Topics in analytic number theory and automorphic forms

Alexandru Pascadi

Corpus Christi College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

2025

Părinților mei,
pentru tot.

Acknowledgements

First of all, I would like to thank my advisor, Professor James Maynard, for four years of friendly guidance, enlightening discussions, and wholehearted support during my DPhil studies in Oxford. Thank you for showing me new ways to understand and enjoy Number Theory, and for making everything feel so simple and beautiful – both in Mathematics and in life.

I would also like to thank Professor Terence Tao for his kind and inspiring mentorship throughout my undergraduate and Master's studies at UCLA, and for setting me on a truly exciting path in mathematical research. In addition, I feel very lucky for the guidance of Professors Ciprian Manolescu, Sorin Popa, and Ken Ono during my college years, and for studying with several amazing teachers during my middle & high school years in Bucharest, including Raluca Mangra and Ovidiu Sontea.

I am deeply grateful to my family – especially my parents Manuela and Mihai and my uncle Mihai – for their unwavering love and support, their invaluable time and countless sacrifices, and for encouraging me to pursue Mathematics in a creative, curious, and pressure-free way. I lovingly thank my partner Alexa, for many of my most beautiful memories in and outside Oxford, and for uplifting me during my more difficult days – thank you for being here for both. I am also really thankful for my friends, particularly my former flatmate Andi, for making my time in Oxford much more fun and exciting. To my sister, Sara, and my cousins, Ruxi and Vlad: I wish you the best of luck as you begin new chapters in your lives, and I trust that you will find as much joy and fulfillment in your work as I have.

I also thank Professor Ben Green, Jori Merikoski, and Lasse Grimmelt, for assisting with my Transfer and Confirmation of Status in Oxford, as well as my co-authors Jesse Thorner and Jared Duker Lichtman, for fruitful collaborations. I am very grateful to Sary Drappeau for productive discussions and for his hospitality during my visit in Marseille, and to Professors Valentin Blomer, Farrell Brumley, Peter Sarnak, Peter Humphries, Régis de la Bretèche, and Kannan Soundararajan for many helpful comments.

Finally, I thank EPSRC for their sponsorship of my DPhil studies, and UCLA for the generous Mathematics Undergraduate Merit Scholarship.

Abstract

This thesis concerns improved results in the analytic theory of automorphic forms, as well as their applications to classical problems about the primes and related arithmetic objects.

First, we prove new large sieve inequalities for the Fourier coefficients of exceptional Maass forms of a given level, weighted by sequences with sparse Fourier transforms. 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 combine these results with other ideas to show that both primes and smooth numbers are equidistributed in arithmetic progressions to moduli up to $x^{5/8-o(1)}$, using triply-well-factorable weights for the primes. This completely eliminates the dependency on Selberg's eigenvalue conjecture in previous works of Lichtman and the author, which built in turn on works of Maynard and Drappeau. As applications, we prove refined upper bounds for the counts of twin primes and consecutive smooth numbers.

Next, we obtain density theorems for 'exceptional' cuspidal automorphic representations of GL_n, which fail the generalized Ramanujan conjecture at some place. We depart from approaches based on Kuznetsov-type trace formulae, and instead rely on Rankin–Selberg L-functions. This improves previous density results near the threshold of the pointwise bounds.

Building on these ideas, we develop a new approach to large sieve inequalities for families of automorphic L-functions $L(s)$, improving earlier results and simultaneously handling the Dirichlet coefficients of L, L^{-1}, and $\log L$. Our bounds are sharp in ranges that are complementary to large sieve inequalities based on trace formulae. We apply our results to establish zero density estimates for families of automorphic L-functions.

Contents

Chapter 1

Introduction

Analytic number theory studies quantitative properties of the integers, prime numbers, and related arithmetic objects, using a variety of methods ranging from complex, harmonic, and functional analysis to probability and discrete mathematics. Besides the rich diversity of mathematical techniques involved, analytic number theory stands out through its great number of unsolved problems, many of which have easy-to-grasp statements. The methods developed to attack such problems often find applications to other branches of mathematics, as well as cryptography and mathematical physics.

In this thesis, we study a few problems in analytic number theory that use or concern the theory of 'automorphic forms'. The latter are central objects of interest in number theory, which obey certain symmetries and share additive and multiplicative structure; they often arise in classical problems as tools to estimate special exponential sums. Unfortunately, many properties of these automorphic forms remain conjectural, and a recurring theme in this thesis is showing that there are few 'exceptional' forms that fail these conjectures, which leads to unconditional results.

What follows in this chapter is an informal discussion, meant for a reader with a general mathematical background and some basic knowledge of classical analytic number theory, but no required knowledge of automorphic forms. We will aim to intuitively answer the questions below:

- What are some key open problems and techniques in analytic number theory?
- What are automorphic forms, and how do they show up in these problems?
- What tools can we apply to estimate sums related to automorphic forms?

While we emphasize intuition and general principles here, Chapter 2 will contain formal statements and preliminary results, to be cited in the later chapters.

1.1 Some key problems in analytic number theory

From finding patterns in the primes to finding primes in sparse sets, analytic number theory has no shortage of difficult problems, which quickly begin to require deep tools. Consider, for a start, the following major open questions:

$$\textit{Is } n^2+1 \textit{ infinitely often a prime, or close to a prime, for } n \in \mathbb{Z}? \tag{1.1}$$

$$\textit{How many pairs of twin primes } \{p, p+2\} \textit{ are there up to } x? \tag{1.2}$$

Although a full answer to such questions is generally viewed as far beyond our current technology, there is considerable partial progress. Indeed, for (1.1), while we cannot show that n^2+1 (or any other non-linear irreducible polynomial) takes infinitely many prime values, one can show that it must have a large prime factor infinitely often; we will give such a result in Chapter 3. For (1.2), while we cannot prove any nontrivial lower-bound on the number of twin primes up to x, there are upper bounds of the correct order of magnitude predicted by the Hardy–Littlewood conjectures, up to a small constant; the currently-best constant is given in a corollary from Chapter 4.

1.1.1 Distribution in arithmetic progressions

A key idea in sieve theory is to detect arithmetic structure, such as primality or 'almost-primality' inside a set of integers A, using information about the equidistribution of the elements of A in many arithmetic progressions.

Indeed, by sifting $A = \{n^2+1 : n \le x\}$ or $A = \{p-2 : p \le x \text{ prime}\}$, our best progress toward questions (1.1) and (1.2) relies on estimating sums of the shape

$$\sum_{q \le Q} \lambda_q \sum_{n \le x} \mathbb{1}_{n^2+1 \equiv 0 \pmod q}, \qquad \sum_{q \le Q} \lambda_q \sum_{\substack{p \le x \\ \text{prime}}} \mathbb{1}_{p \equiv 2 \pmod q}, \tag{1.3}$$

where (λ_q) are certain 1-bounded coefficients, and $\mathbb{1}_S$ denotes the truth value (0 or 1) of a statement S. The goal is to win over the trivial bound of $\approx x$ after subtracting a suitable main term, with a *level of distribution* Q that is as large as possible.

The celebrated *Bombieri–Vinogradov* theorem [Bom65; Vin65] achieves this for the second sum in (1.3), when (λ_q) are arbitrary 1-bounded coefficients and Q is just below $\sqrt{x}$; this matches the conditional range implied by the Generalized Riemann Hypothesis for individual moduli q. Surprisingly, it is possible to go beyond this range by averaging with special weights (λ_q). Virtually all such results use a combinatorial decomposition of the indicator function of the primes into various multiplicative convolutions of 1-bounded sequences (α_m), (β_n), which leads to estimating sums like

$$\sum_{q \le Q} \lambda_q \sum_{mn \le x} \alpha_m \beta_n \mathbb{1}_{mn \equiv 2 \pmod q}. \tag{1.4}$$

A key feature of the congruences from (1.3) and (1.4),

$$n^2 + 1 \equiv 0 \pmod{q}, \qquad\qquad mn \equiv 2 \pmod{q},$$

is that they can be interpreted as finding solutions in $r \in \mathbb{Z}$ to the determinant equations

$$\det \begin{pmatrix} n & r \\ q & n \end{pmatrix} = -1, \qquad\qquad \det \begin{pmatrix} m & r \\ q & n \end{pmatrix} = 2.$$

In particular, if $\lambda_q = \alpha_m = \beta_n = 1$ in (1.3) and (1.4), then we are left with a count of integer matrices with a given determinant, which has an approximate invariance under multiplication by matrices in $\mathrm{SL}_2(\mathbb{Z})$. If $\lambda_q = \mathbb{1}_{q_0|q}$ for some positive integer q_0, then we instead have a symmetry under the *congruence subgroup* $\Gamma_0(q_0) := \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) : q_0 \mid c\}$. This suggests that our equidistribution problems have some GL_2 structure, and may be attacked through methods of GL_2 automorphic forms. The usual path (see Figure 1.2) from such *counting problems* to the realm of *automorphic forms* goes through *exponential sums*, which we detail further.

1.1.2 Bounding exponential sums

Let $e(x) := e^{2\pi i x}$ for $x \in \mathbb{R}$ (or $\mathbb{R}/\mathbb{Z}$). A wide class of arithmetic problems reduce, through Fourier analysis (see (1.15)), to proving upper bounds for sums of the shape

$$\sum_{n \le N} e(f(n)),$$

where $f : \mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ is some function that we expect to equidistribute modulo 1. In general, the best-case upper bound for such a sum is around $\sqrt{N}$; this *square-root cancellation* matches the expected behavior of random functions f.

A key type of exponential sum, which comes up when estimating expressions like (1.3) and (1.4), and obeys essentially square-root cancellation, is the *Kloosterman sum*

$$S(m, n; c) := \sum_{\substack{1 \le a, d \le c \\ ad \equiv 1 \pmod{c}}} e\left(\frac{ma + nd}{c}\right), \tag{1.5}$$

for $m, n \in \mathbb{Z}$ and $c \in \mathbb{Z}_+$. Bounding sums of these Kloosterman sums requires a variety of tools from combinatorics, algebraic geometry, and the spectral theory of GL_2 automorphic forms. Indeed, the GL_2 structure is apparent once again in expressions like

$$\sum_{\substack{c \le C \\ c \equiv 0 \pmod{q}}} S(m, n; c) \approx \sum_{\substack{a, b, c, d \le C \\ c \equiv 0 \pmod{q} \\ ad - bc = 1}} e\left(\frac{ma + nd}{c}\right), \tag{1.6}$$

which can be viewed as sums over $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(q)$.

1.2 Three perspectives on automorphy

1.2.1 Automorphic forms as generalized modular forms

In problems that have a symmetry under the action of $\mathrm{SL}_2(\mathbb{Z})$ (or subgroups thereof), it should not be too surprising that one encounters functions which are invariant (or transform nicely) under this action – these are called *automorphic functions*.

In particular, $\mathrm{SL}_2(\mathbb{Z})$ acts on the upper half-plane $\mathbb{H}$ by $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) z = \frac{az+b}{cz+d}$, and we may consider automorphic functions $f : \mathbb{H} \to \mathbb{C}$ which obey

$$f\left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) z\right) = (cz+d)^k f(z), \qquad \forall \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(q) \ \ (\text{so } q \mid c). \tag{1.7}$$

Here, q and k are called the *level* and the *weight* of f. More generally, we can incorporate a factor of $\chi(d)$ in the right-hand side, where χ is a *Dirichlet character* modulo $q_0 \mid q$ (this is a q_0-periodic function $\chi : \mathbb{Z} \to \mathbb{C}$, which is supported on integers coprime with q_0, and completely multiplicative).

Due to spectral decompositions and trace formulae, it is more fruitful to restrict to automorphic functions which also obey some differential equations and growth conditions – these are called *automorphic forms*. Thus in addition to the automorphy condition above, the classical modular forms are holomorphic, while the classical Maass forms (of weight $k = 0$) are eigenfunctions of the *hyperbolic Laplacian*

$$\Delta := -y^2 \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right).$$

To estimate the sums in (1.6), we need to understand both holomorphic forms and Maass forms. The Maass *cusp* forms (which vanish at the cusps of $\Gamma_0(q)\backslash\mathbb{H}$) are often the most problematic, due to very difficult unsolved problems like the following:

Conjecture 1.1 (Selberg)**.** *If f is a Maass cusp form for a congruence subgroup, with $\Delta f = \lambda_f(\infty) f$, then $\lambda_f(\infty) \geq \frac{1}{4}$.*

A closely related conjecture involves the *Hecke operators* T_n, which play a similar role to the hyperbolic Laplacian but at non-Archimedean places. In fact, the spaces of holomorphic and Maass forms are spanned by *Hecke eigenforms* satisfying $T_n f = \lambda_f(n) f$, and the Hecke eigenvalues $\lambda_f(n)$ of a given form are multiplicative.

Conjecture 1.2 (Ramanujan–Petersson)**.** *If $p \nmid q$ is a prime and f is a Maass cusp for $\Gamma_0(q)$ with $T_p f = \lambda_f(p) f$, then $|\lambda_f(p)| \leq 2$.*

Although the corresponding result for holomorphic cusp forms is known by the celebrated work of Deligne [Del71], our inability to prove Conjectures 1.1 and 1.2 affects the best results on many classical questions, including those in Section 1.1.

One can extend these definitions to other congruence subgroups, and more importantly, to the GL_n setting, using functions on a generalized upper half-plane. In the other direction, GL_1 forms are essentially just Dirichlet characters.

1.2.2 Automorphic L-functions as generalized Dirichlet L-functions

Recall that the Riemann zeta function is defined in $\operatorname{Re} s > 1$ by $\zeta(s) := \sum_{n=1}^{\infty} n^{-s}$, and that given a Dirichlet character χ, one similarly defines the Dirichlet L-function $L(s,\chi) := \sum_{n=1}^{\infty} \chi(n) n^{-s}$. More generally, one can collect the Hecke eigenvalues of an automorphic form f (say, a Maass cusp form) into an L-function, defined in $\operatorname{Re} s > 1$ by

$$L(s,f) \approx \sum_{n=1}^{\infty} \frac{\lambda_f(n)}{n^s},$$

up to factors at the *ramified* primes, which divide the level q. Much like the zeta function and the Dirichlet L-functions, these L-functions have:

(i). An Euler product $L(s,f) = \prod_{p \text{ prime}} L_p(s,f)$ in $\operatorname{Re} s > 1$, due to the multiplicativity of Hecke eigenvalues. For GL_2 forms and $p \nmid q$, one can write $L_p(s,f) = (1 - p^{\mu_f(p)-s})^{-1}(1 - p^{-\mu_f(p)-s})^{-1}$, where $p^{\mu_f(p)} + p^{-\mu_f(p)} = \lambda_f(p)$.

(ii). Meromorphic continutation and a functional equation, which stems from the automorphy condition (1.7). Specifically, after incorporating a suitable Gamma factor to form a completed L-function $\Lambda(s,f) = q_f^{s/2} L_\infty(s,f) L(s,f)$, one has

$$\Lambda(s,f) = \varepsilon_f \Lambda(1-s,f), \tag{1.8}$$

for some $q_f \mid q$ and $\varepsilon_f \in \mathbb{C}$ with $|\varepsilon_f| = 1$.

We expect that essentially all L-functions which have both an Euler product and a functional equation come from automorphic forms, and that the arithmetic data associated to these L-functions obeys strong properties. In particular, for a GL_2 Maass cusp form f, each local factor $L_v(s,f)$ (for $v = p$ or $v = \infty$) is defined in terms of local parameters $\mu_f(v)$, which encode (up to a change of variable) either the Hecke eigenvalue $\lambda_f(p)$ if $v = p$, or the Laplacian eigenvalue $\lambda_f(\infty)$ if $v = \infty$. Then the seemingly-unrelated Conjectures 1.1 and 1.2 are together equivalent to the following:

Conjecture 1.3 (Generalized Ramanujan conjecture for GL_2)**.** *If f is a Maass cusp form for $\Gamma_0(q)$, and $v = p \nmid q$ is a prime or $v = \infty$, then $\operatorname{Re} \mu_f(v) = 0$.*

To make these problems even more inter-connected, the best progress towards Conjecture 1.3 (this is $|\operatorname{Re} \mu_f(v)| \leq \frac{7}{64}$, due to Kim–Sarnak [Kim03, Appendix 2]) comes from relating it to the analogous conjecture for GL_5.

Notably, the generalized Ramanujan conjecture would lead to the bound $|\lambda_f(n)| \ll n^{o(1)}$ for the coefficients of cuspidal automorphic L-functions. This is, of course, true for the classical (GL_1) Dirichlet L-functions, when $\lambda_f(n) = \chi(n)$. Beyond the size of the coefficients $\lambda_f(n)$, one can also try to capture their oscillation when varying either f or n. The comparison to Dirichlet characers is once again fruitful: the identity

$$\sum_{\chi \pmod q} \chi(m)\overline{\chi}(n) = \phi(q)\mathbb{1}_{m\equiv n \pmod q} \tag{1.9}$$

corresponds more generally to a trace formula; see (1.18). Writing $n \sim N$ for $N < n \leq 2N$, the duality property

$$\sum_{n\sim N} \chi(n) \approx \varepsilon_\chi \frac{N}{\sqrt{q}} \sum_{h\sim \frac{q}{N}} \overline{\chi}(h), \tag{1.10}$$

for *primitive* characters χ mod q (meaning that χ is not induced by a character mod $d < q$), has an analogue for any L-function; see (1.17). When varying both χ and n with rough coefficients, the orthogonality properties of $\chi(n)$ are encapsulated into *large sieve inequalities* [IK21, §7.5], such as:

$$\sum_{q\leq Q} \sideset{}{^*}\sum_{\chi \pmod q} \left|\sum_{n\leq N} a_n\chi(n)\right|^2 \leq (Q^2+N)\|a\|_2^2, \tag{1.11}$$

where χ varies among the primitive Dirichlet characters mod q, $a_n \in \mathbb{C}$ are arbitrary, and $\|a\|_2^2 = \sum_{n\sim N} |a_n|^2$. Here the $Q^2\|a\|_2^2$ term corresponds to the diagonal terms $n_1 = n_2$ after expanding the square (i.e., to square-root cancellation in the inner sum), while the $N\|a\|_2^2$ term corresponds to the contribution of a single character χ when we choose $a_n = \overline{\chi}(n)$; so this result is essentially best-possible. Obtaining such large sieve inequalities for families of GL_n forms with $n \geq 2$ is an important problem, closely related to our work in this thesis; we discuss this further in Section 1.3.3.

1.2.3 Automorphic representations from the Langlands perspective

It was an important conceptual leap, pioneered by Langlands in a letter to Weil from 1967, to pass from automorphic forms to automorphic *representations*. Very roughly speaking, one can view the automorphic forms f discussed so far as functions on a quotient of $\mathrm{GL}_n(\mathbb{R})$, and then *adelize* them to obtain functions $\widetilde{f}$ on a quotient of $\mathrm{GL}_n(\mathbb{A}_\mathbb{Q})$, where $\mathbb{A}_\mathbb{Q}$ are the *adeles* of $\mathbb{Q}$. Then $\mathrm{GL}_n(\mathbb{A}_\mathbb{Q})$ acts on these adelic automorphic forms by right-translation, which induces automorphic representations.

In particular, if f is a Hecke cusp form, then $\mathrm{GL}_n(\mathbb{A}_\mathbb{Q})$ acts on the span of right-translates of $\widetilde{f}$, and the resulting *cuspidal* automorphic representation π is irreducible; one associates to π the same L-function $L(s,\pi) = L(s,f)$, with Dirichlet coefficients

$\lambda_\pi(n) = \lambda_f(n)$. Since $\mathbb{A}_\mathbb{Q}$ is given by a restricted product over all places v of $\mathbb{Q}$, so is $\pi = \bigotimes_v \pi_v$, and the local factor $L(s, \pi_v)$ is precisely what we called $L_v(s, f)$ in the previous section. This gives a cleaner and broader view of automorphic L-functions.

Now, there is another important source of L-functions that have Euler products and functional equations: Galois representations. We won't cover these, but we mention that they come with some natural operations: one can take symmetric powers of a Galois representation $\rho \mapsto \mathrm{Sym}^k \rho$, or the tensor product of two Galois representations $\rho, \rho' \mapsto \rho \otimes \rho'$, to obtain new Galois representations with associated Artin L-functions.

Langlands famously conjectured that automorphic representations mimic the properties of Galois representations (and that in many cases, they have the same L-functions). In particular, operations on Galois representations should have a counterpart on automorphic representations: one should be able to take 'symmetric powers' $\mathrm{Sym}^k \pi$ and 'products' $\pi \times \pi'$ to obtain new automorphic representations. By analogy with Artin L-functions, we can guess what the corresponding *symmetric power L-functions* and *Rankin–Selberg L-functions* look like, and write them down for large enough $\mathrm{Re}\, s$:

$$L(s, \mathrm{Sym}^k \pi) = \sum_{n=1}^{\infty} \frac{\lambda_{\mathrm{Sym}^k \pi}(n)}{n^s}, \qquad L(s, \pi \times \pi') = \sum_{n=1}^{\infty} \frac{\lambda_{\pi \times \pi'}(n)}{n^s}.$$

Heuristically, one can imagine that the Dirichlet coefficients above are given by

$$\lambda_{\mathrm{Sym}^k \pi}(n) \approx \lambda_\pi(n^k), \qquad \lambda_{\pi \times \pi'}(n) \approx \lambda_\pi(n) \lambda_{\pi'}(n), \tag{1.12}$$

and these identities actually hold when n is a product of distinct unramified primes. Although we do not know in general that $\mathrm{Sym}^k \pi$ and $\pi \times \pi'$ are automorphic, we often know that their L-functions have nice properties, which gives us more flexibility to study the coefficients $\lambda_\pi(n)$ via (1.12). In particular, the generalized Ramanujan conjecture would follow for π if $\mathrm{Sym}^k \pi$ was automorphic for all positive integers k, and the best unconditional progress towards Conjecture 1.3 uses symmetric powers [Kim03]. Thus the Langlands perspective gives us both heuristics and tools to attack problems that involve automorphic forms, even in the classical GL_2 setting.

1.3 Our toolkit

Here we informally describe some useful operations to transform and bound combinatorial, exponential, or automorphic sums. See Chapter 2 for formal statements, which include various error terms, smooth weights, GCD constraints, and factors with sub-polynomial growth.

1.3.1 Involutions

We list a few (approximate) identities, the first couple of which are elementary:

- **Switching moduli.** Let $a, b, \overline{a}, \overline{b}$ be integers such that $a\overline{a} \equiv 1 \pmod{b}$ and $b\overline{b} \equiv 1 \pmod{a}$. Then we have the approximate equality

$$e\left(\frac{\overline{a}}{b}\right) \approx e\left(-\frac{\overline{b}}{a}\right), \tag{1.13}$$

  stemming from the congruence $a\overline{a} + b\overline{b} \equiv 1 \pmod{ab}$. We usually pass to the right-hand side if $a < b$, so that the modulus decreases.

- **Swapping divisors.** Given $f : \mathbb{Z}^2 \to \mathbb{C}$, a positive integer n, $D \geq 1$, and $C := \frac{n}{2D}$, one has

$$\sum_{\substack{D < d \leq 2D \\ d \mid n}} f(d, n) = \sum_{\substack{C \leq c < 2C \\ c \mid n}} f(\tfrac{n}{c}, n), \tag{1.14}$$

  by the simple substitution $c := \frac{n}{d}$. This is usually helpful when $D > \sqrt{n}$, i.e. $C < D$, so that we pass to a shorter sum over a smaller divisor.

- **Truncated Poisson summation.** (See Lemma 2.2.) Up to smooth weights, given $N \geq 1$ and positive integers a, q, one roughly has

$$\sum_{\substack{n \leq N \\ n \equiv a \pmod{q}}} 1 \approx \frac{N}{q} \sum_{|h| \leq \frac{q}{N}} e\left(\frac{ah}{q}\right), \tag{1.15}$$

  the frequency $h = 0$ giving the expected main term of $\frac{N}{q}$. This is usually helpful when $\sqrt{q} < N < q$, so that we pass to a shorter dual sum. Very closely related is the method of *Fourier completion*: if $f : \mathbb{Z} \to \mathbb{C}$ is periodic mod q, then one can deduce from (1.15) that

$$\sum_{n \leq N} f(n) = \sum_{a \pmod{q}} f(a) \sum_{\substack{n \leq N \\ n \equiv a \pmod{q}}} 1 \approx \frac{N}{q} \sum_{|h| \leq \frac{q}{N}} \sum_{a \pmod{q}} e\left(\frac{ah}{q}\right) f(a). \tag{1.16}$$

  One encounters the Kloosterman sums $S(m, n; c)$ from (1.5) this way (as the inner sums over a), when $q = c$, $h = m$, and $f(a) = \sum_{d \pmod{c},\, ad \equiv 1 \pmod{c}} e(\frac{nd}{c})$.

- **Approximate functional equations.** (See Lemmas 5.6 and 6.12.) Generalizing (1.10), one can deduce from (1.8) and some complex analysis that for a cusp form f,

$$\sum_{n \sim N} \lambda_f(n) \approx \varepsilon_f \frac{N}{\sqrt{C_f}} \sum_{h \sim \frac{C_f}{N}} \overline{\lambda}_f(h), \tag{1.17}$$

where C_f is the *analytic conductor* of f (this essentially equals q_f from above (1.8) times an Archimedean factor depending on $L_\infty(s,f)$). Note that the normalization is such that square-root cancellation would give the same bound on both sides of (1.17). Once again, this duality property is usually helpful when $\sqrt{C_f} < N < C_f$, so that the dual sum is shorter. Moreover, one can express values of $L(s,f)$ in the critical strip $\operatorname{Re} s \in (0,1)$ in terms of these dual sums:

$$L(s,f) \approx \sum_{n \le N} \frac{\lambda_f(n)}{n^s} - \varepsilon_f C_f^{\frac{1}{2}-s} \sum_{h \le \frac{C_f}{N}} \frac{\overline{\lambda}_f(h)}{h^{1-s}}.$$

In fact, one can recover the truncated Poisson identity (1.15) from the approximate functional equation (1.17) for Dirichlet characters. Similarly, one can twist a GL_m cusp form f by Dirichlet characters mod q (increasing C_f by roughly q^m), and deduce a *Voronoi summation formula* for additive twists like $\lambda_f(n)\, e(\frac{an}{q})$.

- **Trace formulae.** In the spirit of (1.9), using the spectral theory of automorphic forms, one can obtain identities of the shape

$$\sum_{f \in \mathcal{S}} \lambda_f(m)\overline{\lambda}_f(n) \approx |\mathcal{S}| \mathbb{1}_{m=n} + (\text{Exponential Sums}), \tag{1.18}$$

where $\mathcal{S}$ is a suitable family and $\mathbb{1}_{m=n}$ is 1 or 0 depending on whether $m = n$. In particular, in the GL_2 *Kuznetsov trace formula* (see Proposition 2.6) for automorphic forms of level $q = rs$, $(r,s) = 1$, the exponential sums look like

$$\sum_{c \le C} \frac{1}{c} S(m\overline{r}, n; sc),$$

with $r\overline{r} \equiv 1 \pmod{sc}$ (compare this with (1.6) when $r = 1$, $s = q$). In this case, the shape of the spectral sum from (1.18) will depend on the parameter $X := \frac{C^2 s^2 r}{mn}$; for $X > 1$, the sum will incorporate factors of X^{θ_f}, where $\theta_f \le \frac{7}{64}$ measures the best progress towards Conjecture 1.3. Notably, the Kuznetsov formula is useful both ways: to understand correlations between coefficients of automorphic forms, or to bound sums of Kloosterman sums. Through the latter process, Conjecture 1.3 comes up in applications to classical problems.

- **Duality principle for matrices.** Let (a_m), (b_n), $(C_{m,n})$ denote complex sequences supported on finite sets of m, n. Then

$$\max_{\|a\|_2 = 1} \sum_n \left| \sum_m a_m C_{m,n} \right|^2 = \max_{\|b\|_2 = 1} \sum_m \left| \sum_n b_n C_{m,n} \right|^2, \tag{1.19}$$

since both are equal to the square of the operator norm of the matrix C.

1.3.2 Inequalities

The involutions in Section 1.3.1 can only take one so far before going in circles. At some point in most analytic arguments, one needs to use a genuine inequality, trading some cancellation for more structure. Some examples include:

- **Trivial bounds and counting arguments.** One can ignore all the oscillation in a sum by the triangle inequality, and bound the resulting sum either trivially or by a combinatorial argument. This can be particularly helpful after transformations like (1.17), where the two sides have different trivial bounds. Elementary estimates like the divisor bound $\sum_{d|n} 1 \ll n^{o(1)}$ often come in handy.

- **Cauchy–Schwarz to create smooth sequences.** Given 1-bounded complex sequences (a_m), (b_n), $(C_{m,n})$ supported on $m \leq M$, $n \leq N$, we have

$$\left|\sum_m a_m \sum_n b_n C_{m,n}\right|^2 \leq M \sum_{m \leq M} \left|\sum_n b_n C_{m,n}\right|^2. \tag{1.20}$$

  This eliminates the coefficients (a_m) and allows us to apply transformations like (1.14) or (1.15) to the sum over m, at the expense of forfeiting all initial cancellation in the m-variable. Indeed, the *diagonal* terms $n_1 = n_2$ obtained after expanding the square in the right-hand side of (1.20) already contribute about M^2N, corresponding to square-root cancellation only in the n-variable. Other instances of Hölder's inequality with even exponents are similarly helpful.

- **Positivity.** If we can view a quantity of interest S as a member of a larger family of nonnegative quantities $(S_f)_{f \in \mathcal{F}}$, it may be easier to bound the full sum $\sum_{f \in \mathcal{F}} S_f$. This is beautifully exemplified by the amplification method of Duke, Friedlander and Iwaniec [Fri95], but comes up in many different contexts. Indeed, such a trick will be crucial in our Chapters 5 and 6.

- **Convexity bound.** By the Phragmén–Lindelöf principle, one can bound the values of an L-function in the critical strip $0 < \operatorname{Re} s < 1$ in terms of its values outside the critical strip. By the functional equation, it essentially remains to bound L-functions near the $\operatorname{Re} s = 1$ line, and in particular the residues of L-functions with a pole at $s = 1$; see Lemma 2.14.

- **Algebraic geometry.** For exponential sums over complete variables $n \bmod p$, typically obtained by Fourier completion as in (1.16), methods from algebraic geometry developed by Weil and Deligne can often lead to square-root cancellation. An important case for us is the Weil bound for individual Kloosterman sums (see Lemma 2.1), but we also mention the work of Fouvry, Kowalski, Michel and Sawin on algebraic trace functions and their applications to bilinear sums of Kloosterman sums [FKM14; KMS17; KMS20].

