand side of (5.28) by

$$\mathcal{S} \ll_\varepsilon Z^{O(\varepsilon)} C \iint_{\mathbb{R}^2} \frac{S(\zeta,\xi)\, d\zeta\, d\xi}{(1+\zeta^4)(1+\xi^4)},$$

where

$$\mathcal{S}(\zeta,\xi) := \sum_{q\sim Q} |w_q| \left| \sum_{m\sim M} e\left(m\left(\omega+\frac{\zeta}{M}\right)\right) \sum_{n\sim N} a_{n,q}\, e\left(n\frac{\xi}{N}\right) \sum_{c\in\mathcal{C}_{\infty\mathfrak{b}_q}} \frac{\mathcal{S}_{\infty\mathfrak{b}_q}(m,\pm n;c)}{c} \varphi_{\zeta,\xi,q}\left(\frac{4\pi\sqrt{mn}}{c}\right)\right|,$$

and $\varphi_{\zeta,\xi,q}(z)$ are smooth functions supported in $z \asymp X^{-1}$, satisfying $\varphi^{(\ell)}_{\zeta,\xi,q} \ll_\ell X^\ell$ for $X := \frac{C}{\sqrt{MN}}$. After applying the Kuznetsov formula, we bound the contribution of the regular spectrum to $\mathcal{S}(\zeta,\xi)$ pointwise in q, as in the previous proofs (leading only to an extra factor of $\|w_q A_{N,q}\|_1 \le \|w_q A_{N,q}\|_2\sqrt{Q}$ instead of A_N). As in (5.24), the contribution of the exceptional spectrum is

$$\mathcal{S}_{\mathcal{M},\mathrm{exc}}(\zeta,\xi) \ll \frac{1}{1+X^{-1}} \sum_{q\sim Q} |w_q| \sum_{\lambda_j<1/4} \frac{1+X^{2\theta_j(q)}}{\cosh(\pi\kappa_j)} \left|\sum_{m\sim M} e\left(m\left(\omega+\frac{\zeta}{M}\right)\right)\overline{\rho_{j\infty_q}(m)}\right| \left|\sum_{n\sim N} a_{n,q}\, e\left(n\frac{\xi}{N}\right)\rho_{j\mathfrak{b}_q}(n)\right|.$$

We then apply Cauchy–Schwarz in the double sum over q and j, splitting $X = X_0\sqrt{X_1X_2}$ for $X_2(\xi)$ as in (5.25); but this time we choose

$$X_1 := \max\left(M, \frac{Q^2}{M}\right), \tag{5.34}$$

corresponding to the allowable range in Theorem 3.11. Keeping $|w_q|$ only in the second sum, this yields

$$\mathcal{S}_{\mathcal{M},\mathrm{exc}}(\zeta,\xi) \ll \sqrt{\frac{(1+X_0)^{2\theta_{\max}}}{1+X^{-1}} \mathcal{S}_M(\zeta,\xi)\, \mathcal{S}_N(\zeta,\xi)},$$

where

$$\mathcal{S}_M(\zeta,\xi) := \sum_{q\sim Q} \sum_{\lambda_j<1/4} \frac{X_1^{2\theta_j(q)}}{\cosh(\pi\kappa_j)} \left|\sum_{m\sim M} e\left(m\left(\omega+\frac{\zeta}{M}\right)\right)\overline{\rho_{j\infty_q}(m)}\right|^2,$$

$$\mathcal{S}_N(\zeta,\xi) := \sum_{q\sim Q} |w_q|^2 \sum_{\lambda_j<1/4} \frac{X_2^{2\theta_j(q)}}{\cosh(\pi\kappa_j)} \left|\sum_{n\sim N} a_{n,q}\, e\left(n\frac{\xi}{N}\right)\rho_{j\mathfrak{b}_q}(n)\right|^2.$$

The treatment of $\mathcal{S}_N$ remains the same as before, pointwise in q, leading to an extra factor of $\|w_q A_{N,q}\|_2^2$ instead of A_N^2. For $\mathcal{S}_M$, we apply Theorem 3.11 (which allowed the choice of X_1 from (5.34)), leading to an extra factor of $\sqrt{Q}$. Overall, instead of (5.30), we obtain

$$\mathcal{S} \ll_\varepsilon (QMNCZ)^{O(\varepsilon)} \left(1+\frac{X}{\sqrt{X_1X_2(0)}}\right)^{2\theta_{\max}} \frac{C}{1+X^{-1}} \times \left(1+X^{-2}+\frac{M}{Q}\right)^{1/2} \left(1+X^{-2}+\frac{N}{Q}\right)^{1/2} \sqrt{QM}\|w_q A_{N,q}\|_2,$$

and plugging in the values of X, X_1, X_2 yields (5.31).

To prove (5.32), let $Q := RS$. By the divisor bound, the left-hand side is at most

$$x^{o(1)} \sum_{Q<q\le 4Q} \max_{\substack{r\sim R\\ s\sim S\\ (r,s)=1\\ rs=q}} |w_{r,s}| \left| \sum_{m\sim M} e(m\omega) \sum_{n\sim N} a_{n,r,s} \sum_{(c,r)=1} \Phi_{r,s}\left(\frac{m}{M}, \frac{n}{N}, \frac{c}{C}\right) S(m\bar{r}, \pm n; sc)\right|,$$

where we interpret any empty maximum as 0. For each q, let $r = r(q), s = s(q)$ attain the maximum (if there are no such r, s, pick $w_q := 0$ and disregard the rest of this paragraph). Then let $w_q := w_{r,s}$, $a_{n,q} := a_{n,r,s}$, $\Phi_q(x, y, z) := \Phi_{r,s}(x, y, z\,(S/s)\sqrt{R/r})$, and $\mathfrak{b}_q := 1/s$, with the scaling matrix in (3.9).

Due to Lemma 3.2, after the change of variables $c \leftarrow c/(s\sqrt{r})$, this leaves us with the sum

$$x^{o(1)} \sum_{Q<q\le 4Q} |w_q| \left| \sum_{m\sim M} e(m\omega) \sum_{n\sim N} a_{n,q} \sum_{c\in \mathcal{C}_{\infty\mathfrak{b}_q}} \Phi_q\left(\frac{m}{M}, \frac{n}{N}, \frac{c}{S\sqrt{R}C}\right) S_{\infty\mathfrak{b}_q}(m, \pm n; c)\right|.$$

Incorporating 1-bounded coefficients into (w_q) to remove absolute values, the desired bound now follows from (5.31). We note that the T parameter becomes

$$T \ll \frac{S\sqrt{R}C}{Q^{3/2}\sqrt{Y_N}} \asymp \frac{C}{R\sqrt{SY_N}},$$

as in (5.31). □

As a direct consequence of Corollary 5.11 and standard techniques, one can also deduce a result for sums of incomplete Kloosterman sums, improving [10, Theorem 12].

Corollary 5.14 (Incomplete Kloosterman bounds with averaging over r, s, n, c, d). *Let $R, S, N \ge 1/2$, $C, D, Z \gg 1$, $Y_N > 0$, and $\varepsilon > 0$. For each $r \sim R, s \sim S$ with $\gcd(r, s) = 1$, let the tuple $(rs, N, Z, (a_{n,r,s})_{n\sim N}, A_{N,r,s}, Y_N)$ satisfy Assumption 5.4, $w_{r,s} \in \mathbb{C}$, and $\Phi_{r,s} : (0,\infty)^3 \to \mathbb{C}$ be a smooth function, with $\Phi_{r,s}(x, y, z)$ supported in $x, y, z \asymp 1$, and $\partial_x^j \partial_y^k \partial_z^\ell \Phi_q(x, y, z) \ll_{j,k,\ell,\varepsilon} Z^{j\varepsilon}$ for $j, k, \ell \ge 0$. Then with a consistent choice of the $\pm$ sign, one has*

$$\sum_{\substack{r\sim R\\ s\sim S\\ (r,s)=1}} w_{r,s} \sum_{n\sim N} a_{n,r,s} \sum_{\substack{c,d\\ (rd,sc)=1}} \Phi_{r,s}\left(\frac{n}{N}, \frac{d}{D}, \frac{c}{C}\right) e\left(\pm n\frac{\overline{rd}}{sc}\right) \ll_\varepsilon (RSNCDZ)^{O(\varepsilon)} \|w_{r,s}A_{N,r,s}\|_2 \mathscr{F}, \tag{5.35}$$

where

$$\mathscr{F}^2 := D^2NR + \left(1 + \frac{C^2}{R^2SY_N}\right)^{2\theta_{\max}} CS(C + DR)(RS + N).$$

Proof of Corollary 5.14. This follows from Corollary 5.11 (specifically, (5.32)) by completing Kloosterman sums, passing from the d-variable to a variable m of size $\ll_\varepsilon (CDS)^\varepsilon CS/D$; this is completely analogous to how [10, Theorem 12] follows from [10, Theorem 11] in [10, §9.2]. We note that [10, Theorem 12] has a minor error (replacing D^2NR with D^2NRS^{-1}), which has been corrected in [5]. □

6. The greatest prime factor of $n^2 + 1$

Here we use our new inputs from Section 5.3 in the computations of Merikoski [37] and de la Bretèche–Drappeau [8], in order to prove Theorem 1.1. We begin with a brief informal sketch.