1.3.3 A relevant example

In the spirit of the GL_1 *large sieve inequality* (1.11), one can try to bound more general sums of the shape

$$\sum_{\pi\in\mathcal{S}}\left|\sum_{n\sim N}a_n\lambda_\pi(n)\right|^2, \tag{1.21}$$

where $\mathcal{S}$ is a suitable family of automorphic forms/representations and $\|a\|_2=1$. If the sequence (a_n) is supported on a single n, so the sum becomes $\sum_{\pi\in\mathcal{S}}|\lambda_\pi(n)|^2$, then one can view such a bound as a *density theorem* towards the generalized Ramanujan conjecture (GRC), since it implies that $|\lambda_\pi(n)|$ cannot be too large for too many π's. We essentially have two ways to bound sums like (1.21): via the spectral theory of automorphic forms, or the L-function theory of automorphic representations. These methods give sharp results in different ranges of N and $|\mathcal{S}|$, and are both reflected in this thesis (in Chapters 3 and 4, respectively Chapters 5 and 6).

If the family $\mathcal{S}$ has an associated trace formula, it is natural to expand the square in (1.21) to reach

$$\sum_{\pi\in\mathcal{S}}\left|\sum_{n\sim N}a_n\lambda_\pi(n)\right|^2=\sum_{m,n\sim N}a_m\overline{a_n}\sum_{\pi\in\mathcal{S}}\lambda_\pi(m)\overline{\lambda}_\pi(n),$$

and then apply a trace formula like (1.18). The terms with $\mathbb{1}_{m=n}$ correspond to square-root cancellation in the initial sum over n, while for the remaining Kloosterman-type sums, one can use bounds from algebraic geometry, Fourier analysis, combinatorics, etc. Deshouillers–Iwaniec [DI82c] famously carried this out in the GL_2 setting, with a large number of applications to classical analytic number theory; our Chapters 3 and 4 build on their work. More recently, Blomer and others [Blo23; AB24; BT24] used the GL_n Kuznetsov formula to obtain such results in more general settings.

The second method is to pass, by the duality principle (1.19), from (1.21) to the maximum over sequences (b_π) with $\|b\|_2=1$ of

$$\sum_{n\sim N}\left|\sum_{\pi\in\mathcal{S}}b_\pi\lambda_\pi(n)\right|^2=\sum_{\pi,\pi'\in\mathcal{S}}b_\pi\overline{b_{\pi'}}\sum_{n\sim N}\lambda_\pi(n)\overline{\lambda}_{\pi'}(n).$$

To evaluate the inner sum, one can apply an approximate functional equation like (1.17) for the Rankin–Selberg L-function $L(s,\pi\times\widetilde{\pi}')$, picking up residues at the diagonal terms $\pi=\pi'$, and trivially bounding the resulting dual sums. Duke–Kowalski [DK00] pioneered this approach, which requires formalizing the approximation $\lambda_\pi(n)\overline{\lambda}_{\pi'}(n)\approx\lambda_{\pi\times\widetilde{\pi}'}(n)$ from (1.12). Normally, this argument incurs significant losses unless one assumes GRC for $\pi\in\mathcal{S}$; a key idea in our Chapters 5 and 6 is to pass to $\lambda_{\pi\times\widetilde{\pi}'}(n)$ via a positivity argument instead (using a linear-algebraic formulation), and to avoid unconditional losses this way.

1.4 Structure and related work

In Chapter 2, we formally state some notation and known preliminary results, pertaining to both classical analytic number theory and automorphic forms. Chapters 3 to 6 constitute the main content of this thesis, and can be read independently[1].

In Chapter 3, we build on the aforementioned work of Deshouillers–Iwaniec [DI82c] to prove better large sieve inequalities for the *exceptional* Maass forms (which might fail Conjecture 1.1), and deduce various bounds for sums of Kloosterman sums. We apply this to show that the greatest prime factor of n^2+1 is infinitely-often greater than $n^{1.3}$, building on works of Merikoski [Mer23] and Bretèche–Drappeau [BD20]. This is based on the author's preprint available at [Pas25a].

In Chapter 4, we use the results of Chapter 3 and other ideas to prove better levels of distribution for primes and smooth numbers in arithmetic progressions, matching the best conditional results that assumed Conjecture 1.1. This builds on works of Bombieri–Friedlander–Iwaniec [BFI86], Maynard [May25b], Fouvry–Tenenbaum [FT96], and Drappeau [Dra15]. We then deduce an improved upper bound for the number of twin primes up to x, and a similar result for smooth numbers. This is based on the author's preprint [Pas25b], which built on previous work in [Pas25c].

In Chapter 5, we prove density theorems towards the generalized Ramanujan conjecture (GRC) for GL_n, using Rankin–Selberg L-functions. This was inspired by the pointwise argument of Luo–Rudnick–Sarnak [LRS95], and works well in different ranges than spectral methods [Blo23]. The key new input is a positive semi-definiteness property of Rankin–Selberg coefficients. This is based on joint work with Jared Duker Lichtman, available as a preprint at [LP24].

In Chapter 6, we develop the ideas from Chapter 5 to prove large sieve inequalities for the coefficients of GL_n automorphic forms. We build on the aforementioned approach of Duke–Kowalski [DK00], and obtain sharp results in complementary ranges to spectral methods [Blo23; BT24]. In particular, we improve the previous unconditional results of Thorner–Zaman [TZ21] by completely removing the dependency on progress towards GRC. This is based on joint work-in-progress with Jesse Thorner.

Each of Chapters 3 to 6 has an 'Outline' section which informally explains the key ideas therein, ignoring various technical details much like we did in this chapter.

The ideas in this thesis are also likely to find applications to trilinear sums of Kloosterman fractions, low-lying zeros of Dirichlet L-functions, and smooth values of quadratic polynomials. Other related directions of research include on-average substitutes for GRC at large primes, and improving large sieve inequalities for automorphic forms of varying levels, by combining spectral and L-function methods.

[1] However, Chapter 4 logically depends on Chapter 3, and Chapter 6 depends on Chapter 5.

Finally, we leave the reader with two figures, illustrating the big pictures of how the methods and results in this thesis fit together. Arrows indicate logical implications.[2]

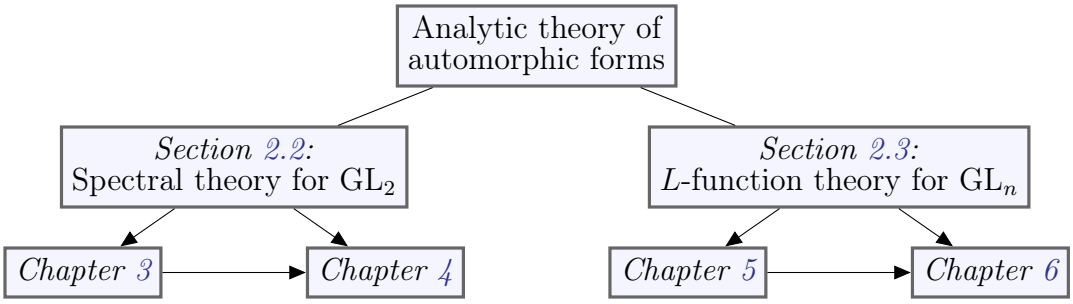

Figure 1.1: How the methods in this thesis are connected

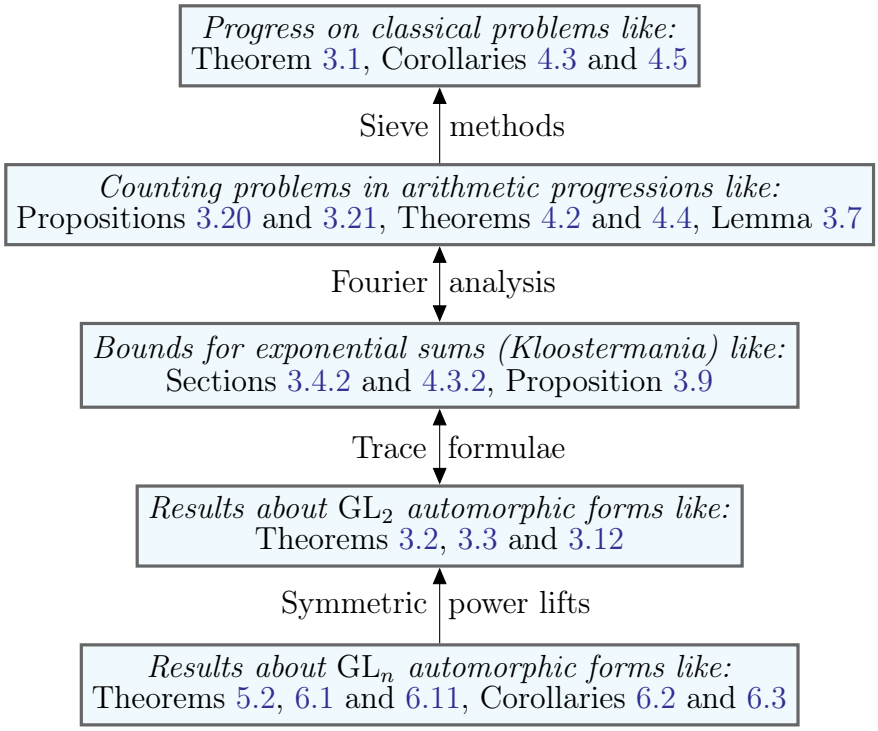

Figure 1.2: How the main types of results in this thesis are connected

[2]Strictly speaking, our new GL_n results in Chapters 5 and 6 have not yet found applications about GL_2 forms through symmetric power lifts, but related results have. For example, Theorem 2.4 relies, by taking a fourth symmetric power, on partial progress towards GRC for GL_5.

Chapter 2

Preliminaries

The notation and preliminary results in this chapter are fairly standard, and formalize some of the intuition given in Chapter 1. Further notation specific to each chapter will be described therein. The reader may prefer to start reading Chapters 3 to 6 directly, and refer back to this chapter when necessary.

2.1 Classical notation and lemmas

2.1.1 Sets, sequences, growth, and some arithmetic

We denote by $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{H}$ the sets of positive integers, rational numbers, real numbers, complex numbers, and 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. Given $n \in \mathbb{Z}_+$ and ring R, we write $\mathrm{GL}_n(R)$, $\mathrm{SL}_n(R)$, $\mathrm{PSL}_n(R)$ for the general, special, and projective special linear groups of degree n. Given $q \in \mathbb{Z}_+$, we write $\Gamma_0^{(n)}(q) \subset \mathrm{SL}_n(\mathbb{Z})$ for the subgroup of matrices with bottom row congruent to $(0, 0, \cdots, 0, *)$ modulo q, which also descends naturally to a subgroup of $\mathrm{PSL}_n(\mathbb{Z})$. When $n = 2$, we may drop the superscript and write $\Gamma_0(q) = \Gamma_0^{(2)}(q)$.

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 $N < n \leq 2N$ (so, e.g., $\mathbb{1}_{n\sim N} = \mathbb{1}_{n\leq 2N} - \mathbb{1}_{n\leq N}$), and interpret sums like $\sum_{n\sim N}$, $\sum_{n\equiv 0 \pmod q}$, or $\sum_{m\leq x}$, $\sum_{d|n}$ with the implied restrictions that $n \in \mathbb{Z}$ and $m, d \in \mathbb{Z}_+$. For $n, k \in \mathbb{Z}_+$, we denote the kth divisor-counting function by $\tau_k(n) := \sum_{d_1\cdots d_k=n} 1$, $\tau(n) := \tau_2(n) = \sum_{d|n} 1$, Euler's totient function by $\varphi(n) := \sum_{m=1}^{n} \mathbb{1}_{(m,n)=1}$, and the Möbius function by $\mu(n)$. 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 $P^+(n) \leq y$.

We use the standard notation $f \asymp_\varepsilon g$, $f \ll_\varepsilon g$, $f = O_\varepsilon(g)$, $f = o(g)$ from analytic number theory, where the subscripts indicate that the implicit constants may depend on the parameter ε. In particular, the statement $f(x) \ll x^{o(1)}g(x)$ is equivalent to the statement that for all $\varepsilon > 0$, we have that $f(x) \ll_\varepsilon x^\varepsilon g(x)$. 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 may 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. The same applies to number fields: we write $(\mathfrak{m}, \mathfrak{n})$ and $[\mathfrak{m}, \mathfrak{n}]$ for the GCD and LCM of two integral ideals.

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 $x \in (\mathbb{Z}/c\mathbb{Z})^\times$, where c is implied from context. We may use the latter notation inside congruences or inside periodic functions modulo c; for example, $x \equiv y\overline{z} \pmod c$ means $xz \equiv y \pmod c$ (where $(z, c) = 1$), and the correct version of (1.13) is

$$e\left(\frac{\overline{a}}{b}\right) = e\left(-\frac{\overline{b}}{a}\right) e\left(\frac{1}{ab}\right).$$

We also recall the classical Kloosterman sums from (1.5), which can be rewritten as

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

for $m, n \in \mathbb{Z}$ and $c \in \mathbb{Z}_+$. The anticipated pointwise bound for Kloosterman sums is:

Lemma 2.1 (Weil and Ramanujan bounds)**.** *For any $m, n \in \mathbb{Z}$ and $c \in \mathbb{Z}_+$,*

$$S(m, n; c) \ll \tau(c)\,(m, n, c)^{1/2}c^{1/2},$$
$$|S(0, n; c)| \le (n, c).$$

Proof. The first bound is due to Weil, and uses algebraic geometry; see [IK21, Corollary 11.12]. The second bound is classical and follows from Möbius inversion. □

2.1.2 The Fourier transform

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

$$\begin{aligned} f : \mathbb{R} \to \mathbb{C} \quad &\rightsquigarrow \quad \widehat{f} : \mathbb{C} \to \mathbb{C}, \qquad \widehat{f}(\xi) := \int_{\mathbb{R}} f(t)\, e(-\xi t)\, dt, \\ \Phi : \mathbb{R} \to \mathbb{C} \quad &\rightsquigarrow \quad \check{\Phi} : \mathbb{C} \to \mathbb{C}, \qquad \check{\Phi}(t) := \int_{\mathbb{R}} \Phi(\xi)\, e(\xi t)\, d\xi, \\ a : \mathbb{Z} \to \mathbb{C} \quad &\rightsquigarrow \quad \widehat{a} : \mathbb{R}/\mathbb{Z} \to \mathbb{C}, \qquad \widehat{a}(\alpha) := \sum_{n \in \mathbb{Z}} a_n\, e(-n\alpha), \\ \varphi : \mathbb{R}/\mathbb{Z} \to \mathbb{C} \quad &\rightsquigarrow \quad \check{\varphi} : \mathbb{Z} \to \mathbb{C}, \qquad \check{\varphi}(n) := \int_{\mathbb{R}/\mathbb{Z}} \varphi(\alpha)\, e(n\alpha)\, d\alpha. \end{aligned} \tag{2.1}$$

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

$$\check{\widehat{\Phi}}\Big|_{\mathbb{R}} = \widehat{\check{\Phi}}\Big|_{\mathbb{R}} = \Phi, \qquad \check{\widehat{a}} = a, \qquad \widehat{\check{\varphi}} = \varphi. \tag{2.2}$$

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,

$$\sum_{n \in \mathbb{Z}} f(n) = \sum_{n \in \mathbb{Z}} \widehat{f}(n) = \sum_{n \in \mathbb{Z}} \check{f}(n). \tag{2.3}$$

In practice, it will be useful to truncate the Poisson summation formula.

Lemma 2.2 (Truncated Poisson with separation of variables)**.** *Let $x \gg 1$ and $1 \ll N, Q \ll x^{O(1)}$, $a \in \mathbb{Z}$, $q \in \mathbb{Z}_+$ with $q \asymp Q$, and $\Phi : (0, \infty) \to \mathbb{C}$ be a smooth function, $\Phi(t)$ supported in $t \asymp 1$, with $\Phi^{(k)} \ll_k 1$ for $k \geq 0$. Then for any $A, \delta > 0$ and $H := x^\delta N^{-1} Q$, one has*

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

where $\Psi_j : (\frac{1}{2}, 2) \to \mathbb{C}$ are some compactly supported functions with $\Psi_j^{(k)} \ll_k 1$ for $k \geq 0$. Note that the integrand is supported in $u \asymp 1$, and that one can rewrite

$$\frac{N}{Q}\Phi\left(\frac{uq}{Q}\right) du = \frac{N}{q}\widetilde{\Phi}\left(\frac{uq}{Q}\right)\frac{du}{u}, \qquad \text{where} \qquad \widetilde{\Phi}(t) := t\Phi(t).$$

Proof. This follows quickly from the Poisson summation formula and a smooth dyadic partition of unity. Note that the variables h and q are separated at the cost of a slowly-varying exponential phase $e(h\omega)$ where $\omega = -uN/Q \ll x^{\delta}H^{-1}$.

Specifically, the Poisson identity (2.3) 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. □

2.1.3 The Mellin transform

Given a bounded smooth function Φ on $(0, \infty)$ with Schwartz decay towards ∞, we define its Mellin transform as

$$\widetilde{\Phi}(s) := \int_0^{\infty} \Phi(x) x^{s-1} dx,$$

in $\operatorname{Re} s > 0$. This is related to the Fourier transform from (2.1) by a change of variables,

$$\widetilde{\Phi}(s) = \widehat{f}\left(-\frac{s}{2\pi i}\right), \qquad f(t) := \Phi(e^t).$$

This function decays rapidly in vertical strips and satisfies the Mellin inversion formula

$$\Phi(x) = \frac{1}{2\pi i}\int_{(\sigma)} x^{-s}\, \widetilde{\Phi}(s)\, ds, \tag{2.4}$$

for any $\sigma > 0$, where the integral is over the vertical line at $\operatorname{Re} s = \sigma$. We recall that the Gamma function is defined as the Mellin transform of e^{-x},

$$\Gamma(s) := \int_0^\infty x^{s-1} e^{-x} \, dx,$$

in $\operatorname{Re} s > 0$, and by meromorphic continuation otherwise. It can be estimated by Stirling's formula,

$$\log \Gamma(s) = \left(s - \frac{1}{2}\right) \log s - s + \frac{\log(2\pi)}{2} + O_\varepsilon(|s|^{-1}),$$

valid in $|\operatorname{Arg} s| < \pi - \varepsilon$. For any $\sigma, t \in \mathbb{R}$ with $\sigma \ll 1$ and $|t| \gg 1$, it follows that

$$\Gamma(\sigma + it) \asymp |t|^{\sigma - \frac{1}{2}} e^{-\frac{\pi}{2}|t|}.$$

(Here the implicit constants are allowed to depend on the constants implied in $\sigma \ll 1$ and $|t| \gg 1$.) Since Γ has no zeros, and poles only at the nonpositive integers, this automatically improves to

$$\begin{cases} \Gamma(\sigma + it) \gg (1 + |t|)^{\sigma - \frac{1}{2}} e^{-\frac{\pi}{2}|t|}, & \text{for } \sigma \ll 1, \\ \Gamma(\sigma + it) \ll (1 + |t|)^{\sigma - \frac{1}{2}} e^{-\frac{\pi}{2}|t|}, & \text{for } \sigma \ll 1, \ \min_{m \in \mathbb{Z}_{\le 0}} |\sigma + it - m| \gg 1. \end{cases} \tag{2.5}$$

It will also be convenient to use the common notation

$$\Gamma_{\mathbb{R}}(s) := \pi^{-s/2} \Gamma(s/2), \qquad \Gamma_{\mathbb{C}}(s) := 2(2\pi)^{-s} \Gamma(s) = \Gamma_{\mathbb{R}}(s) \Gamma_{\mathbb{R}}(s+1). \tag{2.6}$$

From (2.5), it follows that for any complex numbers s, z with $\operatorname{Re} s \ll 1$, $\operatorname{Re} z \ll 1$, and $\operatorname{Re}(1 - s + z) \ge \varepsilon > 0$, one has

$$\frac{\Gamma_{\mathbb{R}}(1 - s + \overline{z})}{\Gamma_{\mathbb{R}}(s + z)} \ll_\varepsilon (1 + |\operatorname{Im}(s + z)|)^{\frac{1}{2} - \operatorname{Re} s}. \tag{2.7}$$

2.2 Spectral theory of GL_2 automorphic forms

2.2.1 Cusps, automorphic forms, and generalized Kloosterman sums

Recall that $\mathrm{PSL}_2(\mathbb{R}) := \mathrm{SL}_2(\mathbb{R})/\{\pm 1\}$ acts on $\mathbb{C} \cup \{\infty\}$ by $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) z := \frac{az+b}{cz+d}$. We will focus on the action of the modular congruence subgroup $\Gamma_0(q) \le \mathrm{PSL}_2(\mathbb{Z})$.

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 [DI82c, Lemma 2.3], the fractions

$$\left\{ \frac{u}{w} \ : \ u, w \in \mathbb{Z}_+, \ (u, w) = 1, \ w \mid q, \ u \le \gcd\left(w, \frac{q}{w}\right) \right\} \tag{2.8}$$

form a maximal set of inequivalent cusps of $\Gamma_0(q)$. Following [DI82c, (1.1)], given a cusp $\mathfrak{a}$ of $\Gamma_0(q)$ and its equivalent representative u/w from (2.8), we denote

$$\mu(\mathfrak{a}) := \frac{\gcd\left(w, \frac{q}{w}\right)}{q}, \tag{2.9}$$

(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 [DI82c, (1.2)], a *scaling matrix* $\sigma_{\mathfrak{a}}$ for a cusp $\mathfrak{a}$ is an element of $\mathrm{PSL}_2(\mathbb{R})$ such that

$$\sigma_{\mathfrak{a}}\infty = \mathfrak{a} \qquad \text{and} \qquad \sigma_{\mathfrak{a}}^{-1}\Gamma_{\mathfrak{a}}\sigma_{\mathfrak{a}} = \Gamma_\infty = \left\{ \left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right) : n \in \mathbb{Z} \right\}. \tag{2.10}$$

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

$$\widetilde{\sigma}_{\mathfrak{a}} = \sigma_{\mathfrak{a}} \left(\begin{smallmatrix} 1 & \alpha \\ 0 & 1 \end{smallmatrix}\right), \tag{2.11}$$

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 [Wat95, p. 195], we will use the canonical choice of scaling matrix

$$\sigma_{\mathfrak{a}} = \tau \cdot \begin{pmatrix} \sqrt{r} & -\overline{s}/\sqrt{r} \\ s\sqrt{r} & \overline{r}\sqrt{r} \end{pmatrix}, \tag{2.12}$$

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 [DI82c, (2.3)], and leads to the simplification of certain extraneous exponential phases. For the cusp $\mathfrak{a} = \infty = \left(\begin{smallmatrix} 1 & 0 \\ -q & 1 \end{smallmatrix}\right)(1/q)$, (2.12) reduces back to the identity matrix.

We refer the reader to the aforementioned work of Deshouillers–Iwaniec [DI82c] for a brief introduction to the classical spectral theory of GL_2 automorphic forms, to [Iwa02; Iwa97; IK21] for a deeper dive into this topic, to [Dra17; Top18; DPR23; BD20; Lic23; Pas25c] for follow-up works and optimizations, and to [Bum97] for the modern viewpoint of automorphic representations. For our purposes, an *automorphic form* of level q, integer weight $k \geq 0$, and trivial nebentypus 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 [DI82c, (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 [DI82c, (1.3)], [Dra17, §4.1.1], [Iwa97]), 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 [DI82c, 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

$$S_{\mathfrak{a}\mathfrak{b}}(m, n; c) := \sum_{\widetilde{d} \in \mathcal{D}_{\mathfrak{a}\mathfrak{b}}(c)} e\left(\frac{m\widetilde{a} + n\widetilde{d}}{c}\right), \tag{2.13}$$

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 [DI82c, p. 239]. We also note the symmetries

$$S_{\mathfrak{a}\mathfrak{b}}(m, -n; c) = \overline{S_{\mathfrak{a}\mathfrak{b}}(-m, n; c)}, \qquad S_{\mathfrak{a}\mathfrak{b}}(m, n; c) = \overline{S_{\mathfrak{b}\mathfrak{a}}(n, m; c)}, \tag{2.14}$$

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 (3.1).

Lemma 2.3 (Explicit Kloosterman sums)**.** *Let $q = rs$ with $r, s \in \mathbb{Z}_+$, $\gcd(r, s) = 1$, and $m, n \in \mathbb{Z}$. Then with the choice of scaling matrices from (2.12), one has $\mathcal{C}_{\infty\, 1/s} \subset \{s\sqrt{r}c : c \in \mathbb{Z}_+,\ (c, r) = 1\}$. In fact, for $c \in \mathbb{Z}_+$ with $(c, r) = 1$, one has*

$$S_{\infty\, 1/s}\left(m, n; s\sqrt{r}c\right) = S(m\overline{r}, n; sc). \tag{2.15}$$

Moreover, let $\mathfrak{a}$ be any cusp of $\Gamma_0(q)$ with $\mu(\mathfrak{a}) = q^{-1}$, and $\sigma_{\mathfrak{a}}$ be as in (2.12). Then $\mathcal{C}_{\mathfrak{a}\mathfrak{a}} \subset q\mathbb{Z}_+$, and for $c \in q\mathbb{Z}_+$, one has

$$S_{\mathfrak{a}\mathfrak{a}}(m, n; c) = S(m, n; c). \tag{2.16}$$

Varying the choice of scaling matrix as in (2.11) would result in an additional factor of $e((n - m)\alpha)$.

Proof. These identities are precisely [Wat95, (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 (2.12) 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 [DI82c, §2]. □

2.2.2 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

$$j(\sigma_{\mathfrak{a}}, z)^{-k} f(\sigma_{\mathfrak{a}} z) = \sum_{n=1}^{\infty} \psi_{\mathfrak{a}}(n)\, e(nz) \tag{2.17}$$

around each cusp $\mathfrak{a}$ of $\Gamma_0(q)$ (see [DI82c, (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 $L^2(\Gamma_0(q)\backslash\mathbb{H})$, consisting of eigenvalues $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \dots$ with no limit point. Around a given cusp $\mathfrak{a}$, Maass cusp forms have Fourier expansions (see [DI82c, (1.15)])

$$u(\sigma_{\mathfrak{a}} z) = y^{1/2} \sum_{n \neq 0} \rho_{\mathfrak{a}}(n) K_{i\kappa}(2\pi|n|y)\, e(mx), \tag{2.18}$$

where $z = x + iy$ and K is the Whittaker function normalized as in [DI82c, p. 264].

(b). *Eisenstein series*, explicitly defined by $E_{\mathfrak{a}}(z; s) := \sum_{\tau \in \Gamma_{\mathfrak{a}} \backslash \Gamma_0(q)} \operatorname{Im}^s(\sigma_{\mathfrak{a}}^{-1} \tau z)$ for $\operatorname{Re} s > 1$, 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 [DI82c, (1.17)] around any cusp $\mathfrak{b}$, involving the Whittaker function and the Kloosterman-resembling coefficients (for $n \in \mathbb{Z}$, $n \neq 0$)

$$\varphi_{\mathfrak{a}\mathfrak{b}}(n; s) := \sum_{c \in \mathcal{C}_{\mathfrak{a}\mathfrak{b}}} c^{-2s} \sum_{\widetilde{d} \in \mathcal{D}_{\mathfrak{a}\mathfrak{b}}(c)} e\left(\frac{n\widetilde{d}}{c}\right). \tag{2.19}$$

We are particularly interested in the *exceptional* Maass cusp forms, which have Laplacian eigenvalues $\lambda_j = \lambda_j(q) \in (0, 1/4)$; there can only be finitely many such forms of each level q, and Selberg conjectured [Sel65] that there are none (this is Conjecture 1.1). With implicit dependencies on q, we denote

$$\kappa_j^2 := \lambda_j - \frac{1}{4} \qquad \text{and} \qquad \theta_j := 2i\kappa_j, \tag{2.20}$$

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

$$\begin{aligned} \theta(q) &:= \sqrt{\max(0, 1 - 4\lambda_1(q))} = \begin{cases} \theta_1(q), & \theta_1(q) \in \mathbb{R}_{>0} \\ 0, & \text{otherwise.} \end{cases}, \\ \theta_{\max} &:= \sup_{q \geq 1} \theta(q), \end{aligned} \tag{2.21}$$

Selberg's eigenvalue conjecture asserts that $\theta_{\max} = 0$, and the best result towards it is due to Kim–Sarnak [Kim03, Appendix 2], based on the automorphy of fourth symmetric power L-functions.

Theorem 2.4 (Kim–Sarnak's eigenvalue bound [Kim03])**.** *One has* $\theta_{\max} \leq \frac{7}{32}$.

Remark. As in [May25a; Pas25c; Lic23], we use Deshouillers–Iwaniec's normalization [DI82c] for the spectral parameters θ_j and the Fourier coefficients $\rho_{j\mathfrak{a}}(n)$ of exceptional Maass forms. In various other works [Top18; Dra17; BD20; Mer23], θ_j and $\rho_{j\mathfrak{a}}(n)$ are rescaled by factors of $1/2$ and $n^{-1/2}$.

Based on earlier work of Kuznetsov [Kuz80], Deshouillers–Iwaniec [DI82c] developed a trace formula relating weighted sums over c of the generalized Kloosterman sums from (2.13) 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

$$\begin{aligned}
\widetilde{\mathcal{B}}_\varphi(r) &:= \int_0^\infty J_r(y)\,\varphi(y)\frac{dy}{y},\\
\widehat{\mathcal{B}}_\varphi(r) &:= \frac{\pi}{\sinh(\pi r)}\int_0^\infty \frac{J_{2ir}(x)-J_{-2ir}(x)}{2i}\varphi(x)\frac{dx}{x},\\
\check{\mathcal{B}}_\varphi(r) &:= \frac{4}{\pi}\cosh(\pi r)\int_0^\infty K_{2ir}(x)\,\varphi(x)\frac{dx}{x},
\end{aligned} \tag{2.22}$$

where K_{it} is the aforementioned Whittaker function, and the Bessel functions J_ℓ, J_{it} are defined as in [DI82c, p. 264–265] (above we slightly departed from the notation in [DI82c; Dra17], to avoid confusion with the Fourier and Mellin transforms). All we will need to know about these transforms are the following bounds.

Lemma 2.5 (Bessel transform bounds [DI82c])**.** *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*

$$\widehat{\mathcal{B}}_\varphi(ir),\ \check{\mathcal{B}}_\varphi(ir) \ll \frac{1+Y^{-2r}}{1+Y}, \qquad \text{for } 0 < r < \frac{1}{2}, \tag{2.23}$$

$$\widetilde{\mathcal{B}}_\varphi(r),\ \widehat{\mathcal{B}}_\varphi(r),\ \check{\mathcal{B}}_\varphi(r) \ll \frac{1+|\log Y|}{1+Y}, \qquad \text{for } r \in \mathbb{R}\setminus\{0\}, \tag{2.24}$$

$$\widetilde{\mathcal{B}}_\varphi(r),\ \widehat{\mathcal{B}}_\varphi(r),\ \check{\mathcal{B}}_\varphi(r) \ll |r|^{-5/2} + |r|^{-3}Y, \qquad \text{for } r \in \mathbb{R},\ |r| \geq 1, \tag{2.25}$$

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

$$\widehat{\mathcal{B}}_\varphi(\kappa) \ll (\kappa^2+1)^{-1}, \qquad \text{for } \kappa \in \mathbb{R}\setminus\{0\}, \tag{2.26}$$

$$\widehat{\mathcal{B}}_\varphi(\kappa) \asymp Y^{-2i\kappa}, \qquad \text{for } \ 0 < i\kappa < \frac{1}{2}. \tag{2.27}$$

Proof. The bounds in (2.23) to (2.25) constitute [DI82c, Lemma 7.1] (note that φ satisfies the requirements in [DI82c, (1.43) and (1.44)] for $(Y, 1)$ in place of (X, Y)). Similarly, (2.26) and the lower bound in (2.27) are [DI82c, (8.2) and (8.3), following from (8.1)], using an appropriate choice of the constants η_1, η_2. The upper bound in (2.27) also follows from [DI82c, (8.1)], but is in fact already covered by (2.23) (using $r = -i\kappa$ and the fact that $\widehat{\mathcal{B}}_\varphi$ is even). □

Finally, we state the Kuznetsov trace formula, following the notation of Deshouillers–Iwaniec [DI82c].

Proposition 2.6 (Kuznetsov trace formula [DI82c; Kuz80])**.** *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*

$$\sum_{c \in \mathcal{C}_{\mathfrak{a}\mathfrak{b}}} \frac{S_{\mathfrak{a}\mathfrak{b}}(m, \mathrm{sgn} \cdot n; c)}{c} \varphi\left(\frac{4\pi\sqrt{mn}}{c}\right) = \begin{cases} \mathcal{H} + \mathcal{M} + \mathcal{E}, & \mathrm{sgn} = 1, \\ \mathcal{M}' + \mathcal{E}', & \mathrm{sgn} = -1, \end{cases} \tag{2.28}$$

with the following notations. Firstly, the holomorphic *contribution is*

$$\mathcal{H} = \frac{1}{2\pi} \sum_{k \in 2\mathbb{Z}_+} \widetilde{\mathcal{B}}_\varphi(k-1) \frac{i^k (k-1)!}{(4\pi\sqrt{mn})^{k-1}} \sum_{j=1}^{h_k(q)} \overline{\psi_{jk\mathfrak{a}}(m)}\, \psi_{jk\mathfrak{b}}(n), \tag{2.29}$$

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 (2.17). Secondly, the Maass *contributions are*

$$\mathcal{M} = \sum_{j=1}^{\infty} \frac{\widehat{\mathcal{B}}_\varphi(\kappa_j)}{\cosh(\pi\kappa_j)} \overline{\rho_{j\mathfrak{a}}(m)}\, \rho_{j\mathfrak{b}}(n), \qquad \mathcal{M}' = \sum_{j=1}^{\infty} \frac{\check{\mathcal{B}}_\varphi(\kappa_j)}{\cosh(\pi\kappa_j)} \rho_{j\mathfrak{a}}(m)\, \rho_{j\mathfrak{b}}(n), \tag{2.30}$$

for any orthonormal basis $(u_j)_j$ of level-q Maass cusp forms, with eigenvalues λ_j (and κ_j, θ_j as in (2.20)), and Fourier coefficients $\rho_{j\mathfrak{a}}(n)$ as in (2.18). Thirdly, the Eisenstein *contributions are*

$$\begin{aligned} \mathcal{E} &= \frac{1}{\pi} \sum_{\mathfrak{c}} \int_{-\infty}^{\infty} \widehat{\mathcal{B}}_\varphi(r) \left(\frac{m}{n}\right)^{-ir} \overline{\varphi_{\mathfrak{c}\mathfrak{a}}\left(m; \frac{1}{2} + ir\right)} \varphi_{\mathfrak{c}\mathfrak{b}}\left(n; \frac{1}{2} + ir\right) dr, \\ \mathcal{E}' &= \frac{1}{\pi} \sum_{\mathfrak{c}} \int_{-\infty}^{\infty} \check{\mathcal{B}}_\varphi(r) (mn)^{ir} \varphi_{\mathfrak{c}\mathfrak{a}}\left(m; \frac{1}{2} + ir\right) \varphi_{\mathfrak{c}\mathfrak{b}}\left(n; \frac{1}{2} + ir\right) dr, \end{aligned} \tag{2.31}$$

where the Fourier coefficients $\varphi_{\mathfrak{a}\mathfrak{b}}(n; s)$ are as in (2.19), and $\mathfrak{c}$ varies over the cusps of $\Gamma_0(q)$.

Proof. This is [DI82c, Theorem 2]. □

Remark. Upon inspecting the Maass contribution (2.30) in light of the bounds (2.23) and (2.27), 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

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

compared to the regular (non-exceptional) spectrum.

2.2.3 Bounds for Fourier coefficients

For holomorphic and Maass cusp forms that are Hecke eigenforms, the nth Fourier coefficient around the cusp $\mathfrak{a} = \infty$ is proportional to the eigenvalue of the Hecke operator T_n, for $(n, q) = 1$. With our normalization, the size of the proportionality factor, which is the first Fourier coefficient, depends on the weight, Laplacian eigenvalue, and level; in particular it grows roughly like $q^{-1/2}$ in the level aspect. In light of the Ramanujan–Petersson conjecture (see Conjecture 1.2), we expect most Fourier coefficients of exceptional Maass forms, which have weight 0 and bounded eigenvalues, to have size $\approx q^{-1/2}$. This will help motivate the bounds in this subsection.

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 [DI82c; Sar95; Kim03; Bum97; GH11a; GH11b]. Notably, the Kim–Sarnak result [Kim03, Appendix 2] also gives the best pointwise bound for Hecke eigenvalues (towards Conjecture 1.2), and thus for individual Fourier coefficients.

Here we are interested in bounding averages over bases of automorphic forms, resembling those that show up in (2.29) to (2.31); 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 including Lemma 2.1.

The first results that we mention keep the index n of the Fourier coefficients fixed, while varying the automorphic form. One can view these results as *density theorems* towards the Ramanujan–Petersson and Selberg eigenvalue conjecture, asserting that the Archimedean parameter X^{θ_j} and the non-Archimedean parameter $\rho_{j\mathfrak{a}}(\mathfrak{n})$ cannot be large for too many forms f_j on $\Gamma_0(q)$. See [Iwa90; Hum18].

Proposition 2.7 (Fourier coefficient bounds with fixed n). *Let $K \gg 1$ and $\varepsilon > 0$. With the notation of Proposition 2.6, each of the three expressions*

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

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

$$\sum_{\lambda_j<1/4} X^{\theta_j}\,|\rho_{j\mathfrak{a}}(n)|^2 \ll_{\varepsilon} (qN)^{\varepsilon}\left(1+(q,n)^{1/2}\,\mu(\mathfrak{a})\,n^{1/2}\right), \tag{2.32}$$

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

Proof. These bounds roughly follow by combining Lemma 2.1 with trace formulas like Proposition 2.6, for $m = n$ and suitable choices of φ. See for example [Top17, 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 [DI82c] was that the bounds in Proposition 2.7 can be improved when averaging over indices $n \sim N$, by exploiting the bilinear structure in m, n of the spectral side of the Kuznetsov formula (2.28). 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 Proposition 2.7.

Proposition 2.8 (Large sieve for the regular spectrum [DI82c]). *Let $K, N \ge 1/2$, $\varepsilon > 0$, and (a_n) be a sequence of complex numbers. With the notation of Proposition 2.6, each of the three expressions*

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

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

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 [DI82c, Theorem 2]. □

Proposition 2.8 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 (2.30), one needs to incorporate factors of X^{θ_j} into the sum over Maass forms, as in (2.32). The following is a preliminary result toward such bounds.

Lemma 2.9 (Preliminary bound for exceptional forms)**.** *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 2.6, one has*

$$\sum_{\lambda_j<1/4} X^{\theta_j} \left|\sum_{n\sim N} a_n \rho_{j\mathfrak{a}}(n)\right|^2 \ll \left|\sum_{c\in\mathcal{C}_{\mathfrak{a}\mathfrak{a}}} \frac{1}{c} \sum_{m,n\sim N} \overline{a_m}\, a_n\, S_{\mathfrak{a}\mathfrak{a}}(m,n;c)\, \Phi\left(\frac{\sqrt{mn}}{c} X\right)\right| + O_\varepsilon\left(1+\mu(\mathfrak{a})N^{1+\varepsilon}\right)\|a_n\|_2^2. \tag{2.33}$$

Proof. This is essentially present in [DI82c] (see [DI82c, first display on p. 271], and [DI82c, (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 Proposition 2.8 with $K = 1/4$, and the bound $\cosh(\pi\kappa) \asymp 1$ for $i\kappa \in [0, 1/4]$ (recall that $i\kappa_j = \theta_j/2 \leq 7/64$ by Theorem 2.4, but the weaker Selberg bound $\theta_j \leq 1/2$ suffices here).

Otherwise, let $\varphi(y) := \Phi(yX(4\pi)^{-1})$, which satisfies all the assumptions in Lemma 2.5 for $Y = 4\pi X^{-1}$; in particular, we have

$$\max(\widehat{\mathcal{B}}_\varphi(r), \widetilde{\mathcal{B}}_\varphi(r)) \ll |r|^{-5/2}, \qquad \text{for } |r| \geq 1, \tag{2.34}$$

$$\widehat{\mathcal{B}}_\varphi(\kappa) \ll \left(\kappa^2+1\right)^{-1}, \qquad \text{for } \kappa \in \mathbb{R}\setminus\{0\}, \tag{2.35}$$

$$\widehat{\mathcal{B}}_\varphi(\kappa) \gg X^{2i\kappa}, \qquad \text{for } 0 < i\kappa < 1/2. \tag{2.36}$$

Now apply Proposition 2.6 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

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

Bounding the contribution of non-exceptional Maass cusp forms, holomorphic cusp forms, and Eisenstein series via (2.34), (2.35), and Proposition 2.8 (in dyadic ranges $K = 2^p$), this reduces to

$$\sum_{c\in\mathcal{C}_{\mathfrak{a}\mathfrak{a}}}\frac{1}{c}\sum_{m,n\sim N}\overline{a_m}\,a_n\,S_{\mathfrak{a}\mathfrak{a}}(m,n;c)\,\Phi\left(\frac{\sqrt{mn}}{c}X\right)=\sum_{\lambda_j<1/4}\frac{\widehat{\mathcal{B}}_\varphi(\kappa_j)}{\cosh(\pi\kappa_j)}\left|\sum_{n\sim N}a_n\,\rho_{j\mathfrak{a}}(n)\right|^2 + O_\varepsilon\left(1+\mu(\mathfrak{a})N^{1+\varepsilon}\right)\|a_n\|_2^2. \tag{2.37}$$

Combining this with the lower bound $\widehat{\mathcal{B}}_\varphi(\kappa_j)\gg X^{\theta_j}$ (due to (2.36)), we recover the desired bound in (2.33). □

From Lemma 2.9, one quickly deduces the following large sieve inequality for exceptional Maass forms, most of which is due to Deshouillers–Iwaniec [DI82c].

Proposition 2.10 (Large sieve for the exceptional spectrum [DI82c])**.** *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 $\theta_j:=\sqrt{1-4\lambda_j}$, one has*

$$\sum_{\lambda_j<1/4}X^{\theta_j}\left|\sum_{n\sim N}a_n\,\rho_{j\mathfrak{a}}(n)\right|^2\ll_\varepsilon(qN)^\varepsilon\left(1+\frac{N}{q}\right)\|a_n\|_2^2, \tag{2.38}$$

for any

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

Proof. It suffices to show that (2.38) holds with $X=1$, $X=\frac{q}{N}$, and $X=\frac{q^2}{N^3}$. The case $X=1$ follows immediately from Proposition 2.8, which is [DI82c, Theorem 2]. Otherwise, by Lemmas 2.3 and 2.9, it suffices to show that

$$\sum_{c\in q\mathbb{Z}_+}\frac{1}{c}\sum_{m,n\sim N}\overline{a_m}\,a_n\,S(m,n;c)\,\Phi\left(\frac{\sqrt{mn}}{c}X\right)\ll_\varepsilon(qN)^\varepsilon\left(1+\frac{N}{q}\right)\|a_n\|_2^2.$$

If $X=\frac{q}{N}$, then by taking $\Phi(t)$ to be supported in $t>2$, the sum vanishes. This is the content of [DI82c, Theorem 5]. Finally, the case $X=\frac{q^2}{N^3}$ follows by bounding the Kloosterman sums pointwise, via Lemma 2.1; see also (3.12). □

Remark. An equivalent (and more common [DI82c; Dra17]) way to phrase results like Proposition 2.10 is that

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

for *any* $X > 0$, and $X_0 = X_0(N, q)$ given by the right-hand side of (3.4). 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 (3.4) for this range is $X \asymp q/N$. But in the critical range $N \asymp q$, Proposition 2.10 is as good as the large sieve inequalities for the full spectrum [DI82c, 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 [DI82c, Theorem 6] improved the admissible range to $X \ll \max(1, (Q/N)^2)$; Lichtman [Lic23] recently refined this to $X \ll \max(1, \min((Q/N)^{32/7}, Q^2/N))$, by making θ-dependencies explicit in [DI82c, §8.2]. We note that these results are still limited at $X \ll 1$ when $N \asymp Q$.

Finally, we will also need the following results of Deshouillers–Iwaniec [DI82c] and Watt [Wat95], which require averaging over the level q, and save more in the X^θ aspect for very special sequences (a_n).

Proposition 2.11 (Large sieve with level averaging 1 [DI82c]). *Let $\varepsilon > 0$, $X > 0$, $N, Q \geq 1/2$, 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 2.10, one has*

$$\sum_{q \sim Q} \sum_{\lambda_j(q) < 1/4} X^{\theta_j(q)} \left| \sum_{n \sim N} e(n\omega)\, \rho_{j\infty_q}(n) \right|^2 \ll_\varepsilon (QN)^\varepsilon\, (Q + N)\, N, \tag{2.39}$$

for any

$$X \ll \max\left(N, \frac{Q^2}{N}\right). \tag{2.40}$$

Proof. This follows immediately from [DI82c, Theorem 1.7] with $X \leftarrow X^{1/2}$. As noted in previous works [Pas25c; BFI87; BD20], although [DI82c, Theorem 7] was only stated for $\alpha = 0$, the same proof holds uniformly in $\alpha \in \mathbb{R}/\mathbb{Z}$. □

Proposition 2.12 (Large sieve with level averaging 2 [Wat95]). *Let ε, $X > 0$, $Q \geq 1/2$, $N_1, N_2, Z \gg 1$, $N := N_1 N_2$, and $\Psi_1(t), \Psi_2(t)$ be smooth functions supported $t \asymp 1$, with $\Phi_i^{(j)} \ll_j Z^j$ for $j \geq 0$. Let $q \in \mathbb{Z}_+$ and ∞_q be the cusp at ∞ of $\Gamma_0(q)$, with the choice of scaling matrix $\sigma_{\infty_q} = Id$. Then with the notation of Proposition 2.10, one has*

$$\sum_{q \sim Q} \sum_{\lambda_j(q) < 1/4} X^{\theta_j(q)} \left| \sum_{n_1, n_2} \Psi_1\left(\frac{n_1}{N_1}\right) \Psi_2\left(\frac{n_2}{N_2}\right) \rho_{j\infty_q}(n_1 n_2) \right|^2 \ll_\varepsilon Q^\varepsilon Z^5\, (Q + N)\, N,$$

for any

$$X \ll \frac{Q^2}{N_1^2 N_2}. \tag{2.41}$$

Proof. This is [Wat95, Theorem 2] with $H = N_1$ and $K = N_2$. In fact, [Wat95, Theorem 2] is stated for functions $\Psi_1(t), \Psi_2(t)$ supported on $t \in [1, 2]$, but the same proof extends to any support $t \asymp 1$ (alternatively, one can use a smooth partition of unity to reduce to functions supported in $[1, 2]$). □

2.3 L-function theory of GL_n representations

We now recall some standard facts about GL_n automorphic representations and their L-functions, where implicit constants may depend on $n \geq 1$. We refer the reader to the books of Bump [Bum97] and Goldfeld–Hundley [GH11a; GH11b] for a comprehensive introduction to these concepts.

Throughout this section, we work over a number field F with ring of integers $\mathcal{O}_F$, absolute norm $\mathrm{N} = \mathrm{N}_{F/\mathbb{Q}}$, absolute discriminant D_F, ring of adeles $\mathbb{A}_F$, and fixed degree $[F : \mathbb{Q}]$. We use $\mathfrak{p}$ (resp. $\mathfrak{n}$) to denote a nonzero prime ideal (resp. a nonzero ideal) of $\mathcal{O}_F$. On a first read, it may help to pretend that $F = \mathbb{Q}$; in particular, we will have $F = \mathbb{Q}$ in Chapter 5, and then our notation may identify ideals $(m) \subset \mathbb{Z} = \mathcal{O}_{\mathbb{Q}}$ with positive integers m.

Let $\mathfrak{F}_n$ denote the family of all cuspidal automorphic representations π of $\mathrm{GL}_n(\mathbb{A}_F)$, normalized to have unitary central characters ω_π which are trivial on the diagonally-embedded positive reals $\mathbb{R}_{>0} \subset \mathbb{A}_{\mathbb{Q}}^\times$. In particular, the representations in $\mathfrak{F}_1$ (which include the trivial representation) correspond to the primitive Dirichlet characters when $F = \mathbb{Q}$. Similarly, the (Maass or holomorphic) primitive Hecke cusp forms discussed in Section 2.2 have associated representations in $\mathfrak{F}_2$. We may express any $\pi \in \mathfrak{F}_n$ as a restricted tensor product $\bigotimes_v \pi_v$ of smooth admissible representations of $\mathrm{GL}_n(F_v)$, where v varies over places of F; finitely many of these may be ramified.

2.3.1 Automorphic L-functions, locally

Let $\pi \in \mathfrak{F}_n$. At the non-Archimedean places, which are given by prime ideals $v = \mathfrak{p}$, the local component $\pi_{\mathfrak{p}}$ is described by n Satake parameters $\alpha_{\pi,1}(\mathfrak{p}), \dots, \alpha_{\pi,n}(\mathfrak{p}) \in \mathbb{C}$, which make up the local L-function

$$L(s, \pi_{\mathfrak{p}}) := \prod_{j=1}^{n} (1 - \alpha_{\pi,j}(\mathfrak{p}) \mathrm{N}\mathfrak{p}^{-s})^{-1} = \sum_{k=0}^{\infty} \frac{\lambda_\pi(\mathfrak{p}^k)}{\mathrm{N}\mathfrak{p}^{ks}}. \tag{2.42}$$

This has an associated local conductor $\mathfrak{q}_{\pi_{\mathfrak{p}}}$, which is a power of $\mathfrak{p}$, and equals $(1) = \mathcal{O}_F$ iff $\pi_{\mathfrak{p}}$ is unramified.

Similarly, at each Archimedean place v, the local component π_v is described by n Langlands parameters $\mu_{\pi,1}(v), \dots, \mu_{\pi,n}(v) \in \mathbb{C}$, from which one defines

$$L(s, \pi_v) := \prod_{j=1}^{n} \Gamma_{F_v}(s - \mu_{j,\pi}(v)),$$

where the localization F_v could be $\mathbb{R}$ or $\mathbb{C}$, and must be $\mathbb{R}$ if π_v is unramified (spherical); recall the notation in (2.6). To bring the Archimedean and non-Archimedean places on a similar footing, we may denote

$$(\mathrm{N}\mathfrak{p})^{\mu_{\pi,j}(\mathfrak{p})} := \alpha_{\pi,j}(\mathfrak{p}), \tag{2.43}$$

where $\operatorname{Im} \mu_{\pi,j}(\mathfrak{p})$ is only defined[1] modulo $\frac{2\pi}{\log \mathrm{N}\mathfrak{p}}$. At the ramified primes we may have $\alpha_{\pi,j}(\mathfrak{p}) = 0$, case in which we take $\mu_{\pi,j}(\mathfrak{p}) = -\infty$.

If $\widetilde{\pi} \in \mathfrak{F}_n$ denotes the contragredient representation to π, then we have $\mathfrak{q}_{\pi_\mathfrak{p}} = \mathfrak{q}_{\widetilde{\pi}_\mathfrak{p}}$ for each $\mathfrak{p}$. Moreover, for every place v (Archimedean or not), we have the equality of multisets

$$\{\mu_{\widetilde{\pi},j}(v) : j \leq n\} = \{\overline{\mu_{\pi,j}(v)} : j \leq n\}.$$

By combining the work in [LRS99; MS04], we know that there exists

$$0 \leq \theta_n \leq \frac{1}{2} - \frac{1}{n^2+1} \tag{2.44}$$

such that

$$\operatorname{Re} \mu_{\pi,j}(v) \leq \theta_n. \tag{2.45}$$

At the non-Archimedean places $v = \mathfrak{p}$, this means that $|\alpha_{\pi,j}(\mathfrak{p})| \leq \mathrm{N}\mathfrak{p}^{\theta_n}$. Remarkably, the fact that $\operatorname{Re} \mu_{\pi,j}(v) < \frac{1}{2}$ follows from purely local considerations [Sar05, (12), (13)]; see also [RS96, (2.2), (2.5)].

At the unramified places v, the fact that the central character ω_π is unitary leads to the further symmetry

$$\{\mu_{\pi,j} : j \leq n\} = \{-\overline{\mu}_{\pi,j} : j \leq n\}. \tag{2.46}$$

One of the biggest open problems in the theory of automorphic forms and representations remains:

Conjecture 2.13 (Generalized Ramanujan conjecture)**.** *One can take $\theta_n = 0$ in (2.45), for all places v. Due to (2.46), if v is unramified, this means that $\operatorname{Re} \mu_{\pi,j} = 0$.*

[1] So $\mu_{\pi,j}(\mathfrak{p}) \in (\mathbb{R} \cup \{-\infty\}) + i\mathbb{R}/\frac{2\pi}{\log \mathrm{N}\mathfrak{p}}\mathbb{Z}$. One can still add and conjugate such quantities.

This includes Conjecture 1.3 when $n = 2$ and $F = \mathbb{Q}$. We remark that Conjecture 2.13 would follow from the automorphy of all symmetric powers [Ser98].

Let us briefly describe how the local parameters of automorphic representations relate to the Laplacian and Hecke eigenvalues of GL_n Maass cusp forms for $n \geq 2$ (the latter are described in Section 2.2 for $n = 2$ and a trivial nebentypus, and are automorphic functions on a generalized upper half-plane for $n \geq 3$). If $F = \mathbb{Q}$ and $\pi \in \mathfrak{F}_n$ is induced by a Maass Hecke cusp form f for $\Gamma_0^{(n)}(q)$, the following hold:

- At $v = \infty$, the local parameters of π correspond to the *Laplacian eigenvalue* $\lambda_f(\infty)$ by [Ter88, p. 49, pp. 185-6]

$$\lambda_f(\infty) = \frac{n^3 - n}{24} - \frac{1}{2}\sum_{j=1}^{n} \mu_{\pi,j}(\infty)^2.$$

  Keeping n fixed, this implies that one has $\lambda_f(\infty) \ll 1$ iff for each j, $\mu_{\pi,j}(\infty) \ll 1$ (since the real parts always satisfy $\operatorname{Re} \mu_{\pi,j}(\infty) \ll 1$).

- At the unramified primes $v = p < \infty$, the local parameters of π correspond to the (appropriately normalized) p^{th} *Hecke eigenvalue* $\lambda_f(p)$ by

$$\lambda_f(p) = \lambda_\pi(p) = \sum_{j=1}^{n} p^{\mu_{\pi,j}(p)}.$$

- The ramified primes $v = p < \infty$, and in fact the local conductors $\mathfrak{q}_{\pi_p}$ (which are positive integers when $F = \mathbb{Q}$), divide the level q [JPS81; Blo23].

- The central character ω_π is an adelization of the *nebentypus* χ_f (indicating how f transforms under the action of $\Gamma_0^{(n)}(q)$).

In particular, when $n = 2$, the Laplacian and Hecke eigenvalues of f uniquely determine the local parameters of π at the unramified places. It is now easy to see that Conjecture 2.13 particularizes to Conjectures 1.1 and 1.2.

2.3.2 Automorphic L-functions, globally

The standard L-function $L(s, \pi)$ associated to $\pi \in \mathfrak{F}_n$ and its associated *arithmetic conductor* are given by

$$L(s,\pi) := \prod_{\mathfrak{p}} L(s, \pi_{\mathfrak{p}}) = \sum_{\mathfrak{n}} \frac{\lambda_\pi(\mathfrak{n})}{\mathrm{N}\mathfrak{n}^s}, \qquad \mathfrak{q}_\pi := \prod_{\mathfrak{p}} \mathfrak{q}_{\pi_{\mathfrak{p}}},$$

so in particular, all ramified primes satisfy $\mathfrak{p} \mid \mathfrak{q}_\pi$. The Euler product and Dirichlet series converge absolutely when $\operatorname{Re} s > 1$. For example, if $n = 1$ and $\pi = \mathbb{1}$ is the trivial representation, then $L(s, \pi) = \zeta_F(s)$ is the Dedekind zeta function for F.

The Archimedean factors are similarly gathered into

$$L(s,\pi_\infty) := \prod_{v|\infty} L(s,\pi_v).$$

Let r_π be the order of the pole of $L(s,\pi)$ at $s=1$; this is 1 iff π is the trivial representation, and 0 otherwise. Then the completed L-function

$$\Lambda(s,\pi) := (s(s-1))^{r_\pi}(D_F^n \mathrm{N}\mathfrak{q}_\pi)^{s/2} L(s,\pi_\infty)L(s,\pi)$$

is entire of order 1, and there exists a complex number ε_π of modulus 1 such that for all $s \in \mathbb{C}$, we have the functional equation (anticipated in (1.8))

$$\Lambda(s,\pi) = \varepsilon_\pi \Lambda(1-s,\widetilde{\pi}). \tag{2.47}$$

Since $\Lambda(s,\pi)$ is entire of order 1, there exist complex numbers a_π and b_π such that

$$\Lambda(s,\pi) = e^{a_\pi + b_\pi s} \prod_{\Lambda(\rho,\pi)=0} \Big(1-\frac{s}{\rho}\Big)e^{s/\rho}.$$

The zeros ρ in the above Hadamard product are the nontrivial zeros of $L(s,\pi)$, and the zeros of $L(s,\pi)$ that arise as poles of $s^{r_\pi}L(s,\pi_\infty)$ are the trivial zeros.

Now let $d(v)=1$ if $F_v=\mathbb{R}$ and $d(v)=2$ if $F_v=\mathbb{C}$. Following Iwaniec–Sarnak [IS00], the completed L-function has an associated *analytic conductor*

$$C_\pi := D_F^n \mathrm{N}\mathfrak{q}_\pi \prod_{v|\infty}\prod_{j=1}^{n} \left(3+|\mu_{\pi,j}(v)|^{d(v)}\right). \tag{2.48}$$

This controls the dual lengths in approximate functional equations, as in (1.17). Relatedly, for $\varepsilon>0$ and $\mathrm{Re}\, s \in (-O(1), \frac{1}{2}-\varepsilon]$, it follows from (2.7) and (2.47) that

$$\frac{L(s,\pi)}{L(1-s,\widetilde{\pi})} = \varepsilon_\pi (D_F^n \mathrm{N}\mathfrak{q}_\pi)^{\frac{1}{2}-s} \frac{L(1-s,\pi_\infty)}{L(s,\pi_\infty)} \ll_\varepsilon (1+|s|)^{O(1)} C_\pi^{\frac{1}{2}-\mathrm{Re}\, s}. \tag{2.49}$$

In Chapter 5, we will also need some closely-related non-standard notation. We define the *total conductor* of π by

$$\mathfrak{C}_\pi := D_F^n \mathrm{N}\mathfrak{q}_\pi \left(\max_{v|\infty}\max_{j\le n}\left(3+|\mu_{\pi,j}(v)|^{d(v)}\right)\right)^{n\#\{v|\infty\}}. \tag{2.50}$$

In particular, we have $C_\pi \le \mathfrak{C}_\pi$. Beyond the arithmetic conductor, the analytic and total conductors C_π, $\mathfrak{C}_\pi$ include an Archimedean factor which capture the growth of $L(s,\pi_\infty)$. In particular, for automorphic representations induced by Maass cusp forms, C_π and $\mathfrak{C}_\pi$ encode the growth of both the level and the Laplacian eigenvalue.

Finally, one can also *twist* any cuspidal automorphic representation π by any Hecke character of the idele class group $F^\times\backslash\mathbb{A}_F^\times$, to obtain another automorphic representation $\pi\otimes\chi$ [Bum97, p. 305]; this multiplies the underlying automorphic forms on $\mathrm{GL}_n(F)\backslash\mathrm{GL}_n(\mathbb{A}_F)$ by $\chi(\det(\cdot))$. If π has central character ω_π, then $\pi\otimes\chi$ has central character $\chi^n\omega_\pi$, so in particular $\pi\otimes\chi$ remains unitary if χ is unitary. As one would expect, the contragredient of $\pi\otimes\chi$ is $\widetilde{\pi}\otimes\overline{\chi}$. Following [RS96, Appendix], when $\chi=|\cdot|^z$ for some $z\in\mathbb{C}$, we may also write

$$\pi[z]:=\pi\otimes|\cdot|^z.$$

The unitary twists by $|\cdot|^{it}$, for $t\in\mathbb{R}$, affect the L-function by $L(s,\pi[it])=L(s+it,\pi)$. Our normalization ensures that only one representation from each family $\{\pi[it]:t\in\mathbb{R}\}$ is included in $\mathfrak{F}_n$, but all definitions so far apply without the normalization assumption. For example, it is helpful to write

$$C_\pi(it)=C_{\pi[it]}=D_F^n\mathrm{N}\mathfrak{q}_\pi\prod_{v|\infty}\prod_{j=1}^{n}\left(3+|\mu_{\pi,j}(v)-it|^{d(v)}\right).$$

The advantage of the total conductor $\mathfrak{C}_\pi$ is that it is more stable under Archimedean twists than C_π.