6.1. Sketch of the argument

We will ultimately prove a lower bound of the shape

$$\sum_{n\sim x}\sum_{\substack{p \text{ prime}\\ p\mid n^2+1\\ p>x^{1.3}}}\log p > \varepsilon\, x\log x,$$

which implies that for some (in fact, for many) $n\sim x$, we must have $P^+(n^2+1) > x^{1.3}$. As in previous works [37, 8, 9, 23], we use an idea of Chebyshev to estimate the full sum

$$\sum_{n\sim x}\sum_{\substack{p \text{ prime}\\ p\mid n^2+1}}\log p \approx \sum_{n\sim x}\sum_{d\mid n^2+1}\Lambda(d) = \sum_{n\sim x}\log(n^2+1) = 2x\log x + O(x),$$

where Λ is the von Mangoldt function. It then remains to upper bound

$$\sum_{n\sim x}\sum_{\substack{p \text{ prime}\\ p\mid n^2+1\\ p\le x^{1.3}}}\log p = \sum_{\substack{p \text{ prime}\\ p\le x^{1.3}}}\log p\sum_{n\sim x}\mathbb{1}_{n^2\equiv -1 \pmod p} \overset{?}{<} (2-\varepsilon)\, x\log x.$$

Following Merikoski [37], we use repeated applications of Buchstab's identity inside the Harman sieve method, to reduce estimating the above sum over primes to bounding "Type I" and "Type II" sums of the form

$$\sum_{d\le D}\lambda_d\sum_{\substack{q\sim Q\\ q\equiv 0 \pmod d}}\left(\sum_{n\sim x}\mathbb{1}_{n^2\equiv -1 \pmod q} - \frac{x}{q}\sum_{\nu \pmod q}\mathbb{1}_{\nu^2\equiv -1 \pmod q}\right),$$

respectively

$$\sum_{q_1\sim Q_1}\lambda_{q_1}\sum_{q_2\sim Q_2}\mu_{q_2}\left(\sum_{n\sim x}\mathbb{1}_{n^2\equiv -1 \pmod{q_1q_2}} - \frac{x}{q_1q_2}\sum_{\nu \pmod{q_1q_2}}\mathbb{1}_{\nu^2\equiv -1 \pmod{q_1q_2}}\right),$$

for various ranges of D, Q, Q_i with $Q_1Q_2 = Q \le x^{1.3}$, aiming to win over the trivial bound of x.

Let us sketch our improvement of the Type I information. As in [8, 37], we separate n into residue classes $\nu \pmod q$ and apply a truncated version of Poisson summation to the sum over n; the principal frequency $h = 0$ cancels with the subtracted main term, leading to a sum like

$$\sum_{d\le D}\lambda_d\sum_{\substack{q\sim Q\\ q\equiv 0 \pmod d}}\sum_{\substack{\nu \pmod q\\ \nu^2\equiv -1 \ (q)}}\sum_{h\sim \frac{Q}{x}} e\left(\frac{h\nu}{q}\right).$$

We can then parametrize the solutions to $\nu^2\equiv -1 \pmod q$ by the Gauss correspondence (see [9, Lemma 2]), which gives $q = r^2+s^2$ and $\frac{\nu}{q}\approx \frac{\overline{r}}{s} \pmod 1$. In the critical range $r, s\sim\sqrt{Q}$, this leaves us with the sum

$$\sum_{d\le D}\lambda_d\sum_{\substack{r,s\sim\sqrt{Q}\\ r^2\equiv -s^2 \pmod d}}\sum_{h\sim \frac{Q}{x}} e\left(\frac{h\overline{r}}{s}\right).$$

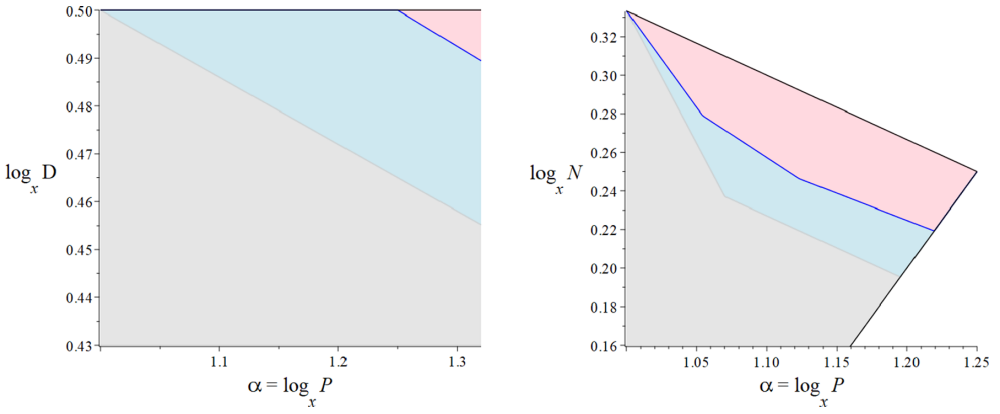

Figure 2. *Type I (left) and Type II (right) ranges. Previous results in gray; our improvements in blue; conditional ranges in red (assuming Selberg's eigenvalue conjecture).*

Separating $r\overline{s}$ into residue classes $\varrho \bmod d$ and Fourier-completing the sum over r yields a dual variable k of size D, and inner sums of the shape

$$\sum_{k\sim D}\sum_{h\sim \frac{Q}{x}} e\left(\frac{k\varrho}{d}\right)\sum_{s\sim\sqrt{Q}} S(h,k\overline{d};s),$$

where ϱ is a solution to $\varrho^2 \equiv -1 \pmod d$; compare these to (1.2). Since the sequence $(e(\frac{k\varrho}{d}))_{k\sim D}$ depends on the level d and its length is as large as the level, the large sieve inequality for general sequences from Theorem 1.2 cannot produce any savings in the X^θ-aspect for this sequence. However, our large sieve inequality for exponential phases from Theorem 1.5 will save a factor of $D^{\theta/2}$, which follows through to the sieve computations. This improves the results of de la Bretèche–Drappeau [8, §8], based in turn on Duke–Friedlander–Iwaniec [16], and is nearly enough to remove the dependency on Selberg's eigenvalue conjecture in the relevant Type I ranges, as illustrated in Fig. 2 (left).

Following Merikoski [37], the Type II sums can be treated similarly using an additional Cauchy–Schwarz step. Once again, this leads to trilinear forms of Kloosterman sums as in (1.2), where both sequences (a_m) and (b_n) depend on the level. Using our large sieve inequalities, we can leverage the fact that (a_m) happen to be exponential phases as in Theorem 1.5, while (b_n) have the shape in Theorem 1.7 (if the sum over h is kept inside the Cauchy–Schwarz step). This ultimately produces three admissible Type II ranges, all gathered in Proposition 6.4 and illustrated in Fig. 2 (right).

By carefully plugging in these Type I and II estimates into Merikoski's Harman sieve computations, which require the numerical calculation of multidimensional integrals, we deduce Theorem 1.1.

6.2. Arithmetic information

We aim to improve the dependency on the θ parameter in the arithmetic information from [37, Propositions 1 and 2]; to do so, we first improve a lemma of de la Bretèche–Drappeau [8, Lemme 8.3].

Lemma 6.1 (De la Bretèche–Drappeau-style exponential sums). *Let $\varepsilon > 0$, $M \gg 1$, and $\theta := \frac{7}{64}$.*

(i). *Let $q, h \in \mathbb{Z}$ and $1 \le |h| \ll q$. Given a smooth function $f : (0,\infty) \to \mathbb{C}$ supported in $v \asymp 1$, with $f^{(j)} \ll_j 1$ for $j \ge 0$, one has*

$$\sum_{(m,q)=1} f\left(\frac{m}{M}\right)\sum_{\nu^2\equiv -1 \ (mod\ mq)} e\left(\frac{h\nu}{mq}\right) \ll_\varepsilon (qhM)^\varepsilon\left(|h| + \sqrt{qM}\left(1+(q,h)^\theta q^{-3\theta/2}M^\theta\right)\right). \quad (6.1)$$

(ii). *Let $Q \geq 1/2$, $1/2 \leq H \ll QM$, and $t \in \mathbb{R}/\mathbb{Z}$. Given smooth functions $(f_q(v))_{q\sim Q}$ supported in $v \asymp 1$, with $f_q^{(j)} \ll_j 1$ for $j \geq 0$, one has*

$$\frac{1}{Q}\sum_{q\sim Q}\left|\frac{1}{H}\sum_{h\sim H} e(th)\sum_{(m,q)=1} f_q\left(\frac{m}{M}\right)\sum_{\nu^2\equiv -1 \ (mod\ mq)} e\left(\frac{h\nu}{mq}\right)\right| \ll_\varepsilon (QHM)^\varepsilon\left(H+\sqrt{M}\left(1+H^{-2\theta}Q^{\theta/2}M^\theta\right)+\sqrt{\frac{QM}{H}}\left(1+Q^{-3\theta/2}M^\theta\right)\right). \tag{6.2}$$

Proof. This is a refinement of the first and third bounds in [8, Lemme 8.3], winning factors of about $q^{\theta/2}$ via our Corollaries 5.5 and 5.11. We only mention what changes from the proof in [8, §8.1], working in the particular case $d = r = 1$, $D = -1$. We note that for $D = -1$, the relevant cusps $\mathfrak{a}$ from [8, §8.1] are equivalent to $0/1$, and thus have $\mu(\mathfrak{a}) = q^{-1}$ (which is also why Merikoski's bounds in [37, §3.8] only require such cusps too).