When $F=\mathbb{Q}$ and χ is induced by a primitive even Dirichlet character of prime conductor $q\nmid\mathfrak{q}_\pi$ (as in [GH11a, Definition 2.1.7]), the twist affects the L-function by [LRS95]

$$L(s,\pi\otimes\chi)=\sum_{m=1}^{\infty}\frac{\lambda_\pi(m)\chi(m)}{m^s},\qquad\qquad L(s,(\pi\otimes\chi)_\infty)=L(s,\pi_\infty),$$

the conductors by

$$\frac{\mathfrak{q}_{\pi\otimes\chi}}{\mathfrak{q}_\pi}=\frac{C_{\pi\otimes\chi}}{C_\pi}=\frac{\mathfrak{C}_{\pi\otimes\chi}}{\mathfrak{C}_\pi}=q^n,$$

and the 'sign' $\varepsilon_{\pi\otimes\chi}$ by an explicit Gauss sum factor.

2.3.3 Rankin–Selberg L-functions

Fix n,n', and let $\pi\in\mathfrak{F}_n$ and $\pi'\in\mathfrak{F}_{n'}$. If the Langlands conjectures are true, there should be a (not necessarily cuspidal) unitary automorphic representation $\pi\times\pi'$ of $\mathrm{GL}_{n^2}(\mathbb{A}_F)$, corresponding to the tensor product of the Galois representations associated to π and π'. Unconditionally, we can still associate to the pair (π,π') an L-function $L(s,\pi\times\pi')$, through the theory of Rankin–Selberg L-functions. This was developed by Jacquet, Piatetski-Shapiro and Shalika [JPS83], Shahidi [Sha81], and Mœglin–Waldspurger [MW89]; we also refer the reader to [Bru06; LRS95; RS96; DK00] for exposition, basic properties, and some applications.

Most properties of automorphic L-functions have counterparts for Rankin–Selberg L-functions. At each prime ideal $\mathfrak{p}$, Jacquet, Piatetski-Shapiro, and Shalika [JPS83] associate $n'n$ Satake parameters $\alpha_{\pi\times\pi',j,j'}(\mathfrak{p})$ to $\pi_\mathfrak{p}$ and $\pi'_\mathfrak{p}$, which make up the local Rankin–Selberg L-function

$$L(s,\pi_\mathfrak{p}\times\pi'_\mathfrak{p}) := \prod_{j=1}^{n}\prod_{j'=1}^{n'}(1-\alpha_{\pi\times\pi',j,j'}(\mathfrak{p})\mathrm{N}\mathfrak{p}^{-s})^{-1} = \sum_{k=0}^{\infty}\frac{\lambda_{\pi\times\pi'}(\mathfrak{p}^k)}{\mathrm{N}\mathfrak{p}^{ks}} \tag{2.51}$$

and a local conductor $\mathfrak{q}_{\pi_\mathfrak{p}\times\pi'_\mathfrak{p}}$. If $\mathfrak{p}\nmid\mathfrak{q}_\pi\mathfrak{q}_{\pi'}$, then $\mathfrak{q}_{\pi_\mathfrak{p}\times\pi'_\mathfrak{p}}$.

At an Archimedean place v of F, Jacquet, Piatetski-Shapiro, and Shalika associate $n'n$ complex Langlands parameters $\mu_{\pi\times\pi',j,j'}(v)$ to π_v and π'_v, from which one defines

$$L(s,\pi_v\times\pi'_v) := \prod_{j=1}^{n}\prod_{j'=1}^{n'}\Gamma_{F_v}(s-\mu_{\pi\times\pi',j,j'}(v)). \tag{2.52}$$

Using the same convention as in (2.43), we may write

$$(\mathrm{N}\mathfrak{p})^{\mu_{\pi\times\pi',j,j'}(\mathfrak{p})} := \alpha_{\pi\times\pi',j,j'}(\mathfrak{p}).$$

For all places v (Archimedean or not), from the explicit descriptions of local parameters from [HB19; ST19], one sees that

$$\operatorname{Re}\mu_{\pi\times\pi',j,j'} \le \theta_n+\theta_{n'}, \tag{2.53}$$

where θ_n is as in (2.45). When $v=\mathfrak{p}$, this reads $|\alpha_{\pi\times\pi',j,j'}(\mathfrak{p})|\le\mathrm{N}\mathfrak{p}^{\theta_n+\theta_{n'}}$.

Moreover, if both π_v and π'_v are unramified, we have equalities

$$\mu_{\pi\times\pi',j,j'}(v) = \mu_{\pi,j}(v)+\mu_{\pi,j'}(v).$$

At the non-Archimedean places $v=\mathfrak{p}$, this reads $\alpha_{\pi\times\pi',j,j'}(\mathfrak{p}) = \alpha_{\pi,j}(\mathfrak{p})\alpha_{\pi',j'}(\mathfrak{p})$.

The Rankin–Selberg L-function $L(s,\pi\times\pi')$ associated to π and π' and its *arithmetic conductor* are

$$L(s,\pi\times\pi') := \prod_{\mathfrak{p}}L(s,\pi_\mathfrak{p}\times\pi'_\mathfrak{p}) = \sum_{\mathfrak{n}}\frac{\lambda_{\pi\times\pi'}(\mathfrak{n})}{\mathrm{N}\mathfrak{n}^s}, \qquad \mathfrak{q}_{\pi\times\pi'} = \prod_{\mathfrak{p}}\mathfrak{q}_{\pi_\mathfrak{p}\times\pi'_\mathfrak{p}},$$

where the convergence is absolute in $\operatorname{Re}s>1$. Similarly, we let

$$L(s,\pi_\infty\times\pi'_\infty) := \prod_{v\mid\infty}L(s,\pi_v\times\pi'_v).$$

Let $r_{\pi\times\pi'} = -\mathrm{ord}_{s=1}L(s,\pi\times\pi')$. By our normalization for the central characters of π and π', we have that $r_{\pi\times\pi'}=0$ if and only if $\pi\ne\widetilde{\pi}'$, and $r_{\pi\times\widetilde{\pi}}=1$ otherwise. The function

$$\Lambda(s,\pi\times\pi') = (s(s-1))^{r_{\pi\times\pi'}}(D_F^{n'n}\mathrm{N}\mathfrak{q}_{\pi\times\pi'})^{s/2}L(s,\pi_\infty\times\pi'_\infty)L(s,\pi\times\pi') \tag{2.54}$$

is entire of order 1, and there exists a complex number $\varepsilon_{\pi\times\pi'}$ of modulus 1 such that we have the functional equation

$$\Lambda(s,\pi\times\pi') := \varepsilon_{\pi\times\pi'}\Lambda(1-s,\widetilde{\pi}\times\widetilde{\pi}'). \tag{2.55}$$

Since $\Lambda(s,\pi\times\pi')$ is entire of order 1, there exist complex numbers $a_{\pi\times\pi'}$ and $b_{\pi\times\pi'}$ such that the Hadamard factorization

$$\Lambda(s,\pi\times\pi') = e^{a_{\pi\times\pi'}+b_{\pi\times\pi's}} \prod_{\Lambda(\rho,\pi\times\pi')=0} \Big(1-\frac{s}{\rho}\Big)e^{s/\rho} \tag{2.56}$$

holds. The zeros ρ in (2.56) are the nontrivial zeros of $L(s,\pi\times\pi')$, and the zeros of $L(s,\pi\times\pi')$ that arise as poles of $s^{r_{\pi\times\pi'}}L(s,\pi_\infty\times\pi'_\infty)$ are the trivial zeros.

As with automorphic L-functions, the completed Rankin–Selberg L-functions have an associated *analytic conductor*

$$C_{\pi\times\pi'} := D_F^{n'n}\mathrm{N}\mathfrak{q}_{\pi\times\pi'}\prod_{v|\infty}\prod_{j=1}^{n}\prod_{j'=1}^{n'}\left(3+|\mu_{\pi\times\pi',j,j'}(v)|^{d(v)}\right), \tag{2.57}$$

which plays a similar role as before. For $\varepsilon>0$ and $\operatorname{Re}s\in(-O(1),-\varepsilon]$, it follows from (2.7) and (2.55) that

$$\begin{aligned}\frac{L(s,\pi\times\pi')}{L(1-s,\widetilde{\pi}\times\widetilde{\pi}')} &= \varepsilon_{\pi\times\pi'}(D_F^{n'n}\mathrm{N}\mathfrak{q}_{\pi\times\pi'})^{\frac12-s}\frac{L(1-s,\pi_\infty\times\pi'_\infty)}{L(s,\pi_\infty\times\pi'_\infty)}\\ &\ll_\varepsilon (1+|s|)^{O(1)}\,C_{\pi\times\pi'}^{\frac12-\operatorname{Re}s}.\end{aligned} \tag{2.58}$$

We also define the *total conductor* (which is, once again, non-standard notation) by

$$\mathfrak{C}_{\pi\times\pi'} := D_F^{n'n}\mathrm{N}\mathfrak{q}_{\pi\times\pi'}\left(\max_{v|\infty}\max_{j\le n}\max_{j'\le n'}\left(3+|\mu_{\pi\times\pi',j,j'}(v)|^{d(v)}\right)\right)^{n'n\#\{v|\infty\}}, \tag{2.59}$$

so that $C_{\pi\times\pi'}\le\mathfrak{C}_{\pi\times\pi'}$. The combined work of Bushnell and Henniart [BH97] and Brumley [HB19, Appendix] yields

$$\mathfrak{q}_{\pi\times\pi'}\ll\mathfrak{q}_\pi^{n'}\mathfrak{q}_{\pi'}^{n}, \qquad\qquad C_{\pi\times\pi'}\ll C_\pi^{n'}C_{\pi'}^{n}, \tag{2.60}$$

and a simpler version of Brumley's computations also gives

$$\mathfrak{C}_{\pi\times\pi'}\ll\mathfrak{C}_\pi^{n'}\mathfrak{C}_{\pi'}^{n}. \tag{2.61}$$

We can also effectively *twist* a Rankin–Selberg L-function by a Hecke character, by twisting one of the two original representations. For twists by $|\cdot|^{it}$, this affects the L-function by $L(s,\pi[it]\times\widetilde{\pi}') = L(s+it,\pi\times\widetilde{\pi}')$. In particular, it is helpful to write

$$C_{\pi\times\pi'}(it) := C_{\pi[it]\times\pi'} = D_F^{n'n}\mathrm{N}\mathfrak{q}_{\pi\times\pi'}\prod_{v|\infty}\prod_{j=1}^{n}\prod_{j'=1}^{n'}(3+|it-\mu_{\pi\times\pi',j,j'}(v)|^{d(v)}),$$

which is immediately bounded by $(3+|t|)^{O(nn'[F:\mathbb{Q}])}C_{\pi\times\pi'}$.

Lemma 2.14 (Li / Harcos–Thorner [HT22]). *For $(\pi, \pi') \in \mathfrak{F}_n \times \mathfrak{F}_{n'}$, consider the holomorphic function*

$$\mathcal{L}(s, \pi \times \pi') = \lim_{s_0 \to s} \left(\frac{s_0 - 1}{s_0 + 1}\right)^{r_{\pi\times\pi'}} L(s_0, \pi \times \pi'), \qquad \operatorname{Re} s > -1. \tag{2.62}$$

If $j \geq 0$, $\sigma \geq 0$, $t \in \mathbb{R}$, and $\varepsilon > 0$, then

$$\mathcal{L}^{(j)}(\sigma + it, \pi \times \pi') \ll_{j,\varepsilon} C_{\pi\times\pi'}(it)^{\max(1-\sigma,0)/2+\varepsilon}. \tag{2.63}$$

In particular, if $r_{\pi\times\pi'} = 0$ or $|t| > 1$, then

$$L^{(j)}(\sigma + it, \pi \times \pi') \ll_{j,\varepsilon} C_{\pi\times\pi'}(it)^{\max(1-\sigma,0)/2+\varepsilon}.$$

Proof. This is [HT22, Lemma 3.2], based on the convexity bound of Li [Li10, Theorem 2]. □

Finally, suppose $F = \mathbb{Q}$, and let χ be a primitive, even Dirichlet character of prime conductor $q \nmid \mathfrak{q}_\pi \mathfrak{q}_{\pi'}$; in this case, the twist by χ affects the Rankin–Selberg L-function[2] by

$$L(s, (\pi\otimes\chi)\times\widetilde{\pi}') = \sum_{m=1}^{\infty} \frac{\lambda_{\pi\times\widetilde{\pi}'}(m)\chi(m)}{m^s}, \quad L(s, (\pi\otimes\chi)_\infty\times\widetilde{\pi}'_\infty) = L(s, \pi_\infty\times\widetilde{\pi}'_\infty), \tag{2.64}$$

and its conductors by

$$\frac{\mathfrak{q}_{(\pi\otimes\chi)\times\widetilde{\pi}'}}{\mathfrak{q}_{\pi\times\pi'}} = \frac{C_{(\pi\otimes\chi)\times\widetilde{\pi}'}}{C_{\pi\times\pi'}} = \frac{\mathfrak{C}_{(\pi\otimes\chi)\times\widetilde{\pi}'}}{\mathfrak{C}_{\pi\times\pi'}} = q^{n'n}.$$

One can also explicitly compute the 'sign' $\varepsilon_{(\pi\otimes\chi)\times\widetilde{\pi}'}$; see [LRS95, Lemma 2.1].

[2] Note that we reserve the notation $\otimes$ for twists, and $\times$ for Rankin–Selberg L-functions.

Chapter 3

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

3.1 Introduction

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

$$S(m,n;c) := \sum_{x \in (\mathbb{Z}/c\mathbb{Z})^\times} e\left(\frac{mx+n\overline{x}}{c}\right), \tag{3.1}$$

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 [BFI86; May25a; May25b; May25c; DI82a; BD20; Mer23; Lic23] and properties of Dirichlet L-functions [DI82b; DI84; Wat95; You11; DPR23; Top18], rely on bounding exponential sums of the form

$$\sum_{m\sim M} a_m \sum_{n\sim N} b_n \sum_{(c,r)=1} g\left(\frac{c}{C}\right) S(m\overline{r}, \pm n; sc), \tag{3.2}$$

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 [Mer23; BD20; May25a]) leverage some additional averaging over r and s, if one of the sequences $(a_m), (b_n)$ is independent of r, s.

Bounds for sums like (3.2) are typically obtained via the spectral theory of automorphic forms [Iwa02; Iwa97], following Deshouillers–Iwaniec [DI82c]; this allows one to bound (3.2) 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 $\lambda \in (0, 1/4)$ of the hyperbolic

Laplacian. This is measured by a parameter $\theta = \max_\lambda \sqrt{\max(0, 1-4\lambda)}$, normalized as in Section 2.2.2; under *Selberg's eigenvalue conjecture* there would be no exceptional eigenvalues [Sel65], so one could take $\theta = 0$. But unconditionally, the record is Kim–Sarnak's bound $\theta \leq 7/32$ [Kim03, Appendix 2], reiterated in Theorem 2.4.

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 [BD20; Mer23], the exponents of distribution of primes [Lic23] and smooth numbers [Pas25c] in arithmetic progressions, and low-lying zeros of Dirichlet L-functions [DPR23]. 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 [DI82c, Theorems 5, 6, 7] and their optimizations in [Dra17; ABL21; Lic23; Pas25c]), 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 (3.2) for completely arbitrary sequences, but only for those arising from variations of Linnik's dispersion method [Lin63; FI85; BFI86; BFI87; BFI89]; these often have the rough form

$$a_m = e(m\alpha) \qquad \text{and} \qquad b_n = \sum_{\substack{h_1, h_2 \sim H \\ h_1\ell_1 - h_2\ell_2 = n}} 1, \tag{3.3}$$

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 chapter 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 [CG11]. The resulting improved bounds for (3.2) can then feed through to the strongest results on several well-studied arithmetic problems.

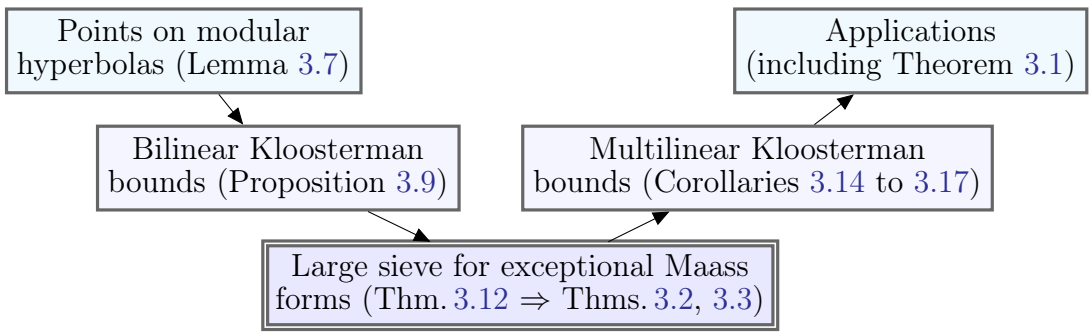

Figure 3.1: Structure of chapter (*arrows signify logical implications*).

Figure 3.1 summarizes the results outlined above, which go from "*counting problems*" (on the top row), to *exponential sums* (middle row), to *automorphic forms* (bot-

tom 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 [Kuz80; DI82c].

Before we dive into the large sieve, let us motivate our discussion with an application.

Theorem 3.1. *For infinitely many $n \in \mathbb{Z}_+$, one has $P^+(n^2+1) > n^{1.3}$.*

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 [Hoo67] proved the same result with an exponent of 1.1001, using the Weil bound for Kloosterman sums. In 1982, Deshouillers–Iwaniec [DI82a] used their bounds on multilinear forms of Kloosterman sums [DI82c] to improve this substantially, up to an exponent of 1.2024. More recently, using Kim–Sarnak's bound $\theta \leq 7/32$ [Kim03, Appendix 2], de la Bretèche and Drappeau [BD20] optimized the exponent to 1.2182. Finally, Merikoski [Mer23] proved a new bilinear estimate (still relying on the bounds of Deshouillers–Iwaniec [DI82c]), 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 3.2 and 3.3), we can improve the arithmetic information due to both Merikoski [Mer23] and de la Bretèche–Drappeau [BD20], leading to the unconditional result in Theorem 3.1. As in [Mer23; BD20], by adapting our proof, it should be possible to obtain similar results for other irreducible quadratic polynomials.

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 [BD20; Har08; Har24] (by improving de la Bretèche–Drappeau's [BD20, Théorème 5.2]), and to enlarging the Fourier support in one-level density estimates for Dirichlet L-functions [DPR23].

3.1.1 The large sieve inequalities

We now turn to our main technical results. The sums of Kloosterman sums from (3.2) are related to the Fourier coefficients of GL_2 automorphic forms of *level* $q = rs$ by the Kuznetsov trace formula [Kuz80; DI82c] for the congruence group $\Gamma_0(q)$.

More precisely, the spectral side of the Kuznetsov formula (see Proposition 2.6) contains three terms, corresponding to the contribution of holomorphic cusp forms, Maass cusp forms, and Eisenstein series. The (conjecturally inexistent) *exceptional* Maass forms, which have Laplacian eigenvalues $\lambda_j \in (0, 1/4)$, typically produce losses of the form X^{θ_j}, where X is a large parameter and $\theta_j = \sqrt{1-4\lambda_j}$. The aforementioned large sieve inequalities for exceptional Maass forms can help alleviate this loss, by

incorporating factors of X^{θ_j}. We recall that Proposition 2.10, which follows from the work of Deshouillers–Iwaniec [DI82c], handles a value of

$$X \ll \max\left(1, \frac{q}{N}, \frac{q^2}{N^3}\right), \tag{3.4}$$

with no losses in the upper bound. Although it seems difficult to improve the range (3.4) in general (see Section 3.2.1), one can hope to do better for special sequences (a_n); for instance, the last term in (3.4) can be improved if the sequence (a_n) is sparse. In this work, we consider the "dual" setting when (a_n) is sparse in frequency space, i.e., 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 3.12, 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 (3.3) (we also incorporate a scalar a in the Fourier coefficients, but on a first read one should take $a = 1$).

Theorem 3.2 (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 Proposition 2.10 and the choice of scaling matrix in (2.12), the bound*

$$\sum_{\lambda_j < 1/4} X^{\theta_j} \left| \sum_{n \sim N} e(n\alpha)\, \rho_{j\mathfrak{a}}(an) \right|^2 \ll_\varepsilon (qaN)^\varepsilon \left(1 + \frac{aN}{q}\right) N \tag{3.5}$$

holds for all

$$X \ll \frac{\max\left(N, \frac{q}{a}\right)}{\min_{t \in \mathbb{Z}_+} (t + N\|t\alpha\|)}. \tag{3.6}$$

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 $\Phi^{(j)} \ll_j 1$.

We recall that $\|\alpha\|$ denotes the distance from α to 0 inside $\mathbb{R}/\mathbb{Z}$; the fact that the worst ("minor-arc") range covered by (3.6) 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 3.2 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. As detailed in Section 2.2.1, altering the scaling matrix $\sigma_{\mathfrak{a}}$ in bounds like (3.5) is equivalent to altering the phase α; the canonical choice in (2.12) leads to several simplifications in practice.

When $a = 1$, $\mathfrak{a} = \infty$, and α is independent of q, Deshouillers–Iwaniec [DI82c] (see Proposition 2.11) showed that the bound in (3.5) 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 [Wat95] (see Proposition 2.12), which saves roughly $X = Q^2/N^{3/2}$ when a_n is a divisor-type function.

For the second sequences mentioned in (3.3), 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 3.3 (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 $\Phi_i^{(j)} \ll_j 1$ for all $j \geq 0$. Then with the notation of Proposition 2.10 and the choice of scaling matrix in (2.12), if $q \gg L^2$, one has

$$\begin{aligned}\sum_{\lambda_j<1/4} X^{\theta_j}\left|\sum_{n\sim N} a_n\,\rho_{j\mathfrak{a}}(an)\right|^2 &\ll_\varepsilon (qaH)^\varepsilon\left(1+\frac{aN}{q}\right)\\ &\quad\times\left(\|a_n\|_2^2 + \gcd(a,q)N\left(\frac{H}{L}+\frac{H^2}{L^2}\right)\right),\end{aligned} \tag{3.7}$$

whenever

$$X \ll \max\left(1, \frac{q}{aN}\right)\max\left(1, \frac{NH}{(H+L)LM}\right), \qquad M := \min_{\substack{t\in\mathbb{Z}_+\\ i\in\{1,2\}}}\left(t + \frac{N}{L}\|t\alpha_i\|\right). \tag{3.8}$$

Remark. In Theorem 3.3, 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 (3.7) produces no important losses over the regular-spectrum bound of $(qN)^\varepsilon(1 + \frac{aN}{q})\,\|a_n\|_2^2$.

Remark. Some instances of the dispersion method [DPR23; Dra17; ABL21] use coefficients roughly of the shape

$$b_n = \sum_{\substack{h\sim H\\ h(\ell_1-\ell_2)=n}} 1, \tag{3.9}$$

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 (3.3) (treated by Theorem 3.3), one should actually handle this case using Theorem 3.2, 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 (3.5) is dominant, and the range in (3.6) becomes $X \ll \ell_1\ell_2/|\ell_1 - \ell_2|$.

Remark. 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 [DI82c]. However, our work should admit two independent extensions, which are relevant for some applications. The first is handling Maass forms with a nebentypus, following Drappeau [Dra17]; this leads to bounds for sums like (3.2) with c restricted to an arithmetic progression. The second is to consider exceptional Hecke–Maass forms for 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 can either follow Assing–Blomer–Li [ABL21] to 'factor out' a from $\rho_{j\mathfrak{a}}(an)$, and apply Kim–Sarnak's bound at places dividing a [Kim03] 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 is better when a is well-factorable).

3.2 Outline

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.

3.2.1 Large sieve with general sequences

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

$$\sum_{\lambda_j<1/4} X^{\theta_j} \left| \sum_{n\sim N} a_n \, \rho_{j\infty}(n) \right|^2 \leq \left(1 + \frac{N}{q}\right) \|a_n\|_2^2 \tag{3.10}$$

of the large sieve inequality from Proposition 2.10, 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 [DI82c, Theorem 2] when $X = 1$, but we need larger values of X to temper the contribution of exceptional eigenvalues. The Kuznetsov trace formula [Kuz80] in Proposition 2.6, combined with large sieve inequalities for the regular spectrum [DI82c, Theorem 2], essentially reduces the problem to bounding (a smoothed variant of) the sum

$$\sum_{\substack{c\sim NX \\ c\equiv 0 \pmod q}} \frac{1}{c} \sum_{m\sim N} \overline{a_m} \sum_{n\sim N} a_n \, S(m, n; c) \tag{3.11}$$

by the same amount as in the right-hand side of (3.10) – see Lemma 2.9 for a formal statement in this direction. The left-hand side vanishes for $X < q/(2N)$, so we immediately obtain (3.10) for $X \ll q/N$, which is the content of [DI82c, 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

$$\frac{NX}{q}\frac{1}{NX}N\|a_n\|_2^2\sqrt{NX} = \frac{N^{3/2}X^{1/2}}{q}\|a_n\|_2^2. \tag{3.12}$$

This is acceptable in (3.10) provided that $X \leq q^2/N^3$, which completes the range from Proposition 2.10.

Improving the range $X \leq \max(1, q/N, q^2/N^3)$ 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 $(a_m), (b_n)$ [KMS17; KMS20; Ker23; Xi18]. For example, an extension of the work of Kowalski–Michel–Sawin [KMS17] to general moduli should improve Proposition 2.10 when $q \approx N^2$, but even then the final numerical savings would be relatively small.

The other critical case encountered in applications is $q \approx N$, where Proposition 2.10 gives no non-trivial 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 (3.10) is at least $X^{\theta(q)}N^2$, while the right-hand side is $(1+\frac{N}{q})qN$; this limits the most optimistic savings for general sequences at $X = (1+\frac{q}{N})^{1/\theta(q)}$.

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 [Lin63]. Often, such sequences have sparse Fourier transforms, and using Fourier analysis on the corresponding exponential sums leads to a combinatorial problem.

3.2.2 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 (3.11) and Fourier-completing in m, n leads to a variant of the identity

$$\sum_{m\sim N} e(-m\alpha)\sum_{n\sim N} e(n\alpha)\, S(m,n;c) \approx N^2 \sum_{\substack{|x-c\alpha|\leq c/N\\ |y+c\alpha|\leq c/N}} e\left(\frac{N(x+y)}{c}\right)\mathbb{1}_{xy\equiv 1 \pmod c}. \tag{3.13}$$

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

$$\sum_{\substack{|x-c\alpha|\leq X\\ |y+c\alpha|\leq X}} \mathbb{1}_{xy\equiv 1 \pmod c}, \tag{3.14}$$

for $c \sim NX$, which is just a count of points on a modular hyperbola in short intervals (as considered in [CG11]). When $\alpha = 0$, one can directly use the divisor bound to write

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

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

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

(This type of bound was also observed by Shparlinski and Zhang [SZ16].) Overall, we roughly obtain

$$\sum_{\substack{c\sim NX \\ c\equiv 0 \pmod q}} \frac{1}{c} \sum_{m,n\sim N} S(m,n;c) \leq \frac{NX + N^2}{q}, \tag{3.15}$$

which is at most $(1 + \frac{N}{q})N$, as required in (3.10), provided that

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

This gives the best-case range from (3.6) (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 [CG11]. For an arbitrary value of α, a pigeonhole argument (Dirichlet approximation) leads to a bound of the shape

$$\sum_{\substack{c\sim NX \\ c\equiv 0 \pmod q}} \frac{1}{c} \sum_{m\sim N} e(-m\alpha) \sum_{n\sim N} e(n\alpha)\, S(m,n;c) \leq \frac{N^{3/2}X + N^2}{q}, \tag{3.16}$$

and ultimately to the range $X \leq \max(\sqrt{N}, q/\sqrt{N})$, which is the worst (and average) case in (3.6) 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. A consequence of not leveraging the exponential phases in the right-hand side of (3.13) 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 (3.15) with a linear growth in X (as opposed to the square-root growth from (3.12), coming from the Weil bound).

3.2.3 Sequences with frequency concentration

It will probably not come as a surprise that one can extend the preceeding 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 3.2 for exponential phases as a black-box, one can directly obtain a bound like

$$\sum_{\lambda_j<1/4} X^{\theta_j} \left| \sum_{n\sim N} a_n\, \rho_{j\mathfrak{a}}(n) \right|^2 \leq \left(1 + \frac{N}{q}\right) N|\mu|(\mathbb{R}/\mathbb{Z})^2, \tag{3.17}$$

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 Proposition 2.10 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^θ.

The alternative approach is to expand the square in the left-hand side of (3.17), pass to a sum of Kloosterman sums as in (3.11) by Kuznetsov, and only then Fourier-expand (two instances of) the sequence (a_n). Using similar combinatorial ideas as for (3.16), we can then essentially bound

$$\sum_{\substack{c\sim NX \\ c\equiv 0 \pmod q}} \frac{1}{c} \sum_{m\sim N} e(m\alpha) \sum_{n\sim N} e(n\beta)\, S(m,n;c) \leq \frac{N^{5/3}X + N^2}{q}, \tag{3.18}$$

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

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

which is acceptable in (3.10), in particular, whenever $X < N^{1/3}$ and $\sqrt{N}|\mu|(\mathbb{R}/\mathbb{Z}) \leq \sqrt{q/N}\|a_n\|_2$. 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 3.12, which particularizes to Theorems 3.2 and 3.3.

What is perhaps more surprising, though, is that strong-enough frequency concentration (i.e., $\sqrt{N}|\mu|(\mathbb{R}/\mathbb{Z})^2 \leq \sqrt{q/N}\|a_n\|_2$) arises in applications, beyond the case

of exponential sequences. A key observation is that the aforementioned dispersion coefficients

$$a_n = \sum_{\substack{h_1,h_2\sim H\\ h_1\ell_1-h_2\ell_2=n}} 1, \tag{3.19}$$

with $\ell_1 \asymp \ell_2 \asymp L$, come from a convolution of two sequences supported on arithmetic progressions, of the form $\mathbb{1}_{n\equiv 0 \pmod{\ell_i}}\mathbb{1}_{n\sim H\ell_i}$. 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 ($\leq 1+\frac{L}{H}$) of rational points (and thus, in frequency concentration on a set of size $\frac{1}{HL}+\frac{1}{H^2}$); see Lemma 3.10.

3.2.4 Multilinear forms of Kloosterman sums

Consider once again the sums (3.2), in the ranges

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

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

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

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 $X^{\theta(q)}$ 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 $(1+X_0)^{\theta(q)}$, and applying Cauchy–Schwarz, we reach an upper bound of

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

Above, we can choose X_1 and X_2 as the maximal values that can be fully incorporated in large sieve inequalities like (3.10) 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

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

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 3.2, which ultimately saves a factor of $N^{\theta/4}$. Similarly, if (b_n) are of the form in (3.19), where $H \asymp L \asymp \sqrt{N}$, then by Theorem 3.3 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 [DI82c, 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 Proposition 2.11 below) with our savings in the X_2-aspect. Following [DI82c, 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 3.17. Such bounds feed directly into the dispersion method and its applications, as we shall see in Section 3.5.

3.2.5 Structure

Section 3.3 only contains elementary arguments, from counting points on modular hyperbolas in Lemma 3.7 (following Cilleruelo–Garaev [CG11]), to the bilinear Kloosterman bounds in Proposition 3.9 (which may be of independent interest to the reader). In Section 3.4.1, we combine these combinatorial inputs with the Deshouillers–Iwaniec setup [DI82c] to prove a general large sieve inequality in Theorem 3.12, which can be viewed as our main technical result; we then deduce Theorems 3.2 and 3.3 from it. Section 3.4.2 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 3.5 we will use these bounds to prove Theorem 3.1, building on the work of Merikoski [Mer23] and de la Bretèche–Drappeau [BD20].

3.3 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), and use a combinatorial argument inspired by [CG11]; the latter was also used, e.g., in [Ker23].

We highlight the following non-standard notation.

Notation 3.4 (Rational approximation)**.** Given $M, N > 0$, let $T_{M,N} : (\mathbb{R}/\mathbb{Z})^2 \to \mathbb{R}$ denote the function

$$\begin{aligned} T_{M,N}(\alpha,\beta) &:= \min_{t\in\mathbb{Z}_+} (t + M\|\alpha t\| + N\|\beta t\|) \\ (\text{abbreviating } T_N &:= T_{N,N},\ T_N(\alpha) := T_N(\alpha,\alpha)), \end{aligned} \tag{3.20}$$

measuring 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 3.5 (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*

$$T_N(\alpha, \beta \pm \alpha) \asymp T_N(\alpha,\beta). \tag{3.21}$$

Moreover,

$$T_{M,N}(\alpha+\gamma,\beta) \le (1 + M\|\gamma\|)\, T_{M,N}(\alpha,\beta). \tag{3.22}$$

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

$$T_{M,N}\left(\alpha+\gamma,\beta+\delta\right) \asymp T_{M,N}(\alpha,\beta). \tag{3.23}$$

Proof. The first equalities are obvious, and (3.21) 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 (3.22), 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, (3.23) follows immediately from (3.22). □

Lemma 3.6 (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 $N \ge 1/2$, one has

$$T_N(\alpha,\beta) \ll \min\left(\sqrt{N(1 + \|\alpha-\beta\|N)}, N^{2/3}\right). \tag{3.24}$$

Proof. Consider the sequence of points $\{(t\alpha, t\beta)\}_{t \le \lceil A \rceil \lceil B \rceil + 2}$ 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

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

and thus

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

This proves (3.24). □

Lemma 3.7 (Concentration of points on modular hyperbolas, following Cilleruelo—Garaev [CG11])**.** *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*

$$\begin{aligned} &\#\{(x, y) \in (I \cap \mathbb{Z}) \times (J \cap \mathbb{Z}) : xy \equiv \lambda \pmod{c}\} \\ &\ll_\varepsilon c^\varepsilon \left(\frac{XY}{c} T_{\frac{c}{X}, \frac{c}{Y}}(\alpha, \beta) + \gcd(\lambda, c) \right), \end{aligned} \tag{3.25}$$

with $T_{M,N}(\alpha, \beta)$ as in Notation 3.4.

Remark. Lemma 3.7 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; (3.25) essentially recovers this average bound when α and β can be simultaneously approximated by rational numbers with a bounded denominator.

Remark. One can also interpret Lemma 3.7 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 $I \cap \mathbb{Z} = \varnothing$ or $J \cap \mathbb{Z} = \varnothing$, 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

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

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 [CG11, Theorem 1] (and also used, for example, in [Ker23, 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

$$at \equiv a' \pmod{c} \qquad \text{and} \qquad bt \equiv b' \pmod{c}. \tag{3.26}$$

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 $|a - c\alpha| \le X$ and $|b - c\beta| \le Y$, i.e.,

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

So by (3.23), 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 2.1.2. The connection to counting solutions to congruences of the form $xy \equiv 1 \pmod{c}$ comes from the identity

$$\sum_m a_m \sum_n b_n \, S(m, n; c) = \sum_{x,y \pmod{c}} \widehat{a}\left(\frac{x}{c}\right) \widehat{b}\left(\frac{y}{c}\right) \mathbb{1}_{xy \equiv 1 \pmod{c}}, \tag{3.27}$$

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., [May25b, Chapter 6] and [FKM15]), but it turns out to also be a useful input for methods from the spectral theory of automorphic forms.

Proposition 3.8 (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. When $\alpha = \beta = 0$, this recovers a result of Shparlinski and Zhang [SZ16]. A similar argument produces the more general bound

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

for $a, b \in \mathbb{Z} \setminus \{0\}$, $r, s \in \mathbb{Z}$.

Proof. Let $\mathcal{S}$ denote the sum in Proposition 3.8; as in (3.27), we expand $S(am, an; c)$ and swap sums to obtain

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

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 $M\|\alpha + ax/c\|$ into dyadic ranges

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

Proceeding similarly for the sum over n, and writing $a' := \frac{a}{\gcd(a,c)}$, $c' := \frac{c}{\gcd(a,c)}$, we get

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

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 3.7 with $X = c'2^{j+2}M^{-1}$, $Y = c'2^{k+2}N^{-1}$, and $\lambda = a'^2$; since the function $T_{M,N}$ is non-decreasing in M, N, this yields

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

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

Proposition 3.9 (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*

$$\begin{aligned}&\sum_{m\in I} a_m \sum_{n\in J} b_n\, S(am, an; c)\\ &\ll_\varepsilon c^\varepsilon \iint_{(\mathbb{R}/\mathbb{Z})^2} (c\, T_{M,N}(\alpha,\beta) + \gcd(a,c)MN)\, d|\mu|(\alpha)\, d|\nu|(\beta),\end{aligned} \tag{3.28}$$

By (3.24), *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 3.8. □

Remark. Suppose $M = N$ and $a = 1$. By comparison, the pointwise Weil bound would yield a right-hand side in (3.28) of roughly $N\sqrt{c}\, \|a_m\|_2 \|b_n\|_2$, while applying Cauchy–Schwarz after (3.27) gives the bound $c\|a_m\|_2\|b_n\|_2$ (these essentially lead to the ranges in Proposition 2.10). It is a very difficult problem [KMS17; Ker23] 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 3.9 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 (3.28) 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 3.8). 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 3.8 this way).

We will ultimately use Proposition 3.9 for sequences (a_n) of the shape in (3.3), 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 3.3 are more interesting, warranting a separate lemma.

Lemma 3.10 (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 $\Phi_i : (0,\infty) \to \mathbb{C}$ be smooth functions supported in $t \ll 1$, with $\Phi_i^{(j)} \ll_j 1$ 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

$$\widehat{a} \ll H^2, \qquad \widehat{a}(\alpha) \ll_\varepsilon H^{-100} \quad \text{unless} \quad \|\ell_i\alpha - \alpha_i\| \leq H^{\varepsilon-1} \ \forall i \in \{1,2\}. \tag{3.29}$$

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

$$a_n = a(n) = \sum_{m \in \mathbb{Z}} b_1(m)\, b_2(n-m) \qquad \Rightarrow \qquad \widehat{a}(\alpha) = \widehat{b_1}(\alpha) \cdot \widehat{b_2}(\alpha), \tag{3.30}$$

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

$$\widehat{b_i}(\alpha) = \widehat{c_i}(\ell_i\alpha - \alpha_i), \tag{3.31}$$

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 (3.30) and (3.31), we obtain

$$\widehat{a}(\alpha) \ll \begin{cases} H^2, & \max\left(\|\ell_1\alpha - \alpha_1\|, \|\ell_2\alpha - \alpha_2\|\right) \leq H^{\varepsilon-1}, \\ O_\varepsilon(H^{-100}), & \max\left(\|\ell_1\alpha - \alpha_1\|, \|\ell_2\alpha - \alpha_2\|\right) > H^{\varepsilon-1}, \end{cases} \tag{3.32}$$

which proves (3.29). Now suppose that $\max\left(\|\ell_1\alpha-\alpha_1\|,\|\ell_2\alpha-\alpha_2\|\right)\leq H^{\varepsilon-1}$; 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

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

so in particular,

$$\ell_1m_2-\ell_2m_1=\ell_2\alpha_1-\ell_1\alpha_2+O\left(H^{\varepsilon-1}L\right). \tag{3.33}$$

Since $\gcd(\ell_1,\ell_2)=1$, as $m_1,m_2\ll L$ vary, the difference $\ell_1m_2-\ell_2m_1$ 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 (3.33). 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

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

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 (3.32), we conclude that for any $p\geq1$,

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

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

Remark. As in [Shp18], 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.

3.4 Spectral bounds

We now combine the combinatorial arguments from the previous section with techniques from the spectral theory of automorphic forms (inspired by [DI82c]), 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 Section 2.2, especially Section 2.2.3.

3.4.1 Large sieve for exceptional Maass forms

Our generalization of Theorem 3.2 requires the following notation, applied to the Fourier transform of a sequence (a_n).

Notation 3.11 (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 3.4. In general, the bound in (3.24) ensures that

$$\mathcal{I}_N(\mu) \ll \iint_{(\mathbb{R}/\mathbb{Z})^2} \min\left(\sqrt{N(1+\|\alpha-\beta\|N)}, N^{2/3}\right) d|\mu|(\alpha)\, d|\mu|(\beta), \tag{3.34}$$

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

$$|\mu|(\mathbb{R}/\mathbb{Z})^2 \ll \mathcal{I}_N(\mu) \ll N^{2/3}|\mu|(\mathbb{R}/\mathbb{Z})^2. \tag{3.35}$$

We also recall the Fourier-analytic notation from Section 2.1.2.

Theorem 3.12 (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 Proposition 2.10, with $\mu(\mathfrak{a}) = q^{-1}$ and the choice of scaling matrix $\sigma_{\mathfrak{a}}$ in (2.12). Then one has

$$\sum_{\lambda_j<1/4} X^{\theta_j} \left|\sum_{n\sim N} a_n\, \rho_{j\mathfrak{a}}(an)\right|^2 \ll_\varepsilon (qaNX)^{2\varepsilon}\left(1+\frac{aN}{q}\right) A^2, \tag{3.36}$$

[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).

whenever

$$A \gg \|a_n\|_2 + \frac{\sqrt{\gcd(a,q)}N}{\sqrt{q+aN}}|\mu|(\mathbb{R}/\mathbb{Z}), \quad X \ll \max\left(1, \frac{q}{aN}\right)\max\left(1, \frac{A^2}{\mathcal{I}_N(\mu)}\right). \tag{3.37}$$

Remark. Theorem 3.12 obtains a saving over Proposition 2.10 whenever we can take $A \ll (qN)^{o(1)}\|a_n\|_2$ and $X > \max(1, \frac{q}{aN})$. To satisfy (3.37) in this context, assuming $\gcd(a,q) = 1$, we need

$$|\mu|(\mathbb{R}/\mathbb{Z}) \ll (qN)^{o(1)}\frac{\sqrt{q+aN}}{N}\|a_n\|_2 \qquad \text{and} \qquad \mathcal{I}_N(\mu) = o\left(\|a_n\|_2^2\right). \tag{3.38}$$

These should be compared with the lower bound

$$|\mu|(\mathbb{R}/\mathbb{Z}) \gg N^{-1/2}\,\|a_n\|_2, \tag{3.39}$$

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

(1). From (3.39) and the lower bound in (3.35), 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 (3.37) to the best-case scenario $X \ll \max(N, \frac{q}{a})$. This is indeed achieved by Theorem 3.2 when $\alpha = 0$.

(2). When $a \ll 1$ and $q \approx N$, (3.38) 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 (3.19).

(3). Using the upper bound $\mathcal{I}_N(\mu) \ll N^{2/3}|\mu|(\mathbb{R}/\mathbb{Z})^2$ from (3.35) 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 (3.38) 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 3.12. 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 Proposition 2.8, we are immediately done if $X \leq 1$, so assume $X > 1$. Let Φ be a fixed nonnegative smooth function supported in $[2,4]$, with positive integral. Then by Lemma 2.9, it suffices to show that

$$\mathcal{S} := \sum_{c \in \mathcal{C}_{\mathfrak{aa}}} \frac{1}{c} \sum_{m,n \sim N} \overline{a_m}\, a_n\, S_{\mathfrak{aa}}(am, an; c)\, \Phi\left(\frac{a\sqrt{mn}}{c} X\right) \ll_\varepsilon (qaNX)^{2\varepsilon} \left(1 + \frac{aN}{q}\right) A^2, \tag{3.40}$$

in the range (3.37). Since $\mu(\mathfrak{a}) = q^{-1}$, Lemma 2.3 implies that

$$\mathcal{S} = \sum_{\substack{c \in (aNX/4, aNX) \\ c \equiv 0 \pmod q}} \frac{\mathcal{S}(c)}{c}, \tag{3.41}$$

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 $aNX \leq q$, the sum over c is void; so we may assume that $X > \max\left(1, \frac{q}{aN}\right)$, which by (3.37) implies

$$\mathcal{I}_N(\mu) \ll A^2. \tag{3.42}$$

We aim to bound each of the $\asymp aNX/q$ inner sums $\mathcal{S}(c)$ separately, using Proposition 3.9. To this end, we need to separate the variables m, n, c; we can rewrite

$$\mathcal{S}(c) = \sum_{m,n \sim N} \overline{\check{\mu}(m)}\, \check{\mu}(n)\, S(am, an; c)\, \Psi_c\left(\frac{m}{N}, \frac{n}{N}\right), \tag{3.43}$$

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 (3.43) and swapping sums and integrals, we obtain

$$\mathcal{S}(c) = \iint_{\mathbb{R}^2} \widehat{\Psi_c}(t_1, t_2)\, \mathcal{S}(c, t_1, t_2)\, dt_1\, dt_2, \tag{3.44}$$

where

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

Note that translating μ corresponds to multiplying $\check{\mu}(n)$ by exponential factors $e(n\alpha)$, so Proposition 3.9 and a change of variables yield

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

where we recalled that $T_N(\alpha,\beta) = T_N(-\alpha,\beta)$. By (3.22), 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 (3.44) 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 (3.41) we conclude that

$$\mathcal{S} \ll_\varepsilon (aNX)^{2\varepsilon}\left(\frac{aNX}{q}\mathcal{I}_N(\mu) + \frac{\gcd(a,q)N^2}{q}|\mu|(\mathbb{R}/\mathbb{Z})^2\right). \tag{3.45}$$

By the lower bound for A in (3.37), the contribution of the second term is

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

which is acceptable in (3.40). Similarly, the first term in (3.45) is acceptable provided that

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

i.e.,

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

which follows from (3.37) and (3.42). □

In particular, we can now deduce the large sieve inequalities promised in Theorems 3.2 and 3.3.

Proof of Theorem 3.2. 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 (3.37) 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

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

so Theorem 3.12 (i.e., (3.37)) recovers the large sieve range

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

from (3.6). In particular, we can recall from (3.24) 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 3.3. 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 3.10. Thus letting $\varphi : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$ be the Fourier transform of $(a_n)_{n \in \mathbb{Z}}$, and $\mu := \varphi\, d\lambda$ (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 3.10 implies that

$$\varphi \ll H^2, \qquad \varphi(\alpha) \ll_\varepsilon H^{-100} \ \text{ unless } \ \|\ell_i\alpha - \alpha_i\| \le H^{\varepsilon - 1}\ \forall i \in \{1, 2\}, \tag{3.46}$$

and

$$|\mu|(\mathbb{R}/\mathbb{Z}) = \|\varphi\|_{L^1} \ll_\varepsilon H^\varepsilon \left(1 + \frac{H}{L}\right). \tag{3.47}$$

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 (3.46), 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 3.4) 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 (3.47), we conclude that

$$\begin{aligned} \mathcal{I}_N(\mu) = \iint_{(\mathbb{R}/\mathbb{Z})^2} T_N(\alpha,\beta)\, d|\mu|(\alpha)\, d|\mu|(\beta) &\ll_\varepsilon LH^{-90} + H^\varepsilon LM\, |\mu|(\mathbb{R}/\mathbb{Z})^2 \\ &\ll_\varepsilon H^{2\varepsilon} LM \left(1 + \frac{H}{L}\right)^2. \end{aligned} \tag{3.48}$$

We are now in a position to apply Theorem 3.12, 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 (3.47), 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 (3.37) 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 (3.36) 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 (3.48),

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

This proves (3.7). □

3.4.2 Multilinear Kloosterman bounds

In contrast to the "vertical" bilinear averages of Kloosterman sums $S(m, n; c)$ over m, n from Section 3.3 (or from [KMS17; Ker23]), 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 $\theta(q) = \sqrt{\max(0, 1 - 4\lambda_1(q))} \leq 7/32$; we recall that the purpose of large sieve inequalities for the exceptional spectrum, like Theorem 3.12, is to improve the dependency on $\theta(q)$.

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

Assumption 3.13 (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 (2.12), 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*

$$\sum_{\lambda_j < 1/4} X^{\theta_j} \left| \sum_{n\sim N} e\left(\frac{n}{N}\xi\right) a_n \rho_{j\mathfrak{a}}(n) \right|^2 \ll_\varepsilon (qNZ)^\varepsilon \left(1 + \frac{N}{q}\right) A_N^2, \tag{3.49}$$

for all $X \ll \max\left(1, \frac{q}{N}\right) \frac{Y_N}{1+|\xi|^2}$. Here, $\theta_j := \sqrt{1 - 4\lambda_j}$ and $\theta(q) := \max_{\lambda_j < 1/4} \theta_j(q)$.

For example, Proposition 2.10 shows that the tuple $(q, N, 1, (a_n)_{n\sim N}, \|a_n\|_2^2, 1)$ satisfies Assumption 3.13 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 3.2 implies that another suitable choice of parameters is

$$a_n := e(n\alpha), \qquad Y_N := \frac{N}{T_N(\alpha)} \gg \sqrt{N}, \qquad A_N := \sqrt{N}, \tag{3.50}$$

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

(3.22). Likewise, incorporating $\ell_i\xi/N$ into α_i, Theorem 3.3 shows that we can choose

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

$$Y_N := \max\left(1, \frac{NH}{(H+L)L\min_i T_H(\alpha_i)}\right), \qquad A_N := \|a_n\|_2 + \sqrt{N}\sqrt{\frac{H}{L}+\frac{H^2}{L^2}}, \tag{3.51}$$

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 3.13 (and implicitly Theorems 3.2 and 3.3), all arguments in this subsection are fairly standard [DI82c; Dra17; BD20].

Corollary 3.14 (Kloosterman-averaging over n, c)**.** *Let $(q, N, Z, (a_n)_{n\sim N}, A_N, Y_N)$ satisfy Assumption 3.13. 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 (2.12). 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*

$$\sum_{n\sim N} a_n \sum_{c\in\mathcal{C}_{\mathfrak{a}\mathfrak{b}}} \Phi\left(\frac{n}{N},\frac{c}{C}\right) S_{\mathfrak{a}\mathfrak{b}}(m,\pm n;c) \ll_\varepsilon (qmNCZ)^{O(\varepsilon)}(1+T)^{\theta(q)}\frac{C^2A_N}{C+\sqrt{mN}} \times \left(1+\frac{mN}{C^2}+\frac{\sqrt{(q,m)m}}{q}\right)^{1/2}\left(1+\frac{mN}{C^2}+\frac{N}{q}\right)^{1/2}, \tag{3.52}$$

for

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

Remark. 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 3.14 saves a total factor of $N^{\theta/4}$ over previous bounds (and up to $N^{\theta/2}$ 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 (3.52) are typically $\asymp 1$.

Remark. While the smooth weight in the c variable is necessary here (stemming from Proposition 2.6), 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 (3.52). The same remark applies to the next results.

Proof of Corollary 3.14. Let $\mathcal{S}$ be the sum in (3.52). For $\Psi(x;y) := \sqrt{x}\,\Phi(x,\sqrt{x}/y)$, 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

$$X := \frac{C}{\sqrt{mN}}. \tag{3.53}$$

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

$$\mathcal{S} \ll_\varepsilon Z^{O(\varepsilon)}C\int_{\mathbb{R}}\frac{|S(\xi)|\,d\xi}{1+\xi^4}, \tag{3.54}$$

where

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

The inner sum is in a suitable form to apply the Kuznetsov trace formula from Proposition 2.6. 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 (2.24) if $|r| \leq R$ and

(2.25) otherwise, where $R \geq 1$ will be chosen shortly. Together with Cauchy–Schwarz and the bounds in Proposition 2.7 (in m) and Proposition 2.8 (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

$$\begin{aligned}\mathcal{S}_{\mathcal{M},\mathrm{reg}}(\xi) \ll_\varepsilon (qmNC)^{O(\varepsilon)} \frac{1}{1+X^{-1}} \\ \times \left(1 + X^{-2} + \frac{\sqrt{(q,m)m}}{q}\right)^{1/2} \left(1 + X^{-2} + \frac{N}{q}\right)^{1/2} \|a_n\|_2.\end{aligned} \tag{3.55}$$

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

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

Then by (2.23) and Cauchy–Schwarz, we obtain

$$\begin{aligned}&\mathcal{S}_{\mathcal{M},\mathrm{exc}}(\xi) \\ &\ll \frac{1}{1+X^{-1}} \sum_{\lambda_j<1/4} \frac{1+X^{\theta_j}}{\cosh(\pi\kappa_j)} |\rho_{j\mathfrak{a}}(m)| \left|\sum_{n\sim N} e\left(\frac{n}{N}\xi\right) a_n \rho_{j\mathfrak{b}}(n)\right| \\ &\ll \frac{(1+X_0)^{\theta(q)}}{1+X^{-1}} \left(\sum_{\lambda_j<1/4} X_1^{\theta_j} |\rho_{j\mathfrak{a}}(m)|^2\right)^{1/2} \left(\sum_{\lambda_j<1/4} X_2^{\theta_j} \left|\sum_{n\sim N} e\left(\frac{n}{N}\xi\right) a_n \rho_{j\mathfrak{b}}(n)\right|^2\right)^{1/2}.\end{aligned} \tag{3.56}$$

We pick X_1 and X_2 as large as (2.32) and Assumption 3.13 allow, specifically

$$X_1 := \max\left(1, \frac{q^2}{(q,m)m}\right), \qquad X_2(\xi) := \max\left(1, \frac{q}{N}\right) \frac{Y_N}{1+|\xi|^2}. \tag{3.57}$$

Then by Proposition 2.7 and Assumption 3.13, we obtain

$$\begin{aligned}\mathcal{S}_{\mathcal{M},\mathrm{exc}}(\xi) \ll_\varepsilon (qmNC)^{O(\varepsilon)} \left(1 + \frac{X}{\sqrt{X_1X_2(\xi)}}\right)^{\theta(q)} \frac{1}{1+X^{-1}} \\ \times \left(1 + \frac{\sqrt{(q,m)m}}{q}\right)^{1/2} \left(1 + \frac{N}{q}\right)^{1/2} A_N.\end{aligned} \tag{3.58}$$

Putting together (3.55) (and the identical bounds for Eisenstein series and holomorphic cusp forms) with (3.58) and (3.54), while noting that $\|a_n\|_2 \ll A_N$ by Assumption 3.13, we conclude that

$$\begin{aligned}\mathcal{S} \ll_\varepsilon (qmNCZ)^{O(\varepsilon)} \left(1+\frac{X}{\sqrt{X_1X_2(0)}}\right)^{\theta(q)} \frac{C}{1+X^{-1}} \\ \times \left(1+X^{-2}+\frac{\sqrt{(q,m)m}}{q}\right)^{1/2}\left(1+X^{-2}+\frac{N}{q}\right)^{1/2} A_N,\end{aligned} \tag{3.59}$$

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 (3.53) and (3.57). □

Remark. In treating the regular spectrum, we picked a slightly sub-optimal value of R (following [DI82c, p. 268]), to simplify the final bounds; in practice, this does not usually matter since one has $X \gg 1$.

Corollary 3.15 (Kloosterman-avg. 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 3.13. 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 (2.12). 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*

$$\begin{aligned}\sum_{m\sim M} a_m \sum_{n\sim N} b_n \sum_{c\in\mathcal{C}_{\mathfrak{a}\mathfrak{b}}} \Phi\left(\frac{m}{M}, \frac{n}{N}, \frac{c}{C}\right) S_{\mathfrak{a}\mathfrak{b}}(m, \pm n; c) \ll_\varepsilon (qMNCZ)^{O(\varepsilon)} (1+T)^{\theta(q)} \\ \times \frac{C^2 A_M A_N}{C+\sqrt{MN}} \left(1+\frac{MN}{C^2}+\frac{M}{q}\right)^{1/2}\left(1+\frac{MN}{C^2}+\frac{N}{q}\right)^{1/2},\end{aligned} \tag{3.60}$$

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

$$\begin{aligned}\sum_{m\sim M} a_m \sum_{n\sim N} b_n \sum_{(c,r)=1} \Phi\left(\frac{m}{M}, \frac{n}{N}, \frac{c}{C}\right) S(m\overline{r}, \pm n; sc) \ll_\varepsilon (rsMNCZ)^{O(\varepsilon)} \\ \times \left(1+\frac{C}{\sqrt{rY_MY_N}}\right)^{\theta(q)} A_M A_N \\ \times \frac{\left(s\sqrt{r}C+\sqrt{MN}+\sqrt{sM}C\right)\left(s\sqrt{r}C+\sqrt{MN}+\sqrt{sN}C\right)}{s\sqrt{r}C+\sqrt{MN}}.\end{aligned} \tag{3.61}$$

Remark. Once again, T_0 represents the smallest value of T that one could use prior to this work; see [DI82c, Theorem 9]. When $a_m = e(m\alpha)$ and $b_n = e(n\beta)$, Corollary 3.15 saves a factor of $(MN)^{\theta/4}$ over previous bounds (and up to $(MN)^{\theta/2}$ if α, β are close to rational numbers with small denominators).

Proof of Corollary 3.15. We only mention what changes from the proof of Corollary 3.14. We expand the sum $\mathcal{S}$ in the left-hand side of (3.60) as a double integral in ζ, ξ, using the Fourier inversion formula

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

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

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

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

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

We proceed as before, applying the Kuznetsov formula from Proposition 2.6 to the inner sum, then using the Bessel transform bounds from Lemma 2.5. 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., Proposition 2.8 and Assumption 3.13). The resulting bounds are symmetric in M, N, with

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

Instead of (3.59), we thus obtain

$$\begin{aligned} \mathcal{S} \ll_\varepsilon (qMNCZ)^{O(\varepsilon)} &\left(1 + \frac{X}{\sqrt{X_1(0)X_2(0)}}\right)^{\theta(q)} \frac{C}{1+X^{-1}} \\ &\times \left(1 + X^{-2} + \frac{M}{q}\right)^{1/2} \left(1 + X^{-2} + \frac{N}{q}\right)^{1/2} \|a_m\|_2\, \|b_n\|_2, \end{aligned} \tag{3.62}$$

which recovers (3.60) after plugging in the values of X, X_1, X_2.

Finally, to prove (3.61) for $q = rs$, we pick $\mathfrak{a} = \infty$, and $\mathfrak{b} = 1/s$, keeping the scaling matrices in (2.12), and use (2.15) to rewrite $S(m\overline{r}, n; sc)$ 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 (3.60) recovers (3.61) up to minor rearrangements. □

Corollary 3.16 (Kloosterman-avg. 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 3.13, $w_q \in \mathbb{C}$, $\mathfrak{b}_q$ 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 (2.12) and a consistent choice of the $\pm$ sign, one has*

$$\sum_{q\sim Q} w_q \sum_{m\sim M} e(m\omega) \sum_{n\sim N} a_{n,q} \sum_{c\in\mathcal{C}_{\infty\mathfrak{b}_q}} \Phi_q\left(\frac{m}{M}, \frac{n}{N}, \frac{c}{C}\right) S_{\infty\mathfrak{b}_q}(m, \pm n; c) \ll_\varepsilon (QMNCZ)^{O(\varepsilon)}$$
$$\times (1+T)^{\theta_{\max}} \frac{\sqrt{QM}\|w_q A_{N,q}\|_2 C^2}{C + \sqrt{MN}} \left(1 + \frac{MN}{C^2} + \frac{M}{Q}\right)^{1/2} \left(1 + \frac{MN}{C^2} + \frac{N}{Q}\right)^{1/2}, \tag{3.63}$$

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 3.13. Then one has

$$\sum_{\substack{r\sim R\\ s\sim S\\ (r,s)=1}} w_{r,s} \sum_{\substack{m\sim M\\ n\sim N}} e(m\omega) a_{n,r,s} \sum_{(c,r)=1} \Phi_{r,s}\left(\frac{m}{M}, \frac{n}{N}, \frac{c}{C}\right) S(m\overline{r}, \pm n; sc) \ll_\varepsilon (RSMNCZ)^{O(\varepsilon)}$$
$$\times \left(1 + \frac{C}{R\sqrt{SY_N}}\right)^{\theta_{\max}} \sqrt{RSM}\|w_{r,s}A_{N,r,s}\|_2$$
$$\times \frac{\left(S\sqrt{R}C + \sqrt{MN} + \sqrt{SM}C\right)\left(S\sqrt{R}C + \sqrt{MN} + \sqrt{SN}C\right)}{S\sqrt{R}C + \sqrt{MN}}. \tag{3.64}$$

Remark. 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

(3.64) with the bound

$$\begin{aligned}&\frac{\left(S\sqrt{R}C+\sqrt{MN}+\sqrt{SM}C\right)\left(S\sqrt{R}C+\sqrt{MN}+\sqrt{SN}C\right)}{S\sqrt{R}C+\sqrt{MN}}\\&\ll S\sqrt{R}C+\sqrt{MN}+\sqrt{SM}C+\sqrt{SN}C+\frac{\sqrt{SM}C\sqrt{SN}C}{S\sqrt{R}C}\\&\ll\left(\frac{C^2}{R}(M+RS)(N+RS)+MN\right)^{1/2}.\end{aligned}\tag{3.65}$$

Remark. Corollary 3.16 should be compared with [DI82c, Theorem 11], the relevant saving being $Y_N^{\theta_{\max}/2}$. One can state a similar result, to be compared with [DI82c, 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 [DI82c, Theorem 6] (or rather, its optimization in [Lic23]) instead of [DI82c, Theorem 7].