For part (i), we consider the sums of Kloosterman sums from [8, (8.30)], given (with notation to be explained below) by

$$V_N = V_N(q,h) := \sum_{N/2\leq|n|\leq 2N}\sum_{\gamma\in\mathcal{C}_{\infty\mathfrak{a}}} S_{\infty\mathfrak{a}}(h,n;\gamma)\, G_N(\gamma,n).$$

Here, the n-variable came from a completion of Kloosterman sums, and was localized to a dyadic range of size $N \ll q^{1+\eta}M^\eta$ (where $\eta > 0$ is a small parameter), while $G_N(\gamma,n)$ is a smooth function normalized such that

$$\Phi(x,y) := q\, G_N\left(yq\sqrt{M}, xN\right)$$

satisfies the assumptions of Corollary 5.5 with $Z = qM$ and $\varepsilon \asymp \eta$. Also, $\mathfrak{a}$ is a cusp of $\Gamma_0(q)$, and the scaling matrix $\sigma_{\mathfrak{a}}$ used implicitly in the Kloosterman sum $S_{\infty\mathfrak{a}}(h,n;\gamma)$ hides an exponential phase of the form $e(n\alpha_q)$; the value of α_q is arbitrary for our purposes.

We can now apply Corollary 5.5 (equivalently, we can bound $\mathscr{M}_N^{\mathrm{exc}}$ in [8, (8.40)] using Theorem 1.5), using $a_n = e(n\alpha_q)$, $Y_N = A_N = \sqrt{N}$ (corresponding to (5.18)), $C = q\sqrt{M}$, and $m = |h|$. This yields

$$V_N \ll_\eta (qhM)^{O(\eta)}\left(1+\frac{q\sqrt{M}}{q^{3/2}(q,h)^{-1/2}N^{1/4}}\right)^{2\theta}\sqrt{NM},$$

where we used that $q^\eta C = q^{1+\eta}\sqrt{M} \gg \sqrt{hN}$, that $\sqrt{(q,h)|h|} \leq |h| \leq q$, and that $N \ll q^{1+\eta}M^\eta$ (in particular, the 1-term is dominant in the last two parentheses from (1.1), up to factors of $(qhM)^{o(1)}$).

This bound is increasing in N, so using $N \ll q^{1+\eta}M^\eta$ once again, we get

$$V_N \ll_\eta (qhM)^{O(\eta)}\sqrt{qM}\left(1+(q,h)^\theta q^{-3\theta/2}M^\theta\right),$$

which gives the second term claimed in the upper bound from (6.1).

Part (ii) follows similarly using Corollary 5.11 (or equivalently, by bounding $\mathscr{M}_N^{\mathrm{exc}}$ in [8, §8.1.12] using Theorem 1.5 once again). Indeed, with the similar choices $a_{n,q} = e(n\alpha_q)$, $Y_N = A_{N,q} = \sqrt{N}$, $Z = QM$, and $C = Q\sqrt{M}$, our bound (5.31) yields

$$\frac{1}{Q}\sum_{q\sim Q}\frac{1}{H}\left|\sum_{h\sim H} e(th)\, V_N(q,h)\right| \ll_\eta (QHM)^{O(\eta)}\left(1+\frac{Q\sqrt{M}}{\max(Q,H)\,Q^{1/2}N^{1/4}}\right)^{2\theta}\sqrt{\frac{NM}{H}}\left(1+\frac{H}{Q}\right)^{1/2}.$$

Again, this bound is increasing in N, so plugging in $N \ll Q^{1+\eta}M^{\eta}$ gives a right-hand side of

$$\begin{aligned}&\ll_{\eta} (QHM)^{O(\eta)}\sqrt{\frac{M}{H}}\max(Q,H)^{1/2}\left(1+\max(Q,H)^{-2\theta}Q^{\theta/2}M^{\theta}\right)\\&\ll (QHM)^{O(\eta)}\sqrt{\frac{M}{H}}\left(H^{1/2}+Q^{1/2}+\left(H^{(1/2)-2\theta}+Q^{(1/2)-2\theta}\right)Q^{\theta/2}M^{\theta}\right),\end{aligned}$$

which gives all but the first term in the upper bound from (6.2). As in [8], the first terms of $|h|$ and H from our bounds in (6.1) and (6.2) could be improved via partial summation, but we omit this optimization too since it will not be relevant for our computations. □

Notation 6.2 (Set-up for arithmetic information). Let $x \geq 1$, $\alpha \in [1, 3/2)$, and

$$P := x^{\alpha}.$$

Let Φ, Ψ be smooth functions supported in $[1, 4]$, satisfying $\Phi \geq 0$ and $\Phi^{(j)}, \Psi^{(j)} \ll_j 1$ for $j \geq 0$ (in [37, §2.1], Merikoski uses $b(t) = \Phi(t/x)$ and $\Psi(t) \leftarrow \Psi(t/P)$). For $q \in \mathbb{Z}_+$, define

$$|\mathcal{A}_q| := \sum_{n^2 \equiv -1 \pmod q} \Phi\left(\frac{n}{x}\right), \qquad X := \int \Phi\left(\frac{t}{x}\right) dt = x\int \Phi,$$

$$\rho(q) := \#\left\{\nu \in \mathbb{Z}/q\mathbb{Z} : \nu^2 \equiv -1 \pmod q\right\}.$$

We will estimate the difference

$$|\mathcal{A}_q| - X\frac{\rho(q)}{q}$$

in "Type I" and "Type II" sums with $q \asymp P$. The Type I sums average over moduli in arithmetic progressions, say $q \equiv 0 \pmod d$ and $d \leq D$, with arbitrary divisor-bounded coefficients λ_d; the Type II sums average over moduli with a conveniently-sized factor, say $q = mn$ with $n \sim N$ (and $m \asymp P/N$), with divisor-bounded coefficients a_m, b_n. One can also view the Type I sums as special Type II sums where $a_m = 1$, except that Type II estimates typically require a lower bound on N.

The strength of the resulting Type I and Type II information is given by the ranges of parameters D and N (in terms of x and P) for which we can obtain power-savings over the trivial bound – that is, for which the sums over $|\mathcal{A}_q|$ have an asymptotic formula. Fig. 2 illustrates the (previous unconditional, new unconditional, and conditional) admissible choices of $\log_x D$ and $\log_x N$ in terms of $\alpha = \log_x P$; both graphs continue downwards, the second region being lower-bounded by the function $\alpha - 1$. The previous unconditional and the conditional ranges are due to Merikoski [37] and de la Bretèche–Drappeau [8]; our improvements are Propositions 6.3 and 6.4.

Proposition 6.3 (Type I estimate). *For any sufficiently small $\varepsilon > 0$ there exists $\delta > 0$ such that the following holds. With Notation 6.2, $1 \leq \alpha \leq 1.4$, $\theta := \frac{7}{64}$, and $D \geq 1$, one has*

$$\sum_{d \leq D}\lambda_d \sum_{q \equiv 0 \pmod d}\left(|\mathcal{A}_q| - X\frac{\rho(q)}{q}\right)\Psi\left(\frac{q}{P}\right)\log q \ll_{\varepsilon} x^{1-\delta}, \tag{6.3}$$

for any divisor-bounded coefficients (λ_d), provided that

$$D \ll_{\varepsilon} x^{-\varepsilon}\min\left(x^{1/2}, x^{(1-2\theta\alpha)/(2-5\theta)}\right).$$

Proof. This is a refinement of Merikoski's [37, Prop. 1] (which explicitated the computations in de la Bretèche–Drappeau's [8, §8.4]), using our Lemma 6.1.(*ii*) instead of [8, (8.7)]. Indeed, in the first display

on [8, p. 1620], by applying (6.2) for $H \leftarrow PX^{-1+\delta}$ (for $\delta = \delta(\varepsilon)$ to be chosen shortly), $Q \leftarrow D \le x^{1/2}$ and $M \leftarrow P/D$, we instead obtain the bound

$$\begin{aligned} &R_H(x,P,D) \\ &\ll_\delta x^{1+O(\delta)}P^{-1}DH\left(H+\sqrt{\frac{P}{D}}\left(1+H^{-2\theta}D^{\theta/2}\left(\frac{P}{D}\right)^{\theta}\right)+\sqrt{\frac{P}{H}}\left(1+D^{-3\theta/2}\left(\frac{P}{D}\right)^{\theta}\right)\right) \\ &= x^{O(\delta)}\left(\frac{PD}{x}+\sqrt{PD}\left(1+x^{2\theta}P^{-\theta}D^{-\theta/2}\right)+\sqrt{x}D\left(1+D^{-5\theta/2}P^{\theta}\right)\right). \end{aligned}$$

Here, $R_H(x,P,D)$ resulted from our Type I sum after putting d in dyadic ranges, expanding and Fourier-completing $|\mathcal{A}_q|$; see [8, §8.4] and then [9, §4, 5]. Overall, this bound is acceptable in (6.3) (i.e., $\ll_\varepsilon x^{1-\delta}$) provided that for an absolute constant K, one has

$$\begin{aligned} D &\ll_\varepsilon x^{-K\delta}\min\left(x^2P^{-1}, x^{1/2}, x^{2(1-2\theta)/(1-\theta)}P^{-(1-2\theta)/(1-\theta)}, x^{1/(2-5\theta)}P^{-2\theta/(2-5\theta)}\right) \\ &= x^{-\varepsilon}\min\left(x^{2-\alpha}, x^{1/2}, x^{(2-\alpha)(1-2\theta)/(1-\theta)}, x^{(1-2\theta\alpha)/(2-5\theta)}\right), \end{aligned}$$

where we picked $\delta := \varepsilon/K$ and substituted $P = x^\alpha$. A quick numerical verification shows that for $1 \le \alpha \le 1.4$ and $\theta = \frac{7}{64}$, the first and the third term do not contribute to the minimum. □

Proposition 6.4 (Type II estimate). *For any sufficiently small $\varepsilon > 0$ there exists $\delta > 0$ such that the following holds. With Notation 6.2, $\theta := \frac{7}{64}$, and $MN = P$ with $M, N \ge 1$, one has*

$$\sum_{\substack{m\sim M\\ n\sim N}} a_m b_n \left(|\mathcal{A}_{mn}| - X\frac{\rho(mn)}{mn}\right)\Psi\left(\frac{mn}{P}\right)\log(mn) \ll_\varepsilon x^{1-\delta}, \tag{6.4}$$

for any divisor-bounded coefficients (a_m) and (b_n), provided that one of the following holds:

(i). *(b_n) is supported on square-free integers, and*

$$x^{\alpha-1+\varepsilon} \ll_\varepsilon N \ll_\varepsilon x^{-\varepsilon}\max\left(x^{(2-(1+2\theta)\alpha)/(3-4\theta)}, x^{(2-\alpha)(1-2\theta)/(3-2\theta)}\right); \tag{6.5}$$

(ii). *(b_n) is supported on primes, and*

$$x^{\alpha-1+\varepsilon} \ll_\varepsilon N \ll_\varepsilon x^{(4-3\alpha)/3-\varepsilon}. \tag{6.6}$$

Remark 6.5. The upper range in Proposition 6.4.(*ii*), which completely removes the dependency on Selberg's eigenvalue conjecture, wins over that in Proposition 6.3.(*i*) only for $\alpha < 136/129 \approx 1.054$. As in [37], assuming Selberg's eigenvalue conjecture, the full admissible range in part (*i*) is $N \ll_\varepsilon x^{(2-\alpha)/3}$, which includes the range in part (*ii*).

Proof of Proposition 6.4.(ii), assuming (i). This is a refinement of Merikoski's [37, Prop. 4.(ii)], using our Lemma 6.1.(*i*) instead of de la Bretèche–Drappeau's bound [8, (8.5)].

We briefly recall that in [37, §3], Merikoski expanded and Fourier-completed $|\mathcal{A}_{mn}|$ (resulting in a sum over $1 \le |h| \le H := Px^{-1+\delta}$), removed the smooth cross-conditions in h, m, n, and inserted the condition $(m,n) = 1$ to reach Type II sums $\Sigma(M,N)$. Then they applied Cauchy–Schwarz with the sum over n inside, to obtain $\Sigma(M,N) \ll M^{1/2}\,\Xi(M,N)^{1/2}$, and trivially bounded the 'diagonal' contribution of $n_1 = n_2$ using the condition $N \gg_\varepsilon x^{2(\alpha-1)+\varepsilon}$. To estimate the remaining sum $\Xi_0(M,N)$ from the second-to-last display in [37, §3.10], we apply our bound (6.1) with $q \leftarrow n_1n_2$ and $h \leftarrow h(n_1 - n_2)$;

this gives the refined bound

$$\begin{aligned}\Xi_0(M,N) &\ll_\delta x^{O(\delta)} \sum_{\substack{n_1,n_2\sim N\\(n_1,n_2)=1}} \frac{1}{H} \sum_{1\le|h|\le H} \left(HN+\sqrt{MN^2}\left(1+(n_1n_2,h(n_1-n_2))^\theta N^{-3\theta}M^\theta\right)\right)\\ &\ll_\delta x^{O(\delta)}N^2\left(HN+M^{1/2}N+M^{(1+2\theta)/2}N^{1-3\theta}\right).\end{aligned}$$

This results in a contribution to $\Sigma(M,N)$ of

$$\begin{aligned}&\ll_\delta x^{O(\delta)}M^{1/2}N\left(H^{1/2}N^{1/2}+M^{1/4}N^{1/2}+M^{(1+2\theta)/4}N^{(1-3\theta)/2}\right)\\ &\ll x^{O(\delta)}\left(x^{-1/2}PN+P^{3/4}N^{3/4}+P^{(3+2\theta)/4}N^{(3-8\theta)/4}\right),\end{aligned}$$

which is acceptable (i.e., $\ll_\varepsilon x^{1-\delta}$) provided that for a large enough absolute constant K,

$$N\ll_\varepsilon x^{-K\delta}\min\left(x^{3/2}P^{-1},x^{4/3}P^{-1},x^{4/(3-8\theta)}P^{-(3+2\theta)/(3-8\theta)}\right).$$

Trivially removing the first term, picking $\delta:=\varepsilon/K$, and substituting $P=x^\alpha$, this proves (6.4) in the range

$$x^{2(\alpha-1)+\varepsilon}\ll_\varepsilon N\ll_\varepsilon x^{-\varepsilon}\min\left(x^{(4-3\alpha)/3},x^{(4-(3+2\theta)\alpha)/(3-8\theta)}\right),$$

when (b_n) is supported on primes. The remaining ranges to consider are

$$x^{\alpha-1+\varepsilon}\ll_\varepsilon N\ll_\varepsilon \min\left(x^{2(\alpha-1)+\varepsilon},x^{(4-3\alpha)/3-\varepsilon}\right) \tag{6.7}$$

and

$$\min\left(x^{\alpha-1+\varepsilon},x^{(4-(3+2\theta)\alpha)/(3-8\theta)-\varepsilon}\right)\ll_\varepsilon N\ll_\varepsilon x^{(4-3\alpha)/3-\varepsilon}, \tag{6.8}$$

both of which are (barely) covered by Proposition 6.4.(i). Indeed, for (6.7), a quick numerical verification shows that

$$\min\left(2(\alpha-1),\frac{4-3\alpha}{3}\right)<\frac{2-(1+2\theta)\alpha}{3-4\theta}$$

for $\theta=\frac{7}{64}$ and all α, the smallest gap being ≈ 0.07, at $a=10/9$. In (6.8), we have a nontrivial range only when

$$\frac{4-(3+2\theta)\alpha}{3-8\theta}\le\frac{4-3\alpha}{3}\qquad\Longleftrightarrow\qquad\alpha\ge\frac{16}{15}\ge 1.066,$$

and for such α, we have

$$\frac{4-3\alpha}{3}<\frac{2-(1+2\theta)\alpha}{3-4\theta}.$$

Therefore, (6.4) holds in the full range from (6.6). □

Remark 6.6. As in [37, §3.10], the bound for $\Xi_0(M,N)$ in the proof above does not leverage any cancellation over h. One can attempt to do this using Corollary 5.9 with $a_m=e(m\alpha_q)$ and b_n as in (1.10), but the gain in the H-aspect would be smaller than the loss in the θ-aspect in our computations. This is because Proposition 6.4.(ii) is only relevant for α close to 1, that is, for small values of H.

Proof of Proposition 6.4.(i). This is a refinement of Merikoski's [37, Prop. 4. (i)], using Corollary 5.9 (plus Theorem 1.7) instead of Deshouillers–Iwaniec's bound [10, Theorem 9].

We very briefly recall the relevant parts of Merikoski's argument and the sizes of the parameters therein, pointing the reader to [37, §3] for details. In [37, §3.4], one expanded and Fourier-completed $|\mathcal{A}_{mn}|$, resulting in a sum over $1 \leq |h| \leq H$ with

$$H := Px^{-1+\delta}, \tag{6.9}$$

as before. Then, one removed the smooth cross-conditions in h, m, n, and separated $k = (m, n)$ to reach the type-II sums $\Sigma_k(M, N)$ from the first display on [37, p. 1275]; we need to bound these by $\ll_\varepsilon x^{1-\delta}/k$, for $\delta = \delta(\varepsilon)$ to be chosen.

In [37, §3.5], one applied Cauchy–Schwarz keeping the sums over h, n inside, to obtain

$$\Sigma_k(M, N) \ll \left(\frac{M}{k}\right)^{1/2} \Xi_k(M, N)^{1/2}, \tag{6.10}$$

and trivially bounded the contribution of $h_2 n_1 = h_1 n_2$ to Ξ_k, using the condition $N \gg_\varepsilon x^{\alpha-1+\varepsilon}$; then they separated $n_0 = (n_1, n_2)$ (and let $n_i \leftarrow n_i/n_0$). We note that considering nontrivial values of the GCD-parameters k and n_0 was not necessary in the proof of Proposition 6.4.(ii), since then (b_n) was supported on primes; in a first pass the reader can pretend that $k = n_0 = 1$.