Proof of Corollary 3.16. We proceed as in the proof of Corollary 3.15, swapping the sum over q with the integral to bound the sum $\mathcal{S}$ in the left-hand side of (3.60) 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

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

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 $\|w_qA_{N,q}\|_1\le\|w_qA_{N,q}\|_2\sqrt{Q}$ instead of A_N). As in (3.56), the contribution of the exceptional spectrum is

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

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 (3.57); but this time we choose

$$X_1:=\max\left(M,\frac{Q^2}{M}\right),\tag{3.66}$$

corresponding to the allowable range in Proposition 2.11. Keeping $|w_q|$ only in the second sum, this yields

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

where

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

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

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 Proposition 2.11 (which allowed the choice of X_1 from (3.66)), leading to an extra factor of $\sqrt{Q}$. Overall, instead of (3.62), we obtain

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

and plugging in the values of X, X_1, X_2 yields (3.63).

To prove (3.64), let $Q := RS$. By the divisor bound, the left-hand side is at most

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

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 (2.12).

Due to Lemma 2.3, after the change of variables $c \leftarrow c/(s\sqrt{r})$, this leaves us with the sum

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

Incorporating 1-bounded coefficients into (w_q) to remove absolute values, the desired bound now follows from (3.63). 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 (3.63). □

As a direct consequence of Corollary 3.16 and standard techniques, we also deduce a result for sums of incomplete Kloosterman sums, improving [DI82c, Theorem 12].

Corollary 3.17 (Incomplete Kloosterman bounds with averaging over r, s, n, c, d)**.** *Let $R, S, N \geq 1/2$, $C, D, Z \gg 1$, $Y_N > 0$, 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 3.13, $w_{r,s} \in \mathbb{C}$, 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 $j, k, \ell \geq 0$. Then with a consistent choice of the $\pm$ sign, one has*

$$\begin{aligned}\sum_{\substack{r\sim R\\ s\sim S\\ (r,s)=1}} w_{r,s} \sum_{n\sim N} a_{n,r,s} \sum_{\substack{c,d\\ (rd,sc)=1}} \Phi_{r,s}\left(\frac{n}{N}, \frac{d}{D}, \frac{c}{C}\right) e\left(\pm n\frac{\overline{rd}}{sc}\right) \\ \times \ll_\varepsilon (RSNCDZ)^{O(\varepsilon)} \|w_{r,s}A_{N,r,s}\|_2 \mathscr{I},\end{aligned} \tag{3.67}$$

where

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

Proof of Corollary 3.17. This follows from Corollary 3.16 (specifically, (3.64)) 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 [DI82c, Theorem 12] follows from [DI82c, Theorem 11] in [DI82c, §9.2]. We note that [DI82c, Theorem 12] has a minor error (replacing D^2NR with D^2NRS^{-1}), which has been corrected in [BFI19]. □

3.5 The greatest prime factor of $n^2 + 1$

Here we use our new inputs from Section 3.4.2 in the computations of Merikoski [Mer23] and de la Bretèche–Drappeau [BD20], in order to prove Theorem 3.1. We begin with a brief informal sketch.

3.5.1 Sketch of the argument

We will ultimately prove a lower bound of the shape

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

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 [Mer23; BD20; DI82a; Hoo67], we use an idea of Chebyshev to estimate the full sum

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

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

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

Following Merikoski [Mer23], 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

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

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 $Q_1Q_2 = Q \leq x^{1.3}$, aiming to win over the trivial bound of x. We can then Fourier-complete the sum over $n \equiv \nu \pmod q$, where ν ranges over the solutions to $\nu^2 \equiv -1 \pmod q$; this results in a smooth variable h of size up to Q/x, and the principal frequency $h = 0$ cancels with the subtracted main term. After a potential Cauchy–Schwarz step (for the Type II estimate), one reparametrizes the solutions to $\nu^2 \equiv -1 \pmod q$ by the Gauss correspondence; this leads to sums of incomplete Kloosterman sums, ultimately amenable to our bounds from Section 3.4.2.

To obtain our Type I information from Proposition 3.20, we improve the computations of de la Bretèche–Drappeau [BD20, §8] (based in turn on Duke–Friedlander–Iwaniec [DFI95]) using our large sieve inequality for exponential phases, Theorem 3.2. This is nearly enough to remove the dependency on Selberg's eigenvalue conjecture in the relevant Type I ranges, as illustrated in Figure 3.2 (left).

For the Type II information, we follow Merikoski's arrangement of exponential sums, which relies on bounding trilinear forms of Kloosterman sums as in (3.2); this argument cannot fully exploit the averaging over the "level" variable r, since both sequences (a_m) and (b_n) depend on r. However, using our large sieve inequalities, we can leverage the fact that (a_m) happen to be exponential-phase sequences as in Theorem 3.2, while (b_n) are roughly of the form in Theorem 3.3. The second maximum inside the X-factor from (3.8), combined with different ways of applying Cauchy–Schwarz (i.e., keeping the sum over h inside or outside), lead to three admissible Type II ranges, all gathered in Proposition 3.21. This is also reflected in the blue polygonal line from Figure 3.2 (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 3.1.

3.5.2 Arithmetic information

We aim to improve the dependency on the θ parameter in the arithmetic information from [Mer23, Propositions 1 and 2]; to do so, we first improve a lemma of de la Bretèche–Drappeau [BD20, Lemme 8.3]. We stress again that we use Deshouillers–Iwaniec's original normalization for the θ parameters, with $\theta_{\max} \leq 7/32$ by Theorem 2.4; this differs from the normalizations of de la Bretèche–Drappeau [BD20] and Merikoski [Mer23] by a factor of 2.

Lemma 3.18 (De la Bretèche–Drappeau-style exponential sums)**.** *Let $\varepsilon > 0$, $M \gg 1$, and $\theta := 7/32$.*

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

$$\begin{aligned} &\sum_{(m,q)=1} f\left(\frac{m}{M}\right) \sum_{\nu^2 \equiv -1 \pmod{mq}} e\left(\frac{h\nu}{mq}\right) \\ &\ll_\varepsilon (qhM)^\varepsilon \left(|h| + \sqrt{qM}\left(1 + (q,h)^{\theta/2} q^{-3\theta/4} M^{\theta/2}\right)\right). \end{aligned} \tag{3.68}$$

(*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*

$$\frac{1}{Q}\sum_{q\sim Q}\left|\frac{1}{H}\sum_{h\sim H} e(th)\sum_{(m,q)=1} f_q\left(\frac{m}{M}\right)\sum_{\nu^2\equiv -1 \pmod{mq}} e\left(\frac{h\nu}{mq}\right)\right| \\ \ll_\varepsilon (QHM)^\varepsilon\left(H+\sqrt{M}\left(1+H^{-\theta}Q^{\theta/4}M^{\theta/2}\right)+\sqrt{\frac{QM}{H}}\left(1+Q^{-3\theta/4}M^{\theta/2}\right)\right). \tag{3.69}$$

Proof. This is a refinement of the first and third bounds in [BD20, Lemme 8.3], winning factors of about $q^{\theta/4}$ via our Corollaries 3.14 and 3.16. We only mention what changes from the proof in [BD20, §8.1], working in the particular case $d = r = 1$, $D = -1$. We note that for $D = -1$, the relevant cusps $\mathfrak{a}$ from [BD20, §8.1] are equivalent to $0/1$, and thus have $\mu(\mathfrak{a}) = q^{-1}$ (which is also why Merikoski's bounds in [Mer23, §3.8] only require such cusps too).

For part (i), we consider the sums of Kloosterman sums from [BD20, (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 3.14 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 3.14 (equivalently, we can bound $\mathscr{M}_N^{\mathrm{exc}}$ in [BD20, (8.40)] using Theorem 3.2), using $a_n = e(n\alpha_q)$, $Y_N = A_N = \sqrt{N}$ (corresponding to (3.50)), $C = q\sqrt{M}$, and $m = |h|$. This yields

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

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 (3.1), 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

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

which gives the second term claimed in the upper bound from (3.68).

Part (ii) follows similarly using Corollary 3.16 (or equivalently, by bounding $\mathscr{M}_N^{\mathrm{exc}}$ in [BD20, §8.1.12] using Theorem 3.2 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 (3.63) yields

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

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

$$\begin{aligned}&\ll_{\eta}(QHM)^{O(\eta)}\sqrt{\frac{M}{H}}\max(Q,H)^{1/2}\left(1+\max(Q,H)^{-\theta}Q^{\theta/4}M^{\theta/2}\right)\\ &\ll(QHM)^{O(\eta)}\sqrt{\frac{M}{H}}\left(H^{1/2}+Q^{1/2}+\left(H^{(1/2)-\theta}+Q^{(1/2)-\theta}\right)Q^{\theta/4}M^{\theta/2}\right),\end{aligned}$$

which gives all but the first term in the upper bound from (3.69). As in [BD20], the first terms of $|h|$ and H from our bounds in (3.68) and (3.69) could be improved via partial summation, but we omit this optimization too since it will not be relevant for our computations. □

Notation 3.19 (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 [Mer23, §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 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 – i.e., for which the sums over $|\mathcal{A}_q|$ have an asymptotic formula. Figure 3.2 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 [Mer23] and de la Bretèche–Drappeau [BD20]; our improvements are Propositions 3.20 and 3.21.

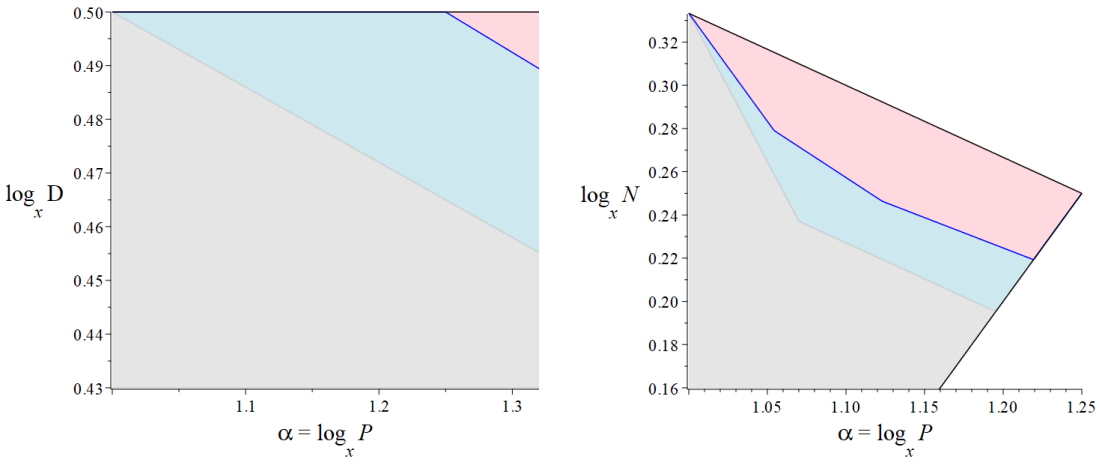

Figure 3.2: Type I (left) and Type II (right) ranges. Previous results in gray; our improvements in blue; conditional ranges in red (assuming Selberg's eigenvalue conjecture).

Proposition 3.20 (Type I estimate)**.** *For any sufficiently small $\varepsilon > 0$ there exists $\delta > 0$ such that the following holds. With Notation 3.19, $1 \leq \alpha \leq 1.4$, $\theta := 7/32$, and $D \geq 1$, one has*

$$\sum_{d \leq D} \lambda_d \sum_{q \equiv 0 \pmod d} \left(|\mathcal{A}_q| - X \frac{\rho(q)}{q} \right) \Psi\left(\frac{q}{P}\right) \log q \ll_\varepsilon x^{1-\delta}, \tag{3.70}$$

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

$$D \ll_\varepsilon x^{-\varepsilon} \min\left(x^{1/2}, x^{2(1-\theta\alpha)/(4-5\theta)}\right).$$

Proof. This is a refinement of Merikoski's [Mer23, Prop. 1] (which explicitated the computations in de la Bretèche–Drappeau's [BD20, §8.4]), using our Lemma 3.18.(*ii*) instead of [BD20, (8.7)]. Indeed, in the first display on [BD20, p. 1620], by applying (3.69) for $H \leftarrow PX^{-1+\delta}$ (for $\delta = \delta(\varepsilon)$ to be chosen shortly), $Q \leftarrow D \leq x^{1/2}$ and $M \leftarrow P/D$, we instead obtain the bound

$$\begin{aligned} R_H(x,P,D) &\ll_\delta x^{1+O(\delta)}P^{-1}DH \\ &\times \left(H + \sqrt{\frac{P}{D}}\left(1 + H^{-\theta}D^{\theta/4}\left(\frac{P}{D}\right)^{\theta/2}\right) + \sqrt{\frac{P}{H}}\left(1 + D^{-3\theta/4}\left(\frac{P}{D}\right)^{\theta/2}\right)\right) \end{aligned}$$

which simplifies to

$$R_H(x,P,D) \ll_\delta x^{O(\delta)}\left(\frac{PD}{x} + \sqrt{PD}\left(1 + x^\theta P^{-\theta/2}D^{-\theta/4}\right) + \sqrt{x}D\left(1 + D^{-5\theta/4}P^{\theta/2}\right)\right).$$

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 [BD20, §8.4] and then [DI82a, §4, 5]. Overall, this bound is acceptable in (3.70) (i.e., $\ll_\varepsilon x^{1-\delta}$) provided that for an absolute constant K, one has

$$\begin{aligned} D &\ll_\varepsilon x^{-K\delta}\min\left(x^2P^{-1}, x^{1/2}, x^{4(1-\theta)/(2-\theta)}P^{-2(1-\theta)/(2-\theta)}, x^{2/(4-5\theta)}P^{-2\theta/(4-5\theta)}\right) \\ &= x^{-\varepsilon}\min\left(x^{2-\alpha}, x^{1/2}, x^{(100-50\alpha)/57}, x^{(64-14\alpha)/93}\right), \end{aligned}$$

where we picked $\delta := \varepsilon/K$ and substituted $\theta = 7/32$, $P = x^\alpha$. A quick numerical verification shows that for $1 \leq \alpha \leq 1.4$, the first and the third term do not contribute to the minimum. □

Proposition 3.21 (Type II estimate)**.** *For any sufficiently small $\varepsilon > 0$ there exists $\delta > 0$ such that the following holds. With Notation 3.19, $\theta := 7/32$, and $MN = P$ with $M, N \geq 1$, one has*

$$\sum_{\substack{m\sim M\\ n\sim N}} a_m b_n \left(|\mathcal{A}_{mn}| - X\frac{\rho(mn)}{mn}\right)\Psi\left(\frac{mn}{P}\right)\log(mn) \ll_\varepsilon x^{1-\delta}, \tag{3.71}$$

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*

$$x^{\alpha-1+\varepsilon} \ll_\varepsilon N \ll_\varepsilon x^{-\varepsilon}\max\left(x^{(2-(1+\theta)\alpha)/(3-2\theta)}, x^{(2-\alpha)(1-\theta)/(3-\theta)}\right); \tag{3.72}$$

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

$$x^{\alpha-1+\varepsilon} \ll_\varepsilon N \ll_\varepsilon x^{(4-3\alpha)/3-\varepsilon}. \tag{3.73}$$

Remark. The upper range in Proposition 3.21.(ii), which completely removes the dependency on Selberg's eigenvalue conjecture, wins over that in Proposition 3.20.(i) only for $\alpha < 136/129 \approx 1.054$. As in [Mer23], 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 3.21.(ii), assuming (i). This is a refinement of Merikoski's result [Mer23, Prop. 4.(ii)], using our Lemma 3.18.(i) instead of Bretèche–Drappeau's bound [BD20, (8.5)].

We briefly recall that in [Mer23, §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 [Mer23, §3.10], we apply our bound (3.68) with $q \leftarrow n_1 n_2$ and $h \leftarrow h(n_1 - n_2)$; with our normalization of θ, this gives the refined bound

$$\Xi_0(M,N) \ll_\delta x^{O(\delta)} \sum_{\substack{n_1,n_2 \sim N \\ (n_1,n_2)=1}} \frac{1}{H} \times \sum_{1 \le |h| \le H} \Big(HN + \sqrt{MN^2}\,\big(1 + (n_1 n_2, h(n_1-n_2))\big)^{\theta/2} N^{-3\theta/2} M^{\theta/2} \Big),$$

which directly leads to

$$\Xi_0(M,N) \ll_\delta x^{O(\delta)} N^2 \left(HN + M^{1/2}N + M^{(1+\theta)/2} N^{(2-3\theta)/2} \right).$$

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

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

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

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

Trivially removing the first term, picking $\delta := \varepsilon/K$, and substituting $P = x^\alpha$, this proves (3.71) in the range

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

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

$$x^{\alpha-1+\varepsilon} \ll_\varepsilon N \ll_\varepsilon \min\left(x^{2(\alpha-1)+\varepsilon}, x^{(4-3\alpha)/3-\varepsilon}\right) \tag{3.74}$$

and

$$\min\left(x^{\alpha-1+\varepsilon}, x^{(4-(3+\theta)\alpha)/(3-4\theta)-\varepsilon}\right) \ll_\varepsilon N \ll_\varepsilon x^{(4-3\alpha)/3-\varepsilon}, \tag{3.75}$$

both of which are (barely) covered by Proposition 3.21.(i). Indeed, for (3.74), a quick numerical verification shows that

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

for $\theta = 7/32$ and all α, the smallest gap being ≈ 0.07, at $a = 10/9$. In (3.75), we have a nontrivial range only when

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

and for such α, we have

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

Thus (3.71) holds in the full range from (3.73). □

Remark. As in [Mer23, §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 3.15 with $a_m = e(m\alpha_q)$ and b_n as in (3.9), but the gain in the H-aspect would be smaller than the loss in the θ-aspect in our computations. This is because Proposition 3.21.(ii) is only relevant for α close to 1, i.e., for small values of H.

Proof of Proposition 3.21.(i). This is a refinement of Merikoski's [Mer23, Prop. 4.(i)], using Corollary 3.15 (plus Theorem 3.3) instead of Deshouillers–Iwaniec's bound [DI82c, Theorem 9].

We very briefly recall the relevant parts of Merikoski's argument and the sizes of the parameters therein, pointing the reader to [Mer23, §3] for details. In [Mer23, §3.4], one expanded and Fourier-completed $|\mathcal{A}_{mn}|$, resulting in a sum over $1 \le |h| \le H$ with

$$H := Px^{-1+\delta}, \tag{3.76}$$

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 [Mer23, p. 1275]; we need to bound these by $\ll_\varepsilon x^{1-\delta}/k$, for $\delta = \delta(\varepsilon)$ to be chosen.

In [Mer23, §3.5], one applied Cauchy–Schwarz keeping the sums over h, n inside, to obtain

$$\Sigma_k(M,N) \ll \left(\frac{M}{k}\right)^{1/2} \Xi_k(M,N)^{1/2}, \tag{3.77}$$

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 3.21.(ii), since then (b_n) was supported on primes; in a first pass the reader can pretend that $k = n_0 = 1$.

In [Mer23, §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 ([Mer23, Lemma 9]), resulting in sums $\Psi_k(R, S)$ of incomplete Kloosterman sums, ranging over r, s of sizes

$$1 \ll R, S \ll \sqrt{\frac{PN}{kn_0}}. \tag{3.78}$$

In [Mer23, §3.7], one completed Kloosterman sums, resulting in a sum over $|t| \leq T$ with

$$T = x^\delta \frac{SdN^2}{Rn_0}, \tag{3.79}$$

and trivially bounded the contribution of $t = 0$. This ultimately leads to the sums of Kloosterman sums $\widetilde{\Psi}_k(R, S)$ from [Mer23, p. 1279], which have a relevant *level* of

$$\varrho := dk^2 n_0 n_1 n_2 \asymp \frac{dN^2}{n_0}. \tag{3.80}$$

Finally, in [Mer23, §3.8], Merikoski used [DI82a, 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 3.15 with $Z = 1$, (c_h) are bounded coefficients,

$$a_m := e\left(-m\frac{\alpha}{\varrho}\right), \qquad b_n := \sum_{\substack{h_1 \sim H_1 \\ h_2 \sim H_2 \\ n = h_1 n_2 - h_2 n_1}} c_{h_1} \overline{c_{h_2}}, \tag{3.81}$$

both of which depend on the level ϱ,

$$\mathcal{M} \ll T, \qquad \mathcal{N} \ll \frac{HN}{kn_0}, \qquad \mathcal{C} \ll S, \tag{3.82}$$

and $1/2 \leq H_2 \leq H_1 \leq H$. 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 [Mer23, p. 1274]) are smooth

functions of h. In fact, expanding $|\mathcal{A}_{mn}| - \frac{\rho(mn)}{mn}$ via Lemma 2.2 and fixing j, u up to a logarithmic loss, we can use the coefficients

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

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. Through Lemma 2.2 we have put $|h|$ in (smooth) dyadic ranges, and then separated 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 (3.81) become

$$b_n := \sum_{\substack{h_1,h_2\in\mathbb{Z}\\ n=h_1n_2-h_2n_1}} c_{h_1}\overline{c_{h_2}},$$

which are in a suitable form to use Theorem 3.3 (see also (3.51)), with $a = 1$, $H = H_j$, $\alpha_i = \pm ux/P \ll x^\delta H^{-1}$, and

$$L := \frac{N}{kn_0} \asymp n_1 \asymp n_2. \tag{3.83}$$

In particular, since $\varrho \ge n_1n_2 \asymp L^2$, the tuple $(\varrho, \mathcal{N}, x, (b_n)_{n\sim\mathcal{N}}, A_\mathcal{N}, Y_\mathcal{N})$ satisfies Assumption 3.13 with

$$Y_\mathcal{N} := \max\left(1, \frac{\mathcal{N}H_j}{(H_j+L)Lx^\delta}\right) \qquad \text{and} \qquad A_\mathcal{N} := \|b_n\mathbb{1}_{n\sim\mathcal{N}}\|_2 + \sqrt{\mathcal{N}}\sqrt{\frac{H_j}{L} + \frac{H_j^2}{L^2}}, \tag{3.84}$$

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 3.2 (see also (3.50)), the tuple $(\varrho, \mathcal{M}, x, (a_m)_{m\sim\mathcal{M}}, A_\mathcal{M}, Y_\mathcal{M})$ satisfies Assumption 3.13, with

$$Y_\mathcal{M} := \sqrt{\mathcal{M}} \qquad \text{and} \qquad A_\mathcal{M} := \sqrt{\mathcal{M}}. \tag{3.85}$$

By Corollary 3.15, specifically (3.61), it follows that

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

and substituting (3.84) and (3.85) gives

$$\begin{aligned}\mathcal{K} \ll_\delta x^{O(\delta)} &\left(1+\frac{\mathcal{C}}{\sqrt{\varrho}\mathcal{M}^{1/4}\max\left(1,\sqrt{\frac{\mathcal{N}H_j}{(H_j+L)L}}\right)}\right)^\theta \\ &\times\sqrt{\mathcal{M}}\left(\|b_n\mathbb{1}_{n\sim\mathcal{N}}\|_2+\sqrt{\mathcal{N}}\sqrt{\frac{H_j}{L}+\frac{H_j^2}{L^2}}\right) \\ &\times\frac{\left(\sqrt{\varrho}\mathcal{C}+\sqrt{\mathcal{M}\mathcal{N}}+\sqrt{\mathcal{M}\mathcal{C}}\right)\left(\sqrt{\varrho}\mathcal{C}+\sqrt{\mathcal{M}\mathcal{N}}+\sqrt{\mathcal{N}\mathcal{C}}\right)}{\sqrt{\varrho}\mathcal{C}+\sqrt{\mathcal{M}\mathcal{N}}}.\end{aligned} \tag{3.86}$$

Since $\mathcal{N}\ll\varrho$ (which follows from $H\ll N$), the term on the second line of (3.86) is at most $\ll\sqrt{\rho}\mathcal{C}+\sqrt{\mathcal{M}\mathcal{C}}+\sqrt{\mathcal{M}\mathcal{N}}$, as in [Mer23, p. 1280]. The resulting bound is non-decreasing in $\mathcal{M},\mathcal{C}$, so we can plug in their upper bounds from (3.82), as well as (3.79) and (3.80) to obtain

$$\begin{aligned}\sum_{d|n_0n_1n_2}\frac{1}{TH^2}\mathcal{K}(d,n_0,n_1,n_2) \ll_\delta x^{O(\delta)}\max_{d\geq1}\frac{Rn_0}{SdN^2H^2}&\left(1+\frac{S\min\left(1,\sqrt{\frac{(H_j+L)L}{\mathcal{N}H_j}}\right)}{\sqrt{\frac{dN^2}{n_0}}\left(\frac{SdN^2}{Rn_0}\right)^{1/4}}\right)^\theta \\ &\times\sqrt{\frac{SdN^2}{Rn_0}}\left(\|b_n\mathbb{1}_{n\sim\mathcal{N}}\|_2+\sqrt{\mathcal{N}}\sqrt{\frac{H_j}{L}+\frac{H_j^2}{L^2}}\right) \\ &\times\left(\sqrt{\frac{dN^2}{n_0}}S+\sqrt{\frac{SdN^2}{Rn_0}}S+\sqrt{\frac{SdN^2}{Rn_0}}\mathcal{N}\right),\end{aligned}$$

where none of the remaining variables have implicit dependencies on d. The right-hand side is seen to be non-increasing 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 [Mer23, bottom of p. 1280] (by [Mer23, 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 non-decreasing 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

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

Summing over n_0 and plugging in the bounds for R, S, L from (3.78) and (3.83), we get

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

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

$$\begin{aligned}\Upsilon_k &\ll_\delta \frac{x^{O(\delta)}}{k} \sum_{n_0 \ll N} \frac{1}{n_0^{2-(\theta/4)}} \left(1 + \frac{\sqrt{PN}}{N^{3/2}} \min\left(1, \frac{\sqrt{N}}{H}\right)\right)^\theta N^2 \left(\frac{\sqrt{PN}}{H} + \sqrt{\frac{N}{H}}\right) \\ &\ll_\delta \frac{x^{O(\delta)}}{k} \left(1 + \min\left(\frac{\sqrt{P}}{N}, \frac{\sqrt{P}}{\sqrt{N} H}\right)\right)^\theta \left(\frac{\sqrt{P} N^{5/2}}{H} + \frac{N^{5/2}}{\sqrt{H}}\right).\end{aligned}$$

Since we have $N \le \sqrt{x} \le \sqrt{P}$ and $\sqrt{N} H \le x^{1/4} P x^{-1+\delta} \le x^{1/4+3/4-1+\delta} \sqrt{P}$ for the ranges in Proposition 3.21.(i), we may ignore the 1-term in the θ-factor; plugging in (3.76), we obtain

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

which improves [Mer23, (3.7)]. In light of (3.77) and $MN = P$, this gives a contribution to $\Sigma_k(M, N)$ of

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

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

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

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

3.5.3 Sieve computations

To complete the proof of Theorem 3.1, it remains to adapt the calculation in [Mer23, §2] with our Type I and Type II information.

Notation 3.22 (Set-up for sieve computations)**.** Further to Notation 3.19, we follow [Mer23, 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 $k \ge 0$. As in [Mer23, p. 1257], we aim to find the greatest ϖ for which

$$\sum_{\substack{x\le P\le x^{\varpi}\\ P=2^j x}} S(x,P) \le (1-\varepsilon)\, X \log x. \tag{3.87}$$

Since $S(x) = X\log x + O(x)$ (see [Mer23, (2.1)]), this will imply the lower bound $P_x \ge x^{\varpi}$.

Following [Mer23, 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 [Mer23, Lemma 1]. Finally, the linear sieve upper bound will require the solutions $F(s), f(s)$ to the delay-differential equation system from [Mer23, p. 1263], while the Harman sieve computations will require the Buchstab function $\omega(u)$, bounded as in [Mer23, (2.5)].

Lemma 3.23 (Linear sieve upper bound)**.** *For any $\varepsilon > 0$ there exists $\delta > 0$ such that the following holds. With $\theta := 7/32$, Notation 3.19, Notation 3.22, and $D := x^{-\varepsilon}\min\left(x^{1/2}, x^{2(1-\theta\alpha)/(4-5\theta)}\right)$, 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 [Mer23, Lemma 2] with the updated parameter D from our Type I information (Proposition 3.20). □

Proposition 3.24 (Asymptotics for Harman sieve sums)**.** *For any $\varepsilon > 0$ there exists $\delta > 0$ such that the following hold. With $\theta := 7/32$, Notation 3.19 and Notation 3.22, let*

$$D := x^{-\varepsilon} \min\left(x^{1/2}, x^{2(1-\theta\alpha)/(4-5\theta)}\right), \qquad U := Dx^{1-\alpha-\varepsilon} =: x^{\xi},$$

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

$$\sigma_0 := \max\left(\frac{2-(1+\theta)\alpha}{3-2\theta}, \frac{(1-\theta)(2-\alpha)}{3-\theta}\right) \tag{3.88}$$

be the exponent from Proposition 3.21.(i).

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

$$\sigma := \max\left(\frac{4-3\alpha}{3}, \sigma_0\right) - \varepsilon, \tag{3.89}$$

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*

$$\gamma := \sigma_0 - (\alpha - 1) - 2\varepsilon, \tag{3.90}$$

one has

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

Proof. These are just [Mer23, Propositions 3 and 4], adapted with our Type II information from Proposition 3.21; the additional term of $(4-3\alpha)/3$ from (3.89) comes from Proposition 3.21.(ii). We note that the proof of [Mer23, Proposition 3] requires

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

which happens for $\alpha < 228/203 - O(\varepsilon)$. Similarly, the proof of [Mer23, Proposition 4] requires

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

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

We are now ready to prove our Theorem 3.1, in a very similar manner to [Mer23, §2.6].

Proof of Theorem 3.1. We follow the Harman sieve computations in [Mer23, §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 3.23 and Proposition 3.24). The five ranges relevant in the proof are now $\alpha < 25/24$, $25/24 \le \alpha < 228/203$, $228/203 \le \alpha < 7/6$, $7/6 \le \alpha < 139/114$, and $\alpha \ge 139/114$. Here, the values $228/203$ and $139/114$ come from Proposition 3.24, 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

$$\xi = \min\left(\frac{1}{2}, \frac{2(1-\theta\alpha)}{4-5\theta}\right) - (\alpha-1) - 2\varepsilon,$$

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

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

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

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

We computed the integrals G_i (for $i \ne 5$) by directly adapting the ranges in Merikoski's Python 3.7 code files (see [Mer23, p. 1268]). In the expression for G_5, we implicitly

used the value $D = x^{1/2-\varepsilon}$ since $\frac{1}{2} < \frac{2(1-\theta\alpha)}{4-5\theta}$ for $\alpha \le 139/114$, and the fact that $1 < 1/(2(\alpha-1)) \le 3$ for $7/6 \le \alpha \le 139/114$. Thus

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

For the remaining range $\alpha \ge 139/114$, we apply Lemma 3.23 to obtain (as in [Mer23, (2.8)])

$$\sum_{\substack{x^{139/114} \le P \le x^{\overline{\omega}} \\ P = 2^j x}} S(x,P) \le \left(4 \int_{139/114}^{1.25} \alpha\, d\alpha + (4-5\theta) \int_{1.25}^{\overline{\omega}} \frac{\alpha}{1-\theta\alpha} d\alpha \right) X \log x,$$

where $\alpha = 1.25 = 5/4$ is the threshold where the expression for D changes (i.e., when $\frac{1}{2} = \frac{2(1-\theta\alpha)}{4-5\theta}$). We conclude that (3.87) holds (for small enough ε) provided that

$$(4-5\theta) \int_{1.25}^{\overline{\omega}} \frac{\alpha}{1-\theta\alpha} d\alpha < 1 - 0.59097 - 4 \int_{139/114}^{1.25} \alpha\, d\alpha,$$

where the right-hand side is at least 0.257406. This inequality (barely) holds true when $\overline{\omega} = 1.30008$, which proves Theorem 3.1. □

Chapter 4

On the exponents of distribution of primes and smooth numbers

4.1 Introduction

Let q be a large positive integer, $a \in \mathbb{Z}$ have $(a, q) = 1$, and $A > 0$. The Siegel–Walfisz theorem gives a pointwise asymptotic for the number of primes up to x which are congruent to a modulo q,

$$\pi(x, q; a) \sim \frac{\pi(x)}{\varphi(q)}, \qquad \text{as } x \to \infty, \text{ for } q \leq (\log x)^A,$$

where $\pi(x) := \#\{\text{prime } p \leq x\}$ and $\pi(x, q; a) := \#\{\text{prime } p \leq x : p \equiv a \pmod q\}$. The small range of moduli $q \leq (\log x)^A$ is an obstruction to many applications, and can be improved substantially to $q \leq x^{1/2}(\log x)^{-B}$ assuming the Generalized Riemann Hypothesis (GRH). Unconditionally, the celebrated Bombieri–Vinogradov theorem [Bom65; Vin65] achieves the same range of moduli in an on-average setting: for $B = B(A)$ large enough in terms of A, one has

$$\sum_{\substack{q \leq x^{1/2}(\log x)^{-B} \\ (q,a)=1}} \left| \pi(x; q, a) - \frac{\pi(x)}{\varphi(q)} \right| \ll_A \frac{x}{(\log x)^A}. \tag{4.1}$$

(In fact, a stronger statement holds true, with a maximum over $a \in (\mathbb{Z}/q\mathbb{Z})^\times$ inside the sum.) This result has been critical to sieve theory methods and their applications, for instance to results on small gaps between primes [May15; Pol14]. Overcoming the square-root barrier at $q < x^{1/2}$, i.e., going "beyond GRH" on average, remains a central open problem in analytic number theory. Elliot–Halberstam [EH68] conjectured that the same estimate holds true in the optimal range of moduli $q \leq x^{1-\varepsilon}$, and Polymath8b [Pol14] showed that a generalization of this conjecture would imply the existence of infinitely many pairs of primes with distance at most 6.

Since the pioneering work of Fouvry [Fou84; Fou87; Fou85; Fou82], Fouvry–Iwaniec [FI80; FI83], and Bombieri–Friedlander–Iwaniec [BFI86; BFI87; BFI89], we have been able to overcome this square-root barrier in special settings [Zha14; May25a; May25b; May25c; Sta25] – in particular, by replacing the absolute values in (4.1) with special weights (λ_q) that arise in sieve theory applications. If such an analogue of (4.1) holds with a weighted sum over all $q \leq x^\vartheta$, for any fixed residue a, then we say that the primes have *exponent of distribution* $\vartheta \in (0,1)$ (or *level of distribution* x^ϑ) with respect to the weights (λ_q).

Motivated by a 'well-factorable' variant of the linear sieve weights [Iwa80; FI83], Bombieri–Friedlander–Iwaniec [BFI86, Theorem 10] considered sequences (λ_q) that can be expressed as a Dirichlet convolution of two sequences of any pre-specified lengths, and achieved an exponent of distribution of $\frac{4}{7} - \varepsilon$ in this setting. More recently, Maynard [May25b] considered the refined setting of 'triply-well-factorable' weights, which we recall from [May25b, Definition 2].

Definition 4.1 (Triply-well-factorable weights [May25b])**.** A complex sequence $(\lambda_q)_{q \leq Q}$ is said to be *triply-well-factorable* of level Q iff for any $Q_1, Q_2, Q_3 \geq 1$ with $Q_1 Q_2 Q_3 = Q$, there exist 1-bounded complex sequences (α_{q_1}), (β_{q_2}), (γ_{q_3}) supported on $q_i \leq Q_i$, such that for all q,

$$\lambda_q = \sum_{q = q_1 q_2 q_3} \alpha_{q_1} \beta_{q_2} \gamma_{q_3}.$$

For such weights, which arise in a slight variant of the β-sieve with $\beta \geq 2$, Maynard [May25b, Theorem 1.1] achieved the exponent of distribution $\frac{3}{5} - \varepsilon$ (i.e., with a level $Q = x^{3/5-\varepsilon}$). His results also implied an improved exponent of $\frac{7}{12} - \varepsilon$ for the well-factorable variant of the upper-bound linear sieve weights (the case $\beta = 1$), which are close to being triply-well-factorable.

Essentially all such results are based on equidistribution estimates for convolutions of sequences in arithmetic progressions, proven using Linnik's dispersion method [Lin63]. Ultimately, these rely on bounding sums of Kloosterman sums via the spectral theory of automorphic forms, following Deshouillers–Iwaniec [DI82c]. In this context, Lichtman [Lic23] used optimized Deshouillers–Iwaniec-style estimates, via Kim–Sarnak's bound [Kim03], to improve the exponent of distribution for triply-well-factorable weights to $\frac{66}{107} - \varepsilon \approx 0.6168$ unconditionally, and up to $\frac{5}{8} - \varepsilon = 0.625 - \varepsilon$ assuming Selberg's eigenvalue conjecture.

Our goal in this work is to completely eliminate the dependency on Selberg's conjecture in several exponent-of-distribution results. For the primes, parts (i) and (ii) of the result below improve on the previous exponents of $\frac{66}{107}$ due to Lichtman [Lic23], respectively $\frac{7}{12}$ due to Maynard [May15].

Theorem 4.2 (Primes in APs to large moduli). *Let $a \in \mathbb{Z} \setminus \{0\}$, $A, \varepsilon > 0$, $x \geq 2$. Assume either:*

(*i*). *$Q \leq x^{5/8-\varepsilon}$, and (λ_q) are triply-well-factorable weights of level Q, or*

(*ii*). *$Q \leq x^{3/5-\varepsilon}$, and (λ_q) are the upper-bound well-factorable linear sieve weights of level Q.*

(See Definition 4.16 for part (ii).) Then one has

$$\sum_{\substack{q \leq Q \\ (q,a)=1}} \lambda_q \left(\pi(x; q, a) - \frac{\pi(x)}{\varphi(q)} \right) \ll_{\varepsilon, A, a} \frac{x}{(\log x)^A}. \tag{4.2}$$

Moreover, in Theorem 4.19, we obtain a similar result applicable to both the upper-bound and the lower-bound well-factorable linear sieve weights, with a variable exponent of distribution depending on the factorization of the modulus q; this refines [Lic23, Proposition 6.6]. As a consequence, we deduce a sharper upper bound for the number of twin primes up to x.

Corollary 4.3 (Count of twin primes). *As $x \to \infty$, one has*

$$\#\{p \leq x : p, p+2 \text{ are prime}\} \leq (3.203 + o(1))\, \Pi_2(x),$$

where $\Pi_2(x) := \frac{2x}{(\log x)^2} \prod_{p>2} \frac{1-2/p}{(1-1/p)^2}$ is the asymptotic predicted by Hardy–Littlewood [HL23].

This improves the constant of 3.229 from [Lic23, Theorem 1.1]; we point the reader to [Lic23, p. 2] for a table of previous results. Once again, the key qualitative feature of Corollary 4.3 is that it cannot be improved directly by assuming Selberg's conjecture.

The analogous equidistribution problem for *smooth* numbers [FT96; Dra15] concerns the quantities

$$\begin{aligned} \Psi(x, y) &:= \#\{n \leq x : P^+(n) \leq y\}, \\ \Psi_q(x, y) &:= \#\{n \leq x : P^+(n) \leq y,\ (n, q) = 1\}, \\ \Psi(x, y; a, q) &:= \#\{n \leq x : P^+(n) \leq y,\ n \equiv a \pmod q\}, \end{aligned} \tag{4.3}$$

where $P^+(n)$ denotes the largest prime factor of n; the key setting is $y \leq x^{1/C}$, with C large. In this context, Granville [Gra93a, Theorem 2] proved a suitable analogue of the Bombieri–Vinogradov theorem, achieving the exponent of distribution $\frac{1}{2} - \varepsilon$ (see also [Wol73a; Wol73b; FT91; Gra93b]). Relying on a triple convolution estimate of Bombieri–Friedlander–Iwaniec [BFI86, Theorem 4], Fouvry–Tenenbaum [FT96] raised the exponent to $\frac{3}{5} - \varepsilon$, with an upper bound of $x(\log x)^{-A}$ as in (4.1). Drappeau later

[Dra15] strengthened the bound back to $\Psi(x,y)(\log x)^{-A}$, with the same exponent of $\frac{3}{5}-\varepsilon$. We remark that all of these results use absolute values (i.e., arbitrary 1-bounded weights λ_q), which is possible beyond the square-root barrier due to the flexible factorization properties of smooth numbers.

Using a different arrangement of exponential sums and optimized Deshouillers–Iwaniec-style estimates, the author [Pas25c] recently showed that smooth numbers have exponent of distribution $\frac{66}{107}-\varepsilon \approx 0.6168$, and up to $\frac{5}{8}-\varepsilon = 0.625-\varepsilon$ assuming Selberg's eigenvalue conjecture. As in the case of primes, we can now fully close the gap between the conditional and unconditional results.

Theorem 4.4 (Smooth numbers in APs to large moduli)**.** *Let $a \in \mathbb{Z}\setminus\{0\}$ and $A,\varepsilon>0$, $x\geq 2$. Then there exists a large enough $C=C(a,A,\varepsilon)>0$ such that for any $y\in[(\log x)^C, x^{1/C}]$ and $Q\leq x^{5/8-\varepsilon}$, one has*

$$\sum_{\substack{q\leq Q\\(q,a)=1}}\left|\Psi(x,y;a,q)-\frac{\Psi_q(x,y)}{\varphi(q)}\right| \ll_{\varepsilon,A,a} \frac{\Psi(x,y)}{(\log x)^A}.$$

Remark. Following Drappeau–Granville–Shao [DGS17], one can deduce a similar result for smooth-supported multiplicative functions in arithmetic progressions, using a slight extension of our triple convolution estimate (Proposition 4.27).

Moreover, in Theorem 4.29, we prove a similar result with a slightly-better saving when the sum over q is supported on smooth moduli; this refines a result of de la Bretèche–Drappeau [BD20, (2.1)]. As a consequence, we improve the exponent of $\frac{3}{5}$ in [BD20, Théorème 4.1] to $\frac{5}{8}$ in Corollary 4.30, which includes the following upper bound for the number of consecutive smooth numbers up to x.

Corollary 4.5 (Count of consecutive smooth numbers)**.** *For any $\varepsilon>0$ there exists $C>0$ such that for any $x\geq 2$ and $y\in[(\log x)^C, x^{1/C}]$, one has*

$$\#\{n\leq x : P^+(n), P^+(n+1)\leq y\} \ll_\varepsilon x\varrho(u)^{1+5/8-\varepsilon},$$

where $u:=(\log x)/\log y$ and ϱ denotes the Dickman function [Hil86].

We note that $\frac{5}{8}-\varepsilon$ is now the best exponent of distribution for both primes and smooth numbers, in essentially any setting relevant for sieve theory. In particular, there does not appear to be a slightly more flexible setting which allows for a better exponent with current methods (e.g., primes with quadruply-well-factorable weights, or smooth numbers with well-factorable weights).

Our improvements stem mainly from a large sieve inequality for exceptional Maass forms from Chapter 3, combined, in the case of primes, with a large sieve inequality of Watt [Wat95, Theorem 2]. These results act as on-average substitutes for Selberg's

eigenvalue conjecture, by improving the dependency on X in bounds for sums of the shape

$$\sum_f X^{\theta_f} \left| \sum_{n \sim N} a_n \, \rho_f(n) \right|^2, \tag{4.4}$$

where f ranges over certain families of automorphic forms, with Fourier coefficients $\rho_f(n)$ and spectral parameters $\theta_f \in [0, \frac{7}{32}]$; here $\theta_f > 0$ only when f fails Selberg's conjecture, and the uniform bound of $\frac{7}{32}$ is due to Kim–Sarnak [Kim03, Appendix 2] (see Theorem 2.4). Importantly, both aforementioned large sieve inequalities use special sequences (a_n) that arise in applications, roughly of the form

$$a_n := \sum_{h,h' \sim H} \mathbb{1}_{h\ell - h'\ell' = n}, \qquad \text{respectively} \qquad a_n := \sum_{\substack{h \sim H \\ k \sim K}} \mathbb{1}_{hk = n}. \tag{4.5}$$

The first sequence in (4.5) comes from an additive convolution, and its additive structure is manifested through a sparse Fourier transform, which was crucial for the large sieve inequalities in Chapter 3. By contrast, the second sequence above comes from a multiplicative convolution, and Watt's argument [Wat95, Section 2] crucially relied on its multiplicative structure. Both arguments use the smoothness of the variables h, h', k, which come from Fourier completion.

Remark. For some applications, it would be interesting to obtain improved large sieve inequalities for sequences which display a mix of additive and multiplicative structure, such as

$$a_n := \sum_{\substack{h,h' \sim H \\ k,k' \sim K}} \mathbb{1}_{hk\ell - h'k'\ell' = n}.$$

In particular, this would be relevant for improving the total length in a mean-value estimate for the squared zeta function times a product of two Dirichlet polynomials, due to Deshouillers–Iwaniec [DI84, Theorem 2] (and refined to an asymptotic by Bettin–Chandee–Radziwill [BCR17]). For an optimal choice of unbalanced ranges $M > N$, assuming Selberg's eigenvalue conjecture, one should reach the threshold $MN \leq T^{5/8}$; the presence of the exponent $\frac{5}{8}$ here is not a coincidence, since these results rely on bounds for exponential sums of the shape in (4.10).

Remark. Although we will focus on equidistribution results with fixed (or small) residues a, similar results are possible in the range $a \ll x^{1+\varepsilon}$; this is relevant, e.g., to upper-bounding counts of Goldbach representations [Lic23]. However, working with large values of a ultimately has the effect of replacing some of the dependency on (progress towards) Selberg's eigenvalue conjecture with its non-Archimedean counterpart, the Ramanujan–Petersson conjecture at primes dividing a. In [Lic23], Lichtman incorporates technology of Assing–Blomer–Li [ABL21] to explicitate the dependency on the Ramanujan–Petersson conjecture; with this approach, the final exponent of

distribution decreases with a. However, using appropriate non-Archimedean analogues of the large sieve inequalities from Chapter 3 and [Wat95], one should be able to match the exponent of distribution $\frac{5}{8} - \varepsilon$ in a larger range of a, and in the full range $a \ll x^{1+\varepsilon}$ if a is well-factorable.

4.2 Outline

Our proofs of Theorems 4.2 and 4.4 build on the arguments of Maynard [May25b] and Lichtman [Lic23], respectively Drappeau [Dra15] and the author [Pas25c], with new inputs in the exceptional automorphic spectrum. Here we give an informal outline of our arguments, ignoring various technical details such as smooth weights, common divisors, and some $x^{o(1)}$ factors.

4.2.1 Reduction to sums of Kloosterman fractions.

Let $Q \in (x^{1/2+o(1)}, x^{5/8-o(1)})$, and fix the residue $a = 1$ for simplicity. In the critical ranges, Theorem 4.2.(i), respectively Theorems 4.4 and 4.29, rely on bounding sums of the form

$$\sum_{q_1 \sim Q_1} \lambda_{q_1} \sum_{q_2 \sim Q_2} \mu_{q_2} \sum_{q_3 \sim Q_3} \nu_{q_3} \sum_{n \sim N} \alpha_n \sum_{m \sim x/N} \beta_m \left(\mathbb{1}_{mn \equiv 1 \pmod{q_1 q_2 q_3}} - \frac{\mathbb{1}_{(mn, q_1 q_2 q_3)=1}}{\varphi(q_1 q_2 q_3)} \right), \tag{4.6}$$

respectively

$$\sum_{q \sim Q} \lambda_q \sum_{n_1 \sim N_1} \alpha_{n_1} \sum_{n_2 \sim N_2} \beta_{n_2} \sum_{n_3 \sim N_3} \gamma_{n_3} \left(\mathbb{1}_{n_1 n_2 n_3 \equiv 1 \pmod{q}} - \frac{\mathbb{1}_{(n_1 n_2 n_3, q)=1}}{\varphi(q)} \right), \tag{4.7}$$

for certain ranges of Q_i, Q, N_i, N with $\prod Q_i \asymp Q$ and $\prod N_i \asymp x$, and for arbitrary divisor-bounded coefficients $(\lambda_q), (\mu_q), (\alpha_n), (\beta_n), (\gamma_n)$. The goal in both cases is to beat the trivial bound of size about x, while making Q as large as possible.

In (4.6) (for primes, with triply-well-factorable weights in the modulus), we are essentially free to factorize $Q = \prod Q_i$ as we wish in terms of x and N, and we will roughly choose

$$Q_1 \approx N, \qquad Q_2 \approx \frac{Q^2}{x}, \qquad Q_3 \approx \frac{x}{QN}. \tag{4.8}$$

Similarly, in (4.7) (for smooth numbers, with arbitrary weights in the modulus), we are free to factorize $N = \prod N_i$ as we wish in terms of x and Q, and we will roughly choose

$$N_1 \approx \frac{x}{Q}, \qquad N_2 \approx \frac{Q^2}{x}, \qquad N_3 \approx \frac{x}{Q}. \tag{4.9}$$

There is a certain duality between the two problems, partly due to the correspondence of sizes $Q_2 \approx N_2$, $NQ_3 \approx N_3$; indeed, the two convolution estimates above reduce to bounding the same sum of Kloosterman fractions, given below in (4.10). This is why the final exponents of distribution are the same – both in previous works [May25b; Dra15], [Lic23; Pas25c], and in our Theorems 4.2 and 4.4. Both proofs rely on Linnik's dispersion method [Lin63; BFI86; BFI87; BFI89], which begins with an application of Cauchy–Schwarz in q_1, m, respectively q, n_1; expanding the square will duplicate the other variables. The main dispersion sums will contain smooth sums over q_1, m, respectively q, n_1, as well as congruences

$$\begin{cases} n \equiv n' \pmod{q_1}, \\ m \equiv \overline{n} \pmod{q_1 q_2 q_3}, \\ m \equiv \overline{n'} \pmod{q_2' q_3'}, \end{cases} \qquad \text{respectively} \qquad \begin{cases} n_2 n_3 \equiv n_2' n_3' \pmod{q}, \\ n_1 \equiv \overline{n_2 n_3} \pmod{q}. \end{cases}$$

One can Fourier-complete the sums in m mod $q_1 q_2 q_3 q_2' q_3'$, respectively n_1 mod q, which introduces a smooth variable h of size

$$|h| \leq \frac{Q_1 (Q_2 Q_3)^2}{x/N} \approx \frac{Q^2}{x}, \qquad \text{respectively} \qquad |h| \leq \frac{Q}{N_1} \approx \frac{Q^2}{x},$$

and the contribution of the principal frequency at $h = 0$ simplifies with other main terms. Moreover, one can pass from q_1, respectively q, to the complementary divisors $\frac{n-n'}{q_1}$, respectively $\frac{n_2 n_3 - n_2' n_3'}{q}$, which have size $\asymp 1$ (so we can ignore them for the moment). In the critical ranges, it essentially remains to bound

$$\sum_{\substack{c \sim Q \\ d \sim x/Q}} \left| \sum_{\ell \sim Q^2/x} v_\ell \sum_{h \sim Q^2/x} e\left(h \frac{\overline{\ell d}}{c} \right) \right| < Q^2, \tag{4.10}$$

where the (c, d, ℓ) variables correspond to $(nq_2'q_3', n'q_3, q_2)$, respectively $(n_2'n_3', n_3, n_2)$, and (v_ℓ) are divisor-bounded coefficients. Note that we need to save a factor of roughly Q^2/x over the trivial bound, corresponding to the loss from Fourier completion.

4.2.2 Reaching the exceptional spectrum

We may now forget about the original structure from (4.6) and (4.7), and focus on the exponential sum in (4.10). After applying Cauchy–Schwarz once again and swapping sums, one is left with proving that

$$\sum_{\ell, \ell' \sim Q^2/x} v_{\ell'} \overline{v}_\ell \sum_{h, h' \sim Q^2/x} \sum_{\substack{c \sim Q \\ d \sim x/Q}} e\left((h\ell - h'\ell') \frac{\overline{\ell \ell' d}}{c} \right) < \frac{Q^4}{x}. \tag{4.11}$$

The diagonal terms with $h\ell = h'\ell'$ are barely acceptable. In the off-diagonal terms, denoting $r := \ell\ell'$, $n := h\ell - h'\ell'$, and[1]

$$a_{n,r} := \sum_{h,h' \sim Q^2/x} \mathbb{1}_{h\ell - h'\ell' = n}, \tag{4.12}$$

and completing Kloosterman sums (passing from d to a variable m of dual size $Q(x/Q)^{-1}$), it remains to show that

$$\sum_{r \sim Q^4/x^2} \left| \sum_{m \sim Q^2/x} \sum_{n \sim Q^4/x^2} a_{n,r} \sum_{c \sim Q} S(m\overline{r}, n; c) \right| < \frac{Q^6}{x^2}. \tag{4.13}$$

At this point, applying the Kuznetsov trace formula [Kuz80; DI82c] to the inner sum over c brings in the Fourier coefficients $\rho_f(m)$, $\rho_f(n)$ of automorphic forms f for the congruence subgroup $\Gamma_0(r)$. The contribution of Maass cusp forms is the most difficult to bound; after moving the sums over m, n inside, we are left to bound

$$\sum_{r \sim Q^4/x^2} \sum_{f}^{\Gamma_0(r)} \sqrt{x}^{\theta_f} \left| \sum_{m \sim Q^2/x} \rho_f(m) \sum_{n \sim Q^4/x^2} a_{n,r} \overline{\rho}_f(n) \right| < \frac{Q^3}{x}, \tag{4.14}$$

where $\theta_f \in [0, 7/32]$ measures the failure of Selberg's eigenvalue conjecture as in (4.4), and $\rho_f(m)$, $\rho_f(n)$ have size about $r^{-1/2}$ on average. Following Deshouillers–Iwaniec [DI82c], one can now apply Cauchy–Schwarz a third time, so that it remains to bound

$$\begin{aligned} &\left(\sum_{r \sim Q^4/x^2} \sum_{f}^{\Gamma_0(r)} \left(\frac{Q^6}{x^3} \right)^{\theta_f} \left| \sum_{m \sim Q^2/x} \rho_f(m) \right|^2 \right) \times \\ &\left(\sum_{r \sim Q^4/x^2} \sum_{f}^{\Gamma_0(r)} \left(\frac{x^4}{Q^6} \right)^{\theta_f} \left| \sum_{n \sim Q^4/x^2} a_{n,r} \overline{\rho}_f(n) \right|^2 \right) < \frac{Q^6}{x^2}. \end{aligned} \tag{4.15}$$

The reason for this arrangement is that the large sieve inequalities from [DI82c] obtain square-root cancellation in the sums over m, n. Moreover, in the exceptional spectrum where $\theta_f > 0$, [DI82c, Theorem 7] can incorporate the factor of $(Q^6/x^3)^{\theta_f}$ from the first sum with no losses.

However, for the exceptional factor of $(x^4/Q^6)^{\theta_f}$ in the second sum, all previous works [May25b; Lic23; Dra15; Pas25c] essentially use an L^∞ bound. Denoting $\theta := \max \theta_f$ and using the aforementioned spectral large sieve inequalities, this would leave us with

$$\frac{Q^4}{x^2} \frac{Q^2}{x} \cdot \left(\frac{x^4}{Q^6} \right)^{\theta} \frac{Q^4}{x^2} \frac{Q^4}{x^2} < \frac{Q^6}{x^2} \qquad \Longleftrightarrow \qquad Q < x^{\frac{5-4\theta}{8-6\theta}},$$

[1] Really, $a_{n,r}$ depends on n, ℓ, ℓ' since there may be more factorizations $r = \ell\ell'$, but let us ignore this here.

where we really need a power-saving in the final bound. Plugging in Selberg's bound $\theta \leq 1/2$ (which was the state of the art in [DI82a]), one reaches the exponent of distribution $\frac{3}{5} - o(1)$ from the works of Maynard [May25b] and Drappeau [Dra15]. Using the celebrated bound $\theta \leq \frac{7}{32}$ of Kim–Sarnak [Kim03, Appendix 2], one reaches the exponent $\frac{66}{107} - o(1)$ from the works of Lichtman [Lic23] and the author [Pas25c]. Conditionally on Selberg's conjecture that $\theta = 0$, the resulting exponent is $\frac{5}{8} - o(1)$.

4.2.3 Our improvements

Naturally, one can hope to win more in the exceptional spectrum using a suitable large sieve inequality for the second sum in (4.15); but until very recently, it was impossible to obtain *any* savings in the θ-aspect for sequences like $(a_{n,r})$ which depend on the level r, when n and r have the same size. This is now possible using our work from Chapter 3, provided that the sequence $(a_{n,r})$ has enough additive structure. Indeed, for the shape of $a_{n,r}$ from (4.12), which matches the left-hand side of (4.5), our Theorem 3.3 saves an additional factor of $(Q^2/x)^{\theta_f}$ in (4.15). This leads to a final bound of

$$\frac{Q^4}{x^2}\frac{Q^2}{x} \cdot \left(\frac{x^5}{Q^8}\right)^{\theta} \frac{Q^4}{x^2}\frac{Q^4}{x^2} < \frac{Q^6}{x^2} \qquad \Longleftrightarrow \qquad Q < x^{\frac{5-5\theta}{8-8\theta}} = x^{5/8},$$

and thus to the unconditional exponent of distribution of $\frac{5}{8} - o(1)$.

This concludes the outline of our results on smooth numbers from Theorems 4.4 and 4.29, up to various technical details. However, the case of primes from Theorem 4.2.(i) presents a significant additional challenge: the triply-well-factorable condition from Definition 4.1 can only really guarantee that $Q_1 \leq N$, $Q_2 \leq \frac{Q^2}{x}$ and $Q_3 \leq \frac{x}{QN}$, as opposed to the double-sided bounds implied in (4.8). The potential gap between Q_1 and N creates a large complementary-divisor factor

$$f := \frac{n - n'}{q_1} \ll F := \frac{N}{Q_1},$$

which ultimately alters the shape of the coefficients $(a_{n,r})$ from (4.12) to

$$a_{n,r} \approx \sum_{f \sim F} \sum_{h,h' \sim Q^2/(xF)} \mathbb{1}_{f(h\ell - h'\ell') = n}.$$

This sequence displays a mix of additive and multiplicative structure, and we do not know how to prove a corresponding large sieve inequality in the exceptional spectrum, generalizing Theorem 3.3 with a good dependency on F. This is a significant issue, since the previous argument could only barely reach the unconditional exponent of $\frac{5}{8} - o(1)$.

We overcome this issue by moving f-variable to the other entry of the Kloosterman sums, by a variant of the identity

$$S(m\overline{r}, fn; c) = S(fm\overline{r}, n; c),$$

which holds when $(f, c) = 1$; working around the latter coprimality constraint is a nontrivial argument in itself, within the proof of Lemma 4.12. In the n-aspect from (4.15), this leaves us with coefficients $(a_{n,r})$ as in the left-hand side of (4.5), which can be handled by Theorem 3.3. In the m-aspect from (4.15), we are left with a multiplicative convolution of two smooth sequences, as in the right-hand side of (4.5). For such sequences, Watt's large sieve inequality [Wat95, Theorem 2], incorporated into our Proposition 4.10, produces nearly-optimal savings when an average over the level r is available. The final dependency of the resulting bounds on F is acceptable, partly because the m-variable is much smaller than the level (so there is enough 'room' for the f-variable).

For Theorem 4.2.(ii) and Theorem 4.19, we mention that Iwaniec's well-factorable linear sieve weights are not very far from being triply-well-factorable – in fact, such results still depend on bounding the sum in (4.6), but with less freedom in choosing the parameters Q_1, Q_2, Q_3. This lower degree of flexibility leads to fairly complicated (but purely elementary) combinatorial optimization problems, which we treat in Section 4.5. Once again, the final levels of distribution match the best conditional results (that one would obtain by assuming Selberg's eigenvalue conjecture in our proofs).