In [37, §3.6], one expanded the condition $(m, n_0 n_1 n_2) = 1$ by Möbius inversion, resulting in a sum over $d \mid n_0 n_1 n_2$ (we switched notation from δ to d). Then, one applied Gauss' lemma ([37, Lemma 9]), resulting in sums $\Psi_k(R, S)$ of incomplete Kloosterman sums, ranging over r, s of sizes

$$1 \ll R, S \ll \sqrt{\frac{PN}{kn_0}}. \tag{6.11}$$

In [37, §3.7], one completed Kloosterman sums, resulting in a sum over $|t| \leq T$ with

$$T = x^\delta \frac{SdN^2}{Rn_0}, \tag{6.12}$$

and trivially bounded the contribution of $t = 0$. This ultimately leads to the sums of Kloosterman sums $\widetilde{\Psi}_k(R, S)$ from [37, p. 1279], which have a relevant *level* of

$$\varrho := dk^2 n_0 n_1 n_2 \asymp \frac{dN^2}{n_0}. \tag{6.13}$$

Finally, in [37, §3.8], Merikoski used [9, Theorem 9] to bound the trilinear sums of Kloosterman sums

$$\mathcal{K} = \mathcal{K}(d, n_0, n_1, n_2) := \max_{\alpha \pmod{\varrho}} \left| \sum_{m \sim \mathcal{M}} a_m \sum_{n \sim \mathcal{N}} b_n \sum_{(c, \varrho)=1} \Phi\left(\frac{m}{\mathcal{M}}, \frac{n}{\mathcal{N}}, \frac{c}{\mathcal{C}}\right) S\left(m\overline{\varrho}, \pm n; c\right) \right|,$$

where Φ is a smooth function as in Corollary 5.9 with $Z = 1$, (c_h) are bounded coefficients,

$$a_m := e\left(-m\frac{\alpha}{\varrho}\right), \qquad b_n := \sum_{\substack{h_1 \sim H_1 \\ h_2 \sim H_2 \\ n = h_1 n_2 - h_2 n_1}} c_{h_1} \overline{c_{h_2}}, \tag{6.14}$$

both of which depend on the level ϱ,

$$\mathcal{M} \ll T, \qquad \mathcal{N} \ll \frac{HN}{kn_0}, \qquad \mathcal{C} \ll S, \tag{6.15}$$

and $1/2 \le H_2 \le H_1 \le H$. We will achieve better bounds for $\mathcal{K}$ by leveraging the structure of the coefficients (a_m) and (b_n). To do so, we note that the coefficients c_h (obtained by removing the cross-condition in h, m, n on [37, p. 1274]) are smooth functions of h. In fact, expanding $|\mathcal{A}_{mn}| - \frac{\rho(mn)}{mn}$ via Lemma 3.1 and fixing j, u up to a logarithmic loss, we can use the coefficients

$$c_h := \Psi_j\left(\frac{|h|}{H_j}\right) e\left(-h\frac{ux}{P}\right),$$

from (3.4), where $1 \le 2^j = H_j \le H = Px^{-1+\delta}$, $u \asymp 1$, and $\Psi_j : (\frac{1}{2}, 2) \to \mathbb{C}$ are compactly-supported smooth functions with bounded derivatives. In particular, through Lemma 3.1 we put $|h|$ in (smooth) dyadic ranges, and then separate into positive and negative values of h, all before applying Cauchy–Schwarz; so the resulting variables h_1, h_2 are of the same size. The coefficients (b_n) from (6.14) become

$$b_n := \sum_{\substack{h_1, h_2 \in \mathbb{Z} \\ n = h_1 n_2 - h_2 n_1}} c_{h_1} \overline{c_{h_2}},$$

which are in a suitable form to use Theorem 1.7 (see also (5.19)), with $a = 1$, $H = H_j$, $\alpha_i = \pm ux/P \ll x^\delta H^{-1}$, and

$$L := \frac{N}{kn_0} \asymp n_1 \asymp n_2. \tag{6.16}$$

In particular, since $\varrho \ge n_1 n_2 \asymp L^2$, the tuple $(\varrho, \mathcal{N}, x, (b_n)_{n\sim\mathcal{N}}, A_\mathcal{N}, Y_\mathcal{N})$ satisfies Assumption 5.4 with

$$Y_\mathcal{N} := \max\left(1, \frac{\mathcal{N}H_j}{(H_j + L)Lx^\delta}\right) \qquad \text{and} \qquad A_\mathcal{N} := \|b_n \mathbb{1}_{n\sim\mathcal{N}}\|_2 + \sqrt{\mathcal{N}}\sqrt{\frac{H_j}{L} + \frac{H_j^2}{L^2}}, \tag{6.17}$$

where we used that $T_{\mathcal{N}/L}(\alpha_i) \ll T_H(\alpha_i) \ll 1 + H|\alpha_i| \ll x^\delta$. On the other hand, by Theorem 1.5 (see also (5.18)), the tuple $(\varrho, \mathcal{M}, x, (a_m)_{m\sim\mathcal{M}}, A_\mathcal{M}, Y_\mathcal{M})$ satisfies Assumption 5.4, with

$$Y_\mathcal{M} := \sqrt{\mathcal{M}} \qquad \text{and} \qquad A_\mathcal{M} := \sqrt{\mathcal{M}}. \tag{6.18}$$

By Corollary 5.9, specifically (5.29), it follows that

$$\mathcal{K} \ll_\delta x^{O(\delta)}\left(1 + \frac{\mathcal{C}}{\sqrt{\varrho}Y_\mathcal{M}Y_\mathcal{N}}\right)^{2\theta} A_\mathcal{M}A_\mathcal{N}\frac{\left(\sqrt{\varrho}\mathcal{C} + \sqrt{\mathcal{M}\mathcal{N}} + \sqrt{\mathcal{M}}\mathcal{C}\right)\left(\sqrt{\varrho}\mathcal{C} + \sqrt{\mathcal{M}\mathcal{N}} + \sqrt{\mathcal{N}}\mathcal{C}\right)}{\sqrt{\varrho}\mathcal{C} + \sqrt{\mathcal{M}\mathcal{N}}},$$

and substituting (6.17) and (6.18) gives

$$\begin{aligned}\mathcal{K} \ll_\delta x^{O(\delta)}&\left(1 + \frac{\mathcal{C}}{\sqrt{\varrho}\mathcal{M}^{1/4}\max\left(1, \sqrt{\frac{\mathcal{N}H_j}{(H_j+L)L}}\right)}\right)^{2\theta}\sqrt{\mathcal{M}}\left(\|b_n\mathbb{1}_{n\sim\mathcal{N}}\|_2 + \sqrt{\mathcal{N}}\sqrt{\frac{H_j}{L} + \frac{H_j^2}{L^2}}\right)\\ &\times\frac{\left(\sqrt{\varrho}\mathcal{C} + \sqrt{\mathcal{M}\mathcal{N}} + \sqrt{\mathcal{M}}\mathcal{C}\right)\left(\sqrt{\varrho}\mathcal{C} + \sqrt{\mathcal{M}\mathcal{N}} + \sqrt{\mathcal{N}}\mathcal{C}\right)}{\sqrt{\varrho}\mathcal{C} + \sqrt{\mathcal{M}\mathcal{N}}}.\end{aligned} \tag{6.19}$$

Since $\mathcal{N} \ll \varrho$ (which follows from $H \ll N$), the term on the second line of (6.19) is at most $\ll \sqrt{\rho}\mathcal{C} + \sqrt{\mathcal{M}}\mathcal{C} + \sqrt{\mathcal{M}\mathcal{N}}$, as in [37, p. 1280]. The resulting bound is nondecreasing in $\mathcal{M}, \mathcal{C}$, so we can

plug in their upper bounds from (6.15), as well as (6.12) and (6.13) to obtain

$$\sum_{d|n_0n_1n_2}\frac{1}{TH^2}\mathcal{K}(d,n_0,n_1,n_2)\ll_\delta x^{O(\delta)}\max_{d\geq 1}\frac{Rn_0}{SdN^2H^2}\left(1+\frac{S\min\left(1,\sqrt{\frac{(H_j+L)L}{\mathcal{N}H_j}}\right)}{\sqrt{\frac{dN^2}{n_0}}\left(\frac{SdN^2}{Rn_0}\right)^{1/4}}\right)^{2\theta}$$
$$\times\sqrt{\frac{SdN^2}{Rn_0}}\left(\|b_n\mathbb{1}_{n\sim\mathcal{N}}\|_2+\sqrt{\mathcal{N}}\sqrt{\frac{H_j}{L}+\frac{H_j^2}{L^2}}\right)$$
$$\times\left(\sqrt{\frac{dN^2}{n_0}}S+\sqrt{\frac{SdN^2}{Rn_0}}S+\sqrt{\frac{SdN^2}{Rn_0}}\mathcal{N}\right),$$

where none of the remaining variables have implicit dependencies on d. The right-hand side is seen to be nonincreasing in d, so we can plug in $d=1$ for an upper bound. Moreover, when summing over $n_1,n_2\sim L=N/(kn_0)$, we have the same bound as in [37, bottom of p. 1280] (by [37, Lemma 7]) for the contribution of $A_{\mathcal{N}}$:

$$\sum_{n_1,n_2\sim L}\left(\|b_n\mathbb{1}_{n\sim\mathcal{N}}\|_2+\sqrt{\mathcal{N}}\sqrt{\frac{H_j}{L}+\frac{H_j^2}{L^2}}\right)\ll\sqrt{\mathcal{N}}\max\left(H_jL,H_j^{1/2}L^{3/2}\right).$$

The resulting bound for $\sum_{n_1,n_2\sim L}\sum_{d|n_0n_1n_2}\frac{1}{TH^2}\mathcal{K}(d,n_0,n_1,n_2)$ is nondecreasing in $\mathcal{N},H_j$, so we can plug in the upper bounds in $\mathcal{N}\ll HL$ and $H_j\ll H$ and simplify the resulting expression to obtain

$$\sum_{n_1,n_2\sim L}\sum_{d|n_0n_1n_2}\frac{1}{TH^2}\mathcal{K}(d,n_0,n_1,n_2)\ll_\delta x^{O(\delta)}\left(1+\frac{S^{3/4}R^{1/4}n_0^{3/4}}{N^{3/2}}\min\left(1,\frac{\sqrt{H+L}}{H}\right)\right)^{2\theta}$$
$$\times\max\left(H^{1/2}L^{3/2},L^2\right)\left(\frac{\sqrt{RS}}{H}+\frac{S}{H}+\sqrt{\frac{L}{H}}\right).$$

Summing over n_0 and plugging in the bounds for R,S,L from (6.11) and (6.16), we get

$$\Upsilon_k:=\sum_{n_0\ll N}\rho(n_0)\sum_{n_1,n_2\sim L}\sum_{d|n_0n_1n_2}\frac{1}{TH^2}\mathcal{K}(d,n_0,n_1,n_2)$$
$$\ll_\delta x^{O(\delta)}\sum_{n_0\ll N}\left(1+\frac{\sqrt{PN}n_0^{1/4}}{\sqrt{k}N^{3/2}}\min\left(1,\frac{\sqrt{H+\frac{N}{kn_0}}}{H}\right)\right)^{2\theta}$$
$$\times\max\left(H^{1/2}\left(\frac{N}{kn_0}\right)^{3/2},\left(\frac{N}{kn_0}\right)^2\right)\left(\frac{\sqrt{PN}}{\sqrt{kn_0}H}+\sqrt{\frac{N}{kn_0H}}\right).$$

Using that $H \ll N$, this further yields

$$\begin{aligned}\Upsilon_k &\ll_\delta \frac{x^{O(\delta)}}{k} \sum_{n_0 \ll N} \frac{1}{n_0^{2-(\theta/2)}} \left(1 + \frac{\sqrt{PN}}{N^{3/2}} \min\left(1, \frac{\sqrt{N}}{H}\right)\right)^{2\theta} N^2 \left(\frac{\sqrt{PN}}{H} + \sqrt{\frac{N}{H}}\right) \\ &\ll_\delta \frac{x^{O(\delta)}}{k} \left(1 + \min\left(\frac{\sqrt{P}}{N}, \frac{\sqrt{P}}{\sqrt{N}H}\right)\right)^{2\theta} \left(\frac{\sqrt{P}N^{5/2}}{H} + \frac{N^{5/2}}{\sqrt{H}}\right).\end{aligned}$$

Since we have $N \le \sqrt{x} \le \sqrt{P}$ and $\sqrt{N}H \le x^{1/4}Px^{-1+\delta} \le x^{1/4+3/4-1+\delta}\sqrt{P}$ for the ranges in Proposition 6.4.(i), we may ignore the 1-term in the θ-factor; plugging in (6.9), we obtain

$$\Upsilon_k \ll_\delta \frac{x^{1+O(\delta)}N^{5/2}}{k\sqrt{P}} \min\left(\frac{\sqrt{P}}{N}, \frac{x}{\sqrt{NP}}\right)^{2\theta},$$

which improves [37, (3.7)]. In light of (6.10) and $MN = P$, this gives a contribution to $\Sigma_k(M, N)$ of

$$\ll_\delta \frac{x^{1/2+O(\delta)}}{k} P^{1/4}N^{3/4} \min\left(\frac{\sqrt{P}}{N}, \frac{x}{\sqrt{NP}}\right)^{\theta},$$

which is acceptable (i.e., $\ll_\varepsilon x^{1-\delta}/k$) provided that for a large enough absolute constant K,

$$N \ll_\varepsilon x^{-K\delta} \max\left(x^{2/(3-4\theta)}P^{-(1+2\theta)/(3-4\theta)}, x^{2(1-2\theta)/(3-2\theta)}P^{-(1-2\theta)/(3-2\theta)}\right).$$

Choosing $\delta := \varepsilon/K$ and substituting $P = x^\alpha$ completes our proof. □

6.3. Sieve computations

To complete the proof of Theorem 1.1, it remains to adapt the calculation in [37, §2] with our Type I and Type II information.

Notation 6.7 (Set-up for sieve computations). Further to Notation 6.2, we follow [37, p. 1257] and let $P_x := P^+\left(\prod_{x\le n\le 2x}(n^2+1)\right)$, then use a smooth dyadic partition of unity to split

$$S(x) := \sum_{\substack{x<p\le P_x \\ p \text{ prime}}} |\mathcal{A}_p| \log p$$

into a sum over $x \le P \le P_x$, $P = P_j = 2^j x$ of

$$S(x, P) := \sum_{p \text{ prime}} \Psi_j\left(\frac{p}{P}\right) |\mathcal{A}_p| \log p,$$

up to an error of $O(x)$. Here Ψ_j are smooth functions supported on $[1, 4]$, with $\Psi_j^{(k)} \ll_k 1$ for all $k \ge 0$. Following [37, p. 1259], given $z \ge 1$ and $u \in \mathbb{Z}_+$, we also let $P(z) := \prod_{\text{prime } p<z} p$ and

$$S(\mathcal{A}(P)_u, z) := \sum_{(n,P(z))=1} |\mathcal{A}_{un}| \Psi\left(\frac{un}{P}\right) \log(un),$$

so that $S(x,P) = S(\mathcal{A}(P), 2\sqrt{P})$ (where dropping the u index means that $u = 1$). This has a corresponding main term of

$$S(\mathcal{B}(P)_u, z) := X \sum_{(n, P(z))=1} \frac{\rho(un)}{un} \Psi\left(\frac{un}{P}\right) \log(un),$$

sums of which can be computed via [37, Lemma 1]. Finally, the linear sieve upper bound will require the solutions $F(s), f(s)$ to the delay-differential equation system from [37, p. 1263], while the Harman sieve computations will require the Buchstab function $\omega(u)$, bounded as in [37, (2.5)].

As in [37, p. 1257], we aim to find the greatest $\overline{\omega}$ for which

$$\sum_{\substack{x \le P \le x^{\overline{\omega}} \\ P = 2^j x}} S(x,P) \le (1-\varepsilon) X \log x. \tag{6.20}$$

Since $S(x) = X \log x + O(x)$ (see [37, (2.1)]), this will imply the lower bound $P_x \ge x^{\overline{\omega}}$.

Lemma 6.8 (Linear sieve upper bound). *For any $\varepsilon > 0$ there exists $\delta > 0$ such that the following holds. With $\theta := \frac{7}{64}$, Notation 6.2, Notation 6.7, and $D := x^{-\varepsilon} \min\left(x^{1/2}, x^{(1-2\theta\alpha)/(2-5\theta)}\right)$, one has*

$$S(\mathcal{A}(P), z) \le (1+\delta) X \int \Psi\left(\frac{u}{P}\right) \frac{\alpha \log x}{e^{\gamma} \log z} F\left(\frac{\log D}{\log z}\right) \frac{du}{u},$$

for any $x^{\varepsilon} < z < D$, where γ is the Euler–Mascheroni constant.

Proof. This is just [37, Lemma 2] with the updated parameter D from our Type I information (Proposition 6.3). □

Proposition 6.9 (Asymptotics for Harman sieve sums). *For any $\varepsilon > 0$ there exists $\delta > 0$ such that the following hold. With $\theta := \frac{7}{64}$, Notation 6.2 and Notation 6.7, let*

$$D := x^{-\varepsilon} \min\left(x^{1/2}, x^{(1-2\theta\alpha)/(2-5\theta)}\right), \qquad U := D x^{1-\alpha-\varepsilon} =: x^{\xi},$$

and (λ_u) be divisor-bounded coefficients. Also, let

$$\sigma_0 := \max\left(\frac{2-(1+2\theta)\alpha}{3-4\theta}, \frac{(1-2\theta)(2-\alpha)}{3-2\theta}\right) \tag{6.21}$$

be the exponent from Proposition 6.4.(i).

(i). *For $1 \le \alpha < 228/203 - O(\varepsilon)$ and*

$$\sigma := \max\left(\frac{4-3\alpha}{3}, \sigma_0\right) - \varepsilon, \tag{6.22}$$

one has

$$\sum_{u \le U} \lambda_u \left(S\left(\mathcal{A}(P)_u, x^{\sigma}\right) - S\left(\mathcal{B}(P)_u, x^{\sigma}\right)\right) \ll_{\varepsilon} x^{1-\delta}.$$

(ii). *For $1 \le \alpha < 139/114 - O(\varepsilon)$ and*

$$\gamma := \sigma_0 - (\alpha - 1) - 2\varepsilon, \tag{6.23}$$

one has

$$\sum_{u\leq U}\lambda_u\left(S\left(\mathcal{A}(P)_u,x^\gamma\right)-S\left(\mathcal{B}(P)_u,x^\gamma\right)\right)\ll_\varepsilon x^{1-\delta}.$$

Proof. These are just [37, Propositions 3 and 4], adapted with our Type II information from Proposition 6.4; the additional term of $(4-3\alpha)/3$ from (6.22) comes from Proposition 6.4.(ii). We note that the proof of [37, Proposition 3] requires

$$2(\alpha-1)<\sigma_0-O(\varepsilon)=\max\left(\frac{2-(1+2\theta)\alpha}{3-4\theta},\frac{(1-2\theta)(2-\alpha)}{3-2\theta}\right)-O(\varepsilon),$$

which happens for $\alpha<228/203-O(\varepsilon)$. Similarly, the proof of [37, Proposition 4] requires

$$\alpha-1<\sigma_0-O(\varepsilon)=\max\left(\frac{2-(1+2\theta)\alpha}{3-4\theta},\frac{(1-2\theta)(2-\alpha)}{3-2\theta}\right)-O(\varepsilon),$$

which happens for $\alpha<139/114-O(\varepsilon)$. □

We are now ready to prove our Theorem 1.1, in a very similar manner to [37, §2.6].