4.2.4 Structure

In Section 4.3, we rely on works of the author (Chapter 3) and Watt [Wat95] to deduce new bounds for multilinear forms of Kloosterman sums. We use these to prove:

- Theorem 4.2 (parts (i) and (ii), resp.) in Sections 4.4 and 4.5.1, building on Maynard [May25b];
- Theorem 4.19 and Corollary 4.3 in Section 4.5.2, building on Lichtman [Lic23];
- Theorem 4.4 in Section 4.6, building on Drappeau [Dra15];
- Theorem 4.29 and Corollary 4.5 in Section 4.7, building on de la Bretèche–Drappeau [BD20].

The figure below summarizes the relationships between the results in this chapter.

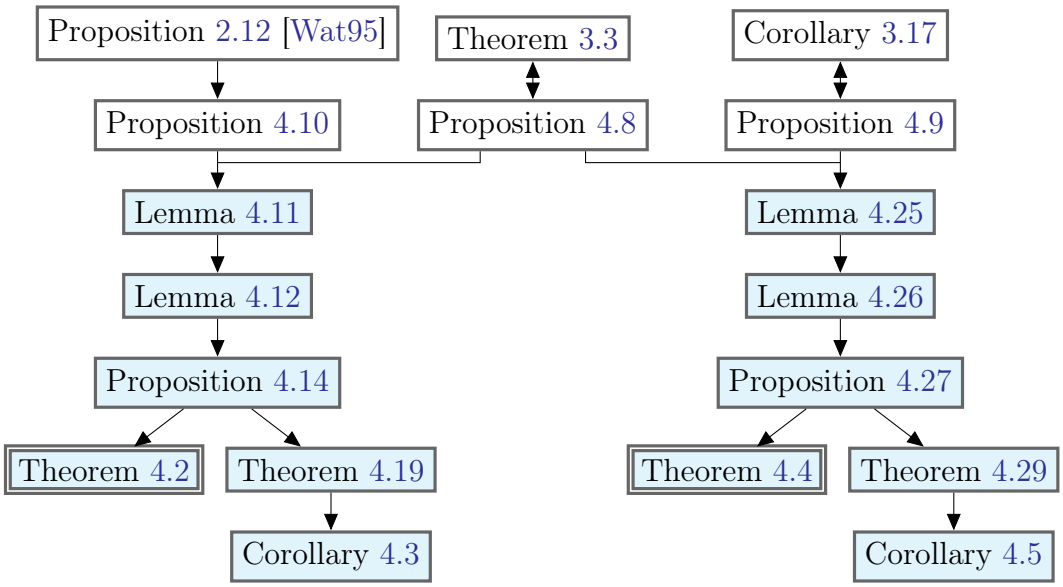

Figure 4.1: Structure of proofs (*arrows represent logical implications*).

4.3 Kloosterman sums in the exceptional spectrum

Here we develop improved bounds for multilinear forms of Kloosterman sums, applicable in particular to the expression in (4.13). This relies on the framework of Deshouillers–Iwaniec [DI82a], with new inputs in the exceptional spectrum.

4.3.1 Large sieve for exceptional forms

Following Deshouillers–Iwaniec [DI82c], in expressions like (4.14) and (4.15), the factors of X^{θ_f} can be tempered through large sieve inequalities for the Fourier coefficients of exceptional Maass forms, of the shape in (4.4). The goal is to incorporate factors of X^{θ_f} with X as large as possible, while matching the essentially-optimal upper bound for the regular-spectrum large sieve inequalities (see Proposition 2.8). Below we recall Assumption 3.13, which is just a framework to state large sieve inequalities and their corollaries succinctly. On a first read, one should pretend that $\xi = 0$ and $A = \|a_n\|_2$.

Assumption 4.6. *We say that a tuple $(q, N, Z, (a_n)_{n\sim N}, A, Y)$, with $q \in \mathbb{Z}_+$, $N \geq 1/2$, $Z \gg 1$, $A \gg \|a_n\|_2$, $Y > 0$, satisfies this assumption iff the following holds. For any $\varepsilon > 0$, $\xi \in \mathbb{R}$, any cusp $\mathfrak{a}$ of $\Gamma_0(q)$ with $\mu(\mathfrak{a}) = q^{-1}$, and any orthonormal basis (f_j) of exceptional Maass cusp forms for $\Gamma_0(q)$, with Laplacian eigenvalues λ_j, $\theta_j := \sqrt{1-4\lambda_j}$, and Fourier coefficients $\rho_{j\mathfrak{a}}(n)$ (using the choice of scaling matrix in (2.12)), one has*

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

for all

$$X \ll \max\left(1, \frac{q}{N}\right) \frac{Y}{1+|\xi|^2}. \tag{4.16}$$

The reason for this notation is that a result of Deshouillers–Iwaniec [DI82c], given in Proposition 4.7 and reiterated below, incorporates a factor of $X = \max(1, \frac{q}{N})$ for an arbitrary sequence (a_n). In fact, this is still the best-known result for general sequences and individual levels, when $N \gg \sqrt{q}$. Therefore, the factor Y in (4.16) represents the additional saving over this result, achieved using the special structure of the sequence (a_n).

Proposition 4.7 (Large sieve with general sequences [DI82c])**.** *Let $q \in \mathbb{Z}_+$, $N \geq 1/2$, and $(a_n)_{n\sim N}$ be any complex sequence. Then the tuple $(q, N, 1, (a_n)_{n\sim N}, \|a_n\|_2, 1)$ satisfies Assumption 4.6.*

Proof. This follows immediately from Proposition 2.10. □

Next, we recall (and slightly rephrase) the large sieve inequality Theorem 3.3, concerning the first type of sequence from (4.5). This will be the main ingredient behind the improvements in Theorems 4.2 and 4.4. Following Notation 3.4 only up to a constant, for $\alpha \in \mathbb{R}/\mathbb{Z}$ and $N > 0$ we use the notation

$$T_N(\alpha) := \min_{t\in\mathbb{Z}_+} (t + N\|t\alpha\|). \tag{4.17}$$

This quantity is nondecreasing in N, and measures the quality of rational approximations to α with small denominators t (recall that $\|\alpha\|$ denotes the distance from 0 to α in $\mathbb{R}/\mathbb{Z}$). In particular, we have $T_N(\alpha) \leq 1 + N\|\alpha\|$, $T_N(\alpha+\beta) \ll (1+N\|\beta\|)T_N(\alpha)$, and $T_N(\alpha) \ll \sqrt{N}$ by a pigeonhole argument (see Lemma 3.6).

Proposition 4.8 (Large sieve with additive convolutions)**.** *Let $N \geq 1/2$, $L, H \gg 1$, $\alpha_1, \alpha_2 \in \mathbb{R}/\mathbb{Z}$, and $q, \ell_1, \ell_2 \in \mathbb{Z}_+$, $a \in \mathbb{Z}$ be such that $q \gg L^2$, $\ell_1, \ell_2 \asymp L$, and $(\ell_1, \ell_2) = 1$. Let $\Phi_i(t) : (0,\infty) \to \mathbb{C}$ be smooth functions supported in $t \ll 1$, with $\Phi_i^{(j)} \ll_j 1$ for all $j \geq 0$, and*

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

Then the tuple $(q, N, H, (a_n)_{n\sim N}, A, Y)$ satisfies Assumption 4.6, where

$$A := \|a_n\|_2 + \sqrt{N\left(\frac{H}{L} + \frac{H^2}{L^2}\right)}, \qquad Y := \max\left(1, \frac{NH}{|a|(H+L)L\min_i T_H(\alpha_i)}\right).$$

Proof. This follows from Theorem 3.3 with $a_n \leftarrow a_{n/|a|}$, $N \leftarrow N/|a|$, $a \leftarrow |a|$, as in (3.51) (we note that the statement is trivial unless $N \ll |a|HL$). One can in fact replace $T_H(\alpha_i)$ with the smaller quantity $T_{N/(|a|L)}(\alpha_i)$, which is helpful in incorporating the phase ξ from Assumption 4.6 via the bound

$$T_{N/(|a|L)}\left(\alpha_i \pm \frac{a\ell_i\xi}{N}\right) \ll \left(1 + \frac{N}{|a|L}\frac{|a\ell_i\xi|}{N}\right) T_{N/(|a|L)}(\alpha_i) \ll (1+|\xi|)T_H(\alpha_i).$$

□

4.3.2 Multilinear forms of Kloosterman sums

We now state a couple of bounds for sums of Kloosterman sums, incorporating the large sieve inequalities with averaging over levels from Propositions 2.11 and 2.12, and also using the framework of Assumption 4.6. We will combine these results with the large sieve inequalities from Propositions 4.7 and 4.8 to remove the dependency on Selberg's eigenvalue conjecture in our results. We recall from (2.21) that $\theta_{\max} \leq \frac{7}{32}$ measures the worst counterexample to Selberg's conjecture.

Proposition 4.9 (Sums of incomplete Kloosterman sums)**.** *Let $R, S, N \geq 1/2$, $C, D, Z \gg 1$, and $Y, \varepsilon > 0$. For all $r \sim R$, $s \sim S$ with $(r,s) = 1$, let:*

- *$w_{r,s} \in \mathbb{C}$;*
- *$\Phi_{r,s} : (0,\infty)^3 \to \mathbb{C}$ be smooth, with $\Phi_{r,s}(x,y,z)$ supported in $x, y, z \asymp 1$, and*

$$\partial_x^j \partial_y^k \partial_z^\ell \Phi_q(x,y,z) \ll_{j,k,\ell,\varepsilon} Z^{j\varepsilon}, \qquad \forall j,k,\ell \geq 0;$$

- *$(rs, N, Z, (a_{n,r,s})_{n\sim N}, A_{r,s}, Y)$ be a tuple satisfying Assumption 4.6.*

Then with a consistent choice of the sign $\pm$, one has

$$\begin{aligned}\sum_{\substack{r\sim R\\ s\sim S\\ (r,s)=1}} w_{r,s} \sum_{n\sim N} a_{n,r,s} \sum_{\substack{c,d\\ (rd,sc)=1}} \Phi_{r,s}\left(\frac{n}{N}, \frac{d}{D}, \frac{c}{C}\right) e\left(\pm n\frac{\overline{rd}}{sc}\right) \\ \ll_\varepsilon (RSNCDZ)^{O(\varepsilon)} \|w_{r,s}A_{r,s}\|_2 \mathscr{I},\end{aligned} \tag{4.18}$$

where

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

Proof. This is Corollary 3.17 from Chapter 3. □

Proposition 4.10 (Sums of Kloosterman sums with multiplicative convolutions)**.** *Let $R, S, N \geq 1/2$, $M_1, M_2, C, Z \gg 1$, $M \asymp M_1 M_2$, and $Y, \varepsilon > 0$. For all $r \sim R$, $s \sim S$ with $(r,s) = 1$, let:*

- *$w_{r,s} \in \mathbb{C}$;*
- *$\Phi_{r,s} : (0,\infty)^4 \to \mathbb{C}$ be smooth, with $\Phi_{r,s}(x_1, x_2, y, z)$ supported in $x_1, x_2, y, z \asymp 1$, and*

$$\partial_{x_1}^{j_1} \partial_{x_2}^{j_2} \partial_y^k \partial_z^\ell \Phi_q(x_1, x_2, y, z) \ll_{j_1, j_2, k, \ell, \varepsilon} Z^{(j_1+j_2+k)\varepsilon}, \qquad \forall j_1, j_2, k, \ell \geq 0;$$

- *$(rs, N, Z, (a_{n,r,s})_{n\sim N}, A_{r,s}, Y)$ be a tuple satisfying Assumption 4.6.*

Then with a consistent choice of the sign $\pm$, it holds that

$$\begin{aligned} \sum_{\substack{r\sim R \\ s\sim S \\ (r,s)=1}} w_{r,s} \sum_{m_1,m_2\in\mathbb{Z}} \sum_{n\sim N} a_{n,r,s} \sum_{(c,r)=1} \Phi_{r,s}\left(\frac{m_1}{M_1}, \frac{m_2}{M_2}, \frac{n}{N}, \frac{c}{C}\right) S(m_1 m_2 \overline{r}, \pm n; sc) \\ \ll_\varepsilon (RSMNCZ)^{O(\varepsilon)} \left(1 + \frac{C\sqrt{M_1}}{R\sqrt{SY}}\right)^{\theta_{\max}} \|w_{r,s} A_{r,s}\|_2 \sqrt{RSM} \\ \times \left(\frac{C^2}{R}(M + RS)(N + RS) + MN\right)^{1/2}. \end{aligned} \tag{4.19}$$

Proof. We closely follow the proof of Corollary 3.16, with minor changes. We start by inserting coefficients $\Psi_i(m_i/M_i)$ in the sum $\mathcal{S}$ from the left-hand side of (4.19); here $\Psi_i(t)$ are smooth functions with $\Psi_i^{(j)} \ll_j 1$, supported in $t \asymp 1$, and equal to 1 on the supports of $x_1, x_2 \asymp 1$ in $\Phi_{r,s}(x_1, x_2, y, z)$. We then separate variables using the Fourier inversion formula

$$\begin{aligned} \Psi_{r,s}(x_1, x_2, y; z) &:= \sqrt{x_1 x_2 y}\, \Phi_{r,s}\left(x_1, x_2, y, \frac{\sqrt{x_1 x_2 y}}{z}\right) \\ &= \iiint_{\mathbb{R}^3} \widehat{\Psi}_{r,s}(\zeta_1, \zeta_2, \xi; z)\, e(x_1\zeta_1 + x_2\zeta_2 + y\xi)\, d\zeta_1\, d\zeta_2\, d\xi, \end{aligned}$$

where the Fourier transform is taken in the first three variables. Thus

$$\begin{aligned} \Phi_{r,s}\left(\frac{m_1}{M_1}, \frac{m_2}{M_2}, \frac{n}{N}, \frac{c}{C}\right) = \frac{\sqrt{M_1 M_2 N}}{\sqrt{m_1 m_2 n}} \iiint_{\mathbb{R}^2} \widehat{\Psi}_{r,s}\left(\zeta_1, \zeta_2, \xi; \frac{C\sqrt{m_1 m_2 n}}{c\sqrt{M_1 M_2 N}}\right) \\ \times e\left(\frac{m_1}{M_1}\zeta_1 + \frac{m_2}{M_2}\zeta_2 + \frac{n}{N}\xi\right) d\zeta_1\, d\zeta_2\, d\xi. \end{aligned}$$

Similarly as in (3.54), this yields

$$\mathcal{S} \ll_\varepsilon Z^{O(\varepsilon)} CS\sqrt{R} \iiint_{\mathbb{R}^2} \frac{S(\zeta_1, \zeta_2, \xi)\, d\zeta_1\, d\zeta_2\, d\xi}{(1+\zeta_1^{100})(1+\zeta_2^{100})(1+\xi^{100})}, \tag{4.20}$$

where

$$\mathcal{S}(\zeta_1,\zeta_2,\xi) := \sum_{\substack{r\sim R\\ s\sim S\\ (r,s)=1}} |w_{r,s}| \Bigg| \sum_{m_1,m_2\in\mathbb{Z}} \Psi_1\left(\frac{m_1}{M_1}\right)\Psi_2\left(\frac{m_2}{M_2}\right) e\left(\frac{m_1}{M_1}\zeta_1+\frac{m_2}{M_2}\zeta_2\right)$$
$$\times \sum_{n\sim N} b_n\, e\left(\frac{n}{N}\xi\right) \sum_{(c,r)=1} \frac{S(m_1m_2\overline{r},\pm n;sc)}{cs\sqrt{r}} \varphi_{\zeta_1,\zeta_2,\xi,r,s}\left(\frac{4\pi\sqrt{m_1m_2n}}{c}\right)\Bigg|,$$

and $\varphi_{\zeta_1,\zeta_2,\xi,r,s}(z)$ is supported in $z \asymp X^{-1}$, and satisfies $\varphi^{(\ell)}_{\zeta_1,\zeta_2,\xi} \ll_\ell X^\ell$ for

$$X := \frac{CS\sqrt{R}}{\sqrt{MN}}. \tag{4.21}$$

We can incorporate the factors $e(t\zeta_i)$ into the functions $\Psi_i(t)$, incurring derivative bounds $\Psi_i^{(j)} \ll_j 1+|\zeta_i|^j$. From here on, the proof is analogous to that of Corollary 3.16 (starting with an application of the Kuznetsov formula for the cusps ∞ and $1/s$), except that we apply Proposition 2.12 instead of Proposition 2.11 in the exceptional spectrum; we use $Z = \max_i(1+|\zeta_i|)$ in Proposition 2.12, which disappears in the integral over ζ_1,ζ_2 from (4.20). Importantly, instead of (3.66) we use

$$X_1 := \frac{Q^2}{M_1^2M_2},$$

as in (2.41). Combining this with the value of X from (4.21) and the value of $X_2(\xi)$ from (4.16) (with $q=rs$) leads to a total exceptional factor of

$$\left(1+\frac{X}{\sqrt{X_1X_2(0)}}\right)^{\theta_{\max}} \ll \left(1+\frac{CS\sqrt{R}}{\sqrt{M_1M_2N}}\frac{M_1\sqrt{M_2}}{Q}\frac{\sqrt{N}}{\sqrt{RSY}}\right)^{\theta_{\max}}$$
$$= \left(1+\frac{C\sqrt{M_1}}{R\sqrt{SY}}\right)^{\theta_{\max}},$$

as in (4.19). Other than this, the right-hand side of (4.19) is identical to that of (3.64), after inserting the follow-up bound from (3.65). □

4.4 Primes with triply-well-factorable weights

Here we prove Theorem 4.2.(*i*) rigorously, building on the arguments of Maynard [May25b]. Compared to the outline in Section 4.2, we will essentially work in reverse, starting from bounds for multilinear forms of Kloosterman sums and building up to a convolution estimate in Proposition 4.14.

We begin with a bound for a sum like in (4.13), which follows from Propositions 4.8 and 4.10.

Lemma 4.11. *Let $\varepsilon > 0$, $a \in \mathbb{Z} \setminus \{0\}$, $1 \ll S, F, K, C, H \ll x^{O(1)}$, $\Phi_i(t)$ be smooth functions supported in $t \asymp 1$ with $\Phi_i^{(j)} \ll_j 1$, and*

$$\phi(h_1, h_2) := \Phi_1\left(\frac{h_1}{H}\right)\Phi_2\left(\frac{h_2}{H}\right) e(h_1\alpha_1 + h_2\alpha_2),$$

where $\alpha_i \in \mathbb{R}/\mathbb{Z}$ have $\min_i T_H(\alpha_i) \ll_\varepsilon x^\varepsilon$ (recall (4.17)). Then for any smooth function $\Phi(x_1, x_2, z)$ supported in $x_i, z \asymp 1$, satisfying $\partial_{x_1}^{j_1}\partial_{x_2}^{j_2}\partial_z^\ell \Phi(x_1, x_2, z) \ll_{j_1, j_2, \ell, \varepsilon} x^{(j_1+j_2)\varepsilon}$, one has

$$\sum_{s_1, s_2 \sim S} \left| \sum_{f,k} \sum_{\substack{h_1, h_2 \\ \ell = h_1 s_1 - h_2 s_2 \neq 0}} \phi(h_1, h_2) \sum_{(c, s_1 s_2) = 1} \Phi\left(\frac{f}{F}, \frac{k}{K}, \frac{c}{C}\right) S(fk\overline{s_1 s_2}, a\ell; c) \right|$$
$$\ll_{\varepsilon, a} x^{O(\varepsilon)} \left(1 + \frac{C\sqrt{F}}{S^2\sqrt{\frac{H^2}{H+S}}}\right)^{\theta_{\max}} \sqrt{H^2 S^3 (H+S) FK}$$
$$\times \left(\frac{C^2}{S}\left(FK + S^2\right)(H+S) + FKHS\right)^{1/2}.$$

Proof. Let $\mathcal{K}_0$ denote the sum in the left-hand side. We first let $s_0 := (s_1, s_2)$, change variables $s_i \leftarrow s_0 s_i$, $\ell \leftarrow s_0 \ell$ for $i \in \{1, 2\}$, and put s_0 into dyadic ranges. This yields

$$\mathcal{K}_0 \ll x^{o(1)} \sup_{S_0 \ll S} \mathcal{K}_1(S_0), \tag{4.22}$$

where after simplifying $S(fk\overline{s_0^2 s_1 s_2}, as_0\ell; c) = S(fk\overline{s_0 s_1 s_2}, a\ell; c)$,

$$\mathcal{K}_1 := \sum_{s_0 \sim S_0} \sum_{\substack{s_1, s_2 \sim S/s_0 \\ (s_1, s_2) = 1}} \left| \sum_{f,k} \sum_{\substack{h_1, h_2 \\ \ell = h_1 s_1 - h_2 s_2 \neq 0}} \phi(h_1, h_2) \sum_{(c, s_0 s_1 s_2) = 1} \Phi\left(\frac{f}{F}, \frac{k}{K}, \frac{c}{C}\right) \right.$$
$$\left. \times S(fk\overline{s_0 s_1 s_2}, a\ell; c) \right|.$$

We then put $n = |a\ell|$ and $r = s_0 s_1 s_2$ in dyadic ranges $n \sim \mathcal{N}$, $r \sim \mathcal{R}$, insert coefficients $\Psi(n/\mathcal{N})$ where $\Psi^{(j)} \ll_j 1$ and $\psi \equiv 1$ on $[1, 2]$, and use the divisor bound to write

$$\mathcal{K}_1 \ll x^{o(1)} \sup_{\substack{\mathcal{N} \ll_a HS/S_0 \\ \mathcal{R} \asymp S^2/S_0}} \mathcal{K}_2(\mathcal{N}, \mathcal{R}), \tag{4.23}$$

for

$$\mathcal{K}_2 := \sum_{r\sim\mathcal{R}} \max_{\substack{s_0\sim S_0\\ s_1,s_2\sim S/s_0\\ (s_1,s_2)=1\\ s_0s_1s_2=r}} \left| \sum_{f,k}\sum_{n\sim\mathcal{N}} \sum_{\substack{h_1,h_2\\ a(h_1s_1-h_2s_2)=\pm n}} \phi(h_1,h_2) \sum_{(c,r)=1} \Psi\left(\frac{n}{\mathcal{N}}\right) \Phi\left(\frac{f}{F},\frac{k}{K},\frac{c}{C}\right) \times S(fk\bar{r},\pm n;c)\right|, \tag{4.24}$$

where the supremum in (4.23) includes the choice of the $\pm$ sign. If the maximum above is attained at some $s_1(r), s_2(r)$, we let

$$\begin{aligned} a_{n,r} &:= \sum_{\substack{h_1,h_2\in\mathbb{Z}\\ a(h_1s_1(r)-h_2s_2(r))=\pm n}} \phi(h_1,h_2) \\ &= \sum_{\substack{h_1,h_2\in\mathbb{Z}\\ \pm a(h_1s_1(r)-h_2s_2(r))=n}} \Phi_1\left(\frac{h_1}{H}\right)\Phi_2\left(\frac{h_2}{H}\right) e(h_1\alpha_1+h_2\alpha_2). \end{aligned}$$

If the maximum is empty, we let $a_{n,r} := 0$. Then we can rewrite (4.24) as

$$\mathcal{K}_2 = \sum_{r\sim\mathcal{R}} \left| \sum_{f,k}\sum_{n\sim\mathcal{N}} a_{n,r} \sum_{(c,r)=1} \Psi\left(\frac{n}{\mathcal{N}}\right) \Phi\left(\frac{f}{F},\frac{k}{K},\frac{c}{C}\right) S(fk\bar{r},\pm n;c)\right|.$$

By Proposition 4.8, the tuple $(r,\mathcal{N},x,(a_{n,r})_{n\sim\mathcal{N}},A_r,Y)$ satisfies Assumption 4.6, where

$$\begin{aligned} Y &:= \frac{\mathcal{N}H}{|a|(H+S/S_0)(S/S_0)\min_i T_H(\alpha_i)}, \\ A_r &:= \left(\sum_{n\sim\mathcal{N}} |a_{n,r}|^2\right)^{1/2} + \sqrt{\mathcal{N}}\sqrt{\frac{HS_0}{S}+\frac{H^2S_0^2}{S^2}}. \end{aligned}$$

Since $\min_i T_H(\alpha_i) \ll_\varepsilon x^\varepsilon$, we further have

$$Y \gg_{\varepsilon,a} x^{-\varepsilon}\frac{\mathcal{N}HS_0}{(H+S)S}.$$

We can now apply Proposition 4.10; specifically, by (4.19), we obtain

$$\begin{aligned} \mathcal{K}_2 \ll_{\varepsilon,a} x^{O(\varepsilon)} &\left(1+\frac{C\sqrt{F}}{\mathcal{R}\sqrt{\frac{\mathcal{N}HS_0}{(H+S)S}}}\right)^{\theta_{\max}} \|A_r\|_2\sqrt{\mathcal{R}FK} \\ &\times\left(\frac{C^2}{\mathcal{R}}(FK+\mathcal{R})(\mathcal{N}+\mathcal{R})+FK\mathcal{N}\right)^{1/2}. \end{aligned}$$

We claim that

$$\|A_r\|_2^2 \ll x^{o(1)}\mathcal{N}(HS+H^2S_0). \tag{4.25}$$

Indeed, this follows from the definition of A_r and the two bounds

$$\sum_{r\sim\mathcal{R}}\mathcal{N}\left(\frac{HS_0}{S}+\frac{H^2S_0^2}{S^2}\right)\asymp\frac{S^2}{S_0}\mathcal{N}\left(\frac{HS_0}{S}+\frac{H^2S_0^2}{S^2}\right)\asymp\mathcal{N}(HS+H^2S_0),$$

$$\begin{aligned}\sum_{r\sim\mathcal{R}}\sum_{n\sim\mathcal{N}}|a_{n,r}|^2&\ll\sum_{n\sim\mathcal{N}}\sum_{s_0\sim S_0}\sum_{\substack{s_1,s_2\sim S/s_0\\(s_1,s_2)=1}}\left(\sum_{\substack{h_1,h_2\asymp H\\h_1s_1-h_2s_2=\pm n/a}}1\right)^2\\&\ll\sum_{n\sim\mathcal{N}}\sum_{s_0\sim S_0}\sum_{\substack{s_1,s_2\sim S/s_0\\(s_1,s_2)=1}}\sum_{\substack{h_1,h_2\asymp H\\h_1s_1-h_2s_2=\pm n/a}}\sum_{\substack{h_1',h_2'\asymp H\\s_1(h_1-h_1')=s_2(h_2-h_2')}}1\\&\ll\sum_{n\sim\mathcal{N}}\sum_{s_0\sim S_0}\sum_{\substack{s_1\sim S/s_0\\h_1\asymp H}}\sum_{\substack{s_2\sim S/s_0\\h_2\asymp H\\h_2s_2=h_1s_1\mp n/a}}\sum_{\substack{h_1'\asymp H\\h_1'\equiv h_1\pmod{s_2}}}\sum_{\substack{h_2'\asymp H\\s_1(h_1-h_1')=s_2(h_2-h_2')}}1\\&\ll x^{o(1)}\mathcal{N}S_0\frac{S}{S_0}H\left(1+\frac{HS_0}{S}\right)\quad=x^{o(1)}\mathcal{N}(HS+H^2S_0).\end{aligned}$$

Using (4.25), we can further bound

$$\begin{aligned}\mathcal{K}_2\ll_{\varepsilon,a}x^{O(\varepsilon)}&\left(1+\frac{C\sqrt{F}}{\mathcal{R}\sqrt{\frac{\mathcal{N}HS_0}{(H+S)S}}}\right)^{\theta_{\max}}\sqrt{\mathcal{N}(HS+H^2S_0)\mathcal{R}FK}\\&\times\left(\frac{C^2}{\mathcal{R}}(FK+\mathcal{R})(\mathcal{N}+\mathcal{R})+FK\mathcal{N}\right)^{1/2},\end{aligned}$$

where the right-hand side is increasing in $\mathcal{N}\ll_a HS/S_0$. Substituting this value of $\mathcal{N}$ and $\mathcal{R}\asymp S^2/S_0$, it follows from (4.23) that

$$\begin{aligned}\mathcal{K}_1\ll_{\varepsilon,a}x^{O(\varepsilon)}&\left(1+\frac{C\sqrt{F}}{\frac{S^2}{S_0}\sqrt{\frac{H^2}{H+S}}}\right)^{\theta_{\max}}\\&\times\sqrt{\frac{HS}{S_0}(HS+H^2S_0)\frac{S^2}{S_0}FK}\left(\frac{C^2S_0}{S^2}\left(FK+\frac{S^2}{S_0}\right)\left(\frac{HS}{S_0}+\frac{S^2}{S_0}\right)+\frac{FKHS}{S_0}\right)^{1/2}.\end{aligned}$$

Since $\theta_{\max}<1/2$, the right-hand side is seen to be decreasing in $S_0\gg1$; substituting S_0 with 1 and plugging this into (4.22), we obtain the desired bound for $\mathcal{K}_0$. □

We use Lemma 4.11 to deduce a power-saving bound for an exponential sum like in (4.10) (before passing to the complementary divisor). This improves [May25b, Lemma 7.1] by allowing larger ranges of Q, R, S in (4.26); as a technical difference, we require that h lies in a smooth dyadic range. We note in passing that the case $h<0$ follows immediately by changing $a\leftrightarrow-a$.

Lemma 4.12 (Exponential sum bound for well-factorable weights)**.** *Let $a \in \mathbb{Z} \setminus \{0\}$, $d \in \mathbb{Z}_+$ with $(a,d) = 1$, $\varepsilon, C > 0$, and $M, N, x, Q, R, S \gg 1$ satisfy $MN \asymp x$ and, with $\theta := 7/32$,*

$$\begin{aligned} N^2R^2S &\le x^{1-8\varepsilon}, \\ N^{\frac{2+\theta}{2-2\theta}}RS^{\frac{4-5\theta}{2-2\theta}} &\le x^{1-16\varepsilon}, \\ N^{\frac{1+\theta}{1-\theta}}Q^{\frac{1-3\theta}{1-\theta}}R^2S^5 &\le x^{2-32\varepsilon}. \end{aligned} \tag{4.26}$$

Let $Q' \in [Q, 2Q]$, $1 \ll H \le x^{o(1)}QR^2S^2/M$, $B_i \gg 1$, and let $\mathscr{N} \subset \mathbb{Z}_+^2$ be such that if $(u;v), (u';v') \in \mathscr{N}$, then $(u,v') = (u',v) = 1$. Finally, let (γ_r), (λ_s), (α_n) be 1-bounded sequences, $\omega \in \mathbb{R}/\mathbb{Z}$ with $T_H(\omega