Proof of Theorem 1.1. We follow the Harman sieve computations in [37, §2.4], applying Buchstab's identity in the same ways (with adapted ranges corresponding to the values of $D,U,\sigma_0,\sigma,\gamma,\xi$ from Lemma 6.8 and Proposition 6.9). The five ranges relevant in the proof are now $\alpha<25/24$, $25/24\leq\alpha<228/203$, $228/203\leq\alpha<7/6$, $7/6\leq\alpha<139/114$, and $\alpha\geq139/114$. Here, the values $228/203$ and $139/114$ come from Proposition 6.9, while $25/24$ and $7/6$ are the thresholds deciding the inequalities $\alpha<\xi+2\sigma$, respectively $2(\alpha-1)<\xi$, up to $o(1)$ factors. Indeed, we recall that

$$\xi=\min\left(\frac{1}{2},\frac{(1-2\theta\alpha)}{2-5\theta}\right)-(\alpha-1)-2\varepsilon,$$

and only the first term in the minimum is relevant for the aforementioned inequalities. We thus obtain

$$\sum_{\substack{x\leq P\leq x^{139/114}\\ P=2^jx}}S(x,P)\leq\left(\frac{7}{6}-1+G_1+G_2+G_3+G_4+G_5-G_6+o(1)\right)X\log x,$$

where

$$\begin{aligned}
G_1&:=\int_1^{25/24}\alpha\left(\int_\sigma^{\alpha-2\sigma}\omega\left(\frac{\alpha}{\beta}-1\right)\frac{d\beta}{\beta^2}+\int_\xi^{\alpha/2}\omega\left(\frac{\alpha}{\beta}-1\right)\frac{d\beta}{\beta^2}\right)d\alpha<0.02093,\\
G_2&:=\int_{25/24}^{228/203}\alpha\int_\sigma^{\alpha/2}\omega\left(\frac{\alpha}{\beta}-1\right)\frac{d\beta}{\beta^2}d\alpha<0.10528,\\
G_3&:=\int_{228/203}^{7/6}\alpha\int_{\sigma_0}^{\alpha/2}\omega\left(\frac{\alpha}{\beta}-1\right)\frac{d\beta}{\beta^2}d\alpha<0.07319,\\
G_4&:=\int f_4\left(\alpha,\vec{\beta}\right)\alpha\,\omega\left(\frac{\alpha-\beta_1-\beta_2-\beta_3}{\beta_3}\right)\frac{d\beta_1\,d\beta_2\,d\beta_3}{\beta_1\beta_2\beta_3^2}d\alpha<0.00163,\\
G_5&:=4\int_{7/6}^{139/114}\alpha\,d\alpha<0.25116,\\
G_6&:=\int_{7/6}^{139/114}\alpha\int_{\alpha-1}^{\sigma_0}\omega\left(\frac{\alpha}{\beta}-1\right)\frac{d\beta}{\beta^2}d\alpha>0.02789.
\end{aligned}$$

Here, f_4 denotes the characteristic function of the set

$$\left\{\frac{228}{203} < \alpha < \frac{7}{6}, \gamma < \beta_3 < \beta_2 < \beta_1 < \alpha - 1, \right.$$
$$\left. \beta_1 + \beta_2, \beta_1 + \beta_3, \beta_2 + \beta_3, \beta_1 + \beta_2 + \beta_3 \notin [\alpha - 1, \sigma_0]\right\}.$$

We computed the integrals G_i (for $i \neq 5$) by directly adapting the ranges in Merikoski's Python 3.7 code files (see [37, p. 1268]). In the expression for G_5, we implicitly used the value $D = x^{1/2-\varepsilon}$ since $\frac{1}{2} < \frac{(1-2\theta\alpha)}{2-5\theta}$ for $\alpha \leq 139/114$, and the fact that $1 < 1/(2(\alpha-1)) \leq 3$ for $7/6 \leq \alpha \leq 139/114$. Thus

$$\sum_{\substack{x \leq P \leq x^{139/114} \\ P = 2^j x}} S(x, P) < 0.59097\, X \log x.$$

For the remaining range $\alpha \geq 139/114$, we apply Lemma 6.8 to obtain (as in [37, (2.8)])

$$\sum_{\substack{x^{139/114} \leq P \leq x^{\overline{\omega}} \\ P = 2^j x}} S(x, P) \leq \left(4 \int_{139/114}^{1.25} \alpha\, d\alpha + (4 - 10\theta) \int_{1.25}^{\overline{\omega}} \frac{\alpha}{1 - 2\theta\alpha} d\alpha\right) X \log x,$$

where $\alpha = 1.25 = 5/4$ is the threshold at which the expression for D changes (i.e., when $\frac{1}{2} = \frac{(1-2\theta\alpha)}{2-5\theta}$). We conclude that (6.20) holds (for small enough ε) provided that

$$(4 - 10\theta) \int_{1.25}^{\overline{\omega}} \frac{\alpha}{1 - 2\theta\alpha} d\alpha < 1 - 0.59097 - 4 \int_{139/114}^{1.25} \alpha\, d\alpha,$$

where the right-hand side is at least 0.257406. This inequality (barely) holds true when $\overline{\omega} = 1.30008$, which proves Theorem 1.1. □

Acknowledgments. The author is grateful to his PhD advisor, James Maynard, for his kind guidance, to Sary Drappeau for many thoughtful discussions, and to Jori Merikoski, Lasse Grimmelt, Jared Duker Lichtman, and the referees, for helpful comments and suggestions.

Competing interest. The author has no competing interests to declare.

Financial support. For the duration of this project, the author was sponsored by EPSRC Scholarship 2580868 at University of Oxford.

References

[1] E. Assing, V. Blomer and J. Li, 'Uniform Titchmarsh divisor problems', *Adv. Math.* **393** (2021), Paper No. 108076, 51. ISSN 0001-8708, 1090–2082. https://doi.org/10.1016/j.aim.2021.108076.
[2] E. Bombieri, J. B. Friedlander and H. Iwaniec, 'Primes in arithmetic progressions to large moduli', *Acta Math.* **156**(3–4) (1986), 203–251. ISSN 0001-5962,1871-2509. https://doi.org/10.1007/BF02399204.
[3] E. Bombieri, J. B. Friedlander and H. Iwaniec, 'Primes in arithmetic progressions to large moduli. II', *Math. Ann.* **277**(3) (1987), 361–393. ISSN 0025-5831,1432-1807. https://doi.org/10.1007/BF01458321.
[4] E. Bombieri, J. B. Friedlander and H. Iwaniec, 'Primes in arithmetic progressions to large moduli. III', *J. Amer. Math. Soc.* **2**(2) (1989), 215–224. ISSN 0894-0347,1088-6834. https://doi.org/10.2307/1990976.
[5] E. Bombieri, J. B. Friedlander and H. Iwaniec, 'Some corrections to an old paper', Preprint, 2019, arXiv:1903.01371.
[6] D. Bump, *Automorphic Forms and Representations*, volume 55 of *Cambridge Studies in Advanced Mathematics* (Cambridge University Press, Cambridge, 1997). ISBN 0-521-55098-X. https://doi.org/10.1017/CBO9780511609572.
[7] J. Cilleruelo and M. Z. Garaev, 'Concentration of points on two and three dimensional modular hyperbolas and applications', *Geom. Funct. Anal.* **21**(4) (2011), 892–904. ISSN 1016-443X,1420-8970. https://doi.org/10.1007/s00039-011-0127-6.
[8] R. de La Bretèche and S. Drappeau, 'Niveau de répartition des polynômes quadratiques et crible majorant pour les entiers friables', *J. Eur. Math. Soc.* **22**(5) (2020), 1577–1624. ISSN 1435-9855,1435-9863. https://doi.org/10.4171/jems/951.

[9] J.-M. Deshouillers and H. Iwaniec, 'On the greatest prime factor of $n^2 + 1$', *Ann. Inst. Fourier (Grenoble)* **32**(4) (1982), 1–11. ISSN 0373-0956,1777-5310. https://doi.org/10.5802/aif.891.

[10] J.-M. Deshouillers and H. Iwaniec, 'Kloosterman sums and Fourier coefficients of cusp forms', *Invent. Math.* **70**(2) (1982), 219–288. ISSN 0020-9910,1432-1297. https://doi.org/10.1007/BF01390728.

[11] J.-M. Deshouillers and H. Iwaniec, 'Power mean values of the Riemann zeta function', *Mathematika* **29**(2) (1982), 202–212. ISSN 0025-5793. https://doi.org/10.1112/S0025579300012298.

[12] J.-M. Deshouillers and H. Iwaniec, 'Power mean-values for Dirichlet's polynomials and the Riemann zeta-function. II', *Acta Arith.* **43**(3) (1984), 305–312. ISSN 0065-1036. https://doi.org/10.4064/aa-43-3-305-312.

[13] S. Drappeau, 'Théorèmes de type Fouvry-Iwaniec pour les entiers friables', *Compos. Math.* **151**(5) (2015), 828–862. ISSN 0010-437X,1570-5846. https://doi.org/10.1112/S0010437X14007933.

[14] S. Drappeau, 'Sums of Kloosterman sums in arithmetic progressions, and the error term in the dispersion method', *Proc. Lond. Math. Soc. (3)* **114**(4) (2017), 684–732. ISSN 0024-6115,1460-244X. https://doi.org/10.1112/plms.12022.

[15] S. Drappeau, K. Pratt and M. Radziwiłł, 'One-level density estimates for Dirichlet L-functions with extended support', *Algebra Number Theory* **17**(4) (2023), 805–830. ISSN 1937-0652,1944-7833. https://doi.org/10.2140/ant.2023.17.805.

[16] W. Duke, J. B. Friedlander and H. Iwaniec, 'Equidistribution of roots of a quadratic congruence to prime moduli', *Ann. of Math. (2)* **141**(2) (1995), 423–441. ISSN 0003-486X,1939-8980. https://doi.org/10.2307/2118527.

[17] E. Fouvry, E. Kowalski and P. Michel, 'On the exponent of distribution of the ternary divisor function', *Mathematika* **61**(1) (2015), 121–144. ISSN 0025-5793,2041-7942. https://doi.org/10.1112/S0025579314000096.

[18] J. B. Friedlander and H. Iwaniec, 'Incomplete Kloosterman sums and a divisor problem', *Ann. of Math. (2)* **121**(2) (1985), 319–350. ISSN 0003-486X,1939-8980. https://doi.org/10.2307/1971175. With an appendix by Bryan J. Birch and Enrico Bombieri.

[19] D. Goldfeld and J. Hundle,. *Automorphic Representations and L-functions for the General Linear Group. Volume I*, volume 129 of *Cambridge Studies in Advanced Mathematics* (Cambridge University Press, Cambridge, 2011). ISBN 978-0-521-47423-8. https://doi.org/10.1017/CBO9780511973628. With exercises and a preface by Xander Faber.

[20] D. Goldfeld and J. Hundley, *Automorphic Representations and L -functions for the General Linear Group. Volume II*, volume 130 of *Cambridge Studies in Advanced Mathematics* (Cambridge University Press, Cambridge, 2011). ISBN 978-1-107-00799-4. https://doi.org/10.1017/CBO9780511973628. With exercises and a preface by Xander Faber.

[21] G. Harman, 'On values of $n^2 + 1$ free of large prime factors', *Arch. Math. (Basel)* **90**(3) (2008), 239–245. ISSN 0003-889X,1420-8938. https://doi.org/10.1007/s00013-007-2404-z.

[22] G. Harman, 'Two problems on the greatest prime factor of $n^2 + 1$', *Acta Arith.* **213**(3) (2024), 273–287. ISSN 0065-1036,1730-6264. https://doi.org/10.4064/aa230710-18-12.

[23] C. Hooley, 'On the greatest prime factor of a quadratic polynomial', *Acta Math.* **117** (1967), 281–299. ISSN 0001-5962,1871-2509. https://doi.org/10.1007/BF02395047.

[24] H. Iwaniec, *Topics in Classical Automorphic Forms*, volume 17 of *Graduate Studies in Mathematics* (American Mathematical Society, Providence, RI, 1997). ISBN 0-8218-0777-3. https://doi.org/10.1090/gsm/017.

[25] H. Iwaniec, *Spectral Methods of Automorphic Forms*, volume 53 of *Graduate Studies in Mathematics* (American Mathematical Society, Providence, RI; Revista Matemática Iberoamericana, Madrid, second edition, 2002). ISBN 0-8218-3160-7. https://doi.org/10.1090/gsm/053.

[26] H. Iwaniec and E. Kowalski, *Analytic Number Theory*, volume 53 (American Mathematical Society, Providence, RI, 2021).

[27] B. Kerr, I. E. Shparlinski, X. Wu and P. Xi, 'Bounds on bilinear forms with Kloosterman sums', *J. Lond. Math. Soc. (2)* **108**(2) (2023), 578–621. ISSN 0024-6107,1469-7750. https://doi.org/10.1112/jlms.12753.

[28] H. H. Kim, 'Functoriality for the exterior square of $_4$ and the symmetric fourth of $_2$', *J. Amer. Math. Soc.* **16**(1) (2003), 139–183. ISSN 0894-0347,1088-6834. https://doi.org/10.1090/S0894-0347-02-00410-1. With appendix 1 by Dinakar Ramakrishnan and appendix 2 by Kim and Peter Sarnak.

[29] E. Kowalski, P. Michel and W. Sawin, 'Bilinear forms with Kloosterman sums and applications', *Ann. of Math. (2)* **186**(2) (2017), 413–500. ISSN 0003-486X,1939-8980. https://doi.org/10.4007/annals.2017.186.2.2.

[30] E. Kowalski, P. Michel and W. Sawin, 'Stratification and averaging for exponential sums: bilinear forms with generalized Kloosterman sums', *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **21** (2020), 1453–1530. ISSN 0391-173X,2036-2145. https://doi.org/10.2422/2036-2145.201805_002.

[31] N. V. Kuznetsov, 'The Petersson conjecture for cusp forms of weight zero and the Linnik conjecture. Sums of Kloosterman sums', *Mat. Sb. (N.S.)* **111**(153)(3) (1980), 334–383, 479. ISSN 0368-8666.

[32] J. D. Lichtman, 'Primes in arithmetic progressions to large moduli, and Goldbach beyond the square-root barrier', Preprint, 2023, arXiv:2309.08522.

[33] Y. V. Linnik, *The Dispersion Method in Binary Additive Problems* (American Mathematical Society, Providence, RI, 1963). Translated by S. Schuur.

[34] J. Maynard, 'Primes in arithmetic progressions to large moduli I: Fixed residue classes', *Mem. Amer. Math. Soc.* **306**(1542) (2025), 1–132. ISSN 0065-9266,1947-6221. https://doi.org/10.1090/memo/1542.

[35] J. Maynard, 'Primes in arithmetic progressions to large moduli II: Well-factorable estimates', *Mem. Amer. Math. Soc.* **306**(1543) (2025), 1–33. ISSN 0065-9266,1947-6221. https://doi.org/10.1090/memo/1543.

[36] J. Maynard, 'Primes in arithmetic progressions to large moduli III: Uniform residue classes', *Mem. Amer. Math. Soc.* **306**(1544) (2025), 1–98. ISSN 0065-9266,1947-6221. https://doi.org/10.1090/memo/1544.

[37] J. Merikoski, 'On the largest prime factor of $n^2 + 1$', *J. Eur. Math. Soc. (JEMS)* **25**(4) (2023), 1253–1284. ISSN 1435-9855,1435-9863. https://doi.org/10.4171/jems/1216.
[38] A. Pascadi, 'On the exponents of distribution of primes and smooth numbers', Preprint, 2025, arXiv:2505.00653.
[39] A. Pascadi, 'Smooth numbers in arithmetic progressions to large moduli', *Compos. Math.* **161**(8) (2025), 1923–1974. ISSN 0010-437X,1570-5846. https://doi.org/10.1112/S0010437X2500747X.
[40] P. Sarnak, 'Selberg's eigenvalue conjecture', *Notices Amer. Math. Soc.* **42**(11) (1995), 1272–1277. ISSN 0002-9920,1088-9477.
[41] A. Selberg, 'On the estimation of Fourier coefficients of modular forms', in *Proc. Sympos. Pure Math ., Vol.* VIII (American Mathematical Society, Providence, RI, 1965), 1–15.
[42] I. E. Shparlinski, 'Character sums with smooth numbers', *Arch. Math. (Basel)* **110**(5) (2018), 467–476. ISSN 0003-889X,1420-8938. https://doi.org/10.1007/s00013-018-1168-y.
[43] I. E. Shparlinski and T. Zhang, 'Cancellations amongst Kloosterman sums', *Acta Arith.* **176**(3) (2016), 201–210. ISSN 0065-1036,1730-6264. https://doi.org/10.4064/aa8365-6-2016.
[44] B. Topacogullari, 'On a certain additive divisor problem', *Acta Arith.* **181**(2) (2017), 143–172. ISSN 0065-1036,1730-6264. https://doi.org/10.4064/aa8643-5-2017.
[45] B. Topacogullari, 'The shifted convolution of generalized divisor functions', *Int. Math. Res. Not.* **2018**(24) (2018), 7681–7724. ISSN 1073-7928,1687-0247. https://doi.org/10.1093/imrn/rnx111.
[46] N. Watt, 'Kloosterman sums and a mean value for Dirichlet polynomials', *J. Number Theory* **53**(1) (1995), 179–210. ISSN 0022-314X,1096-1658. https://doi.org/10.1006/jnth.1995.1086.
[47] P. Xi, 'Ternary divisor functions in arithmetic progressions to smooth moduli', *Mathematika* **64**(3) (2018), 701–729. ISSN 0025-5793,2041-7942. https://doi.org/10.1112/s0025579318000220.
[48] M. P. Young, 'The fourth moment of Dirichlet L-functions', *Ann. of Math. (2)* **173**(1) (2011), 1–50. ISSN 0003-486X,1939-8980. https://doi.org/10.4007/annals.2011.173.1.1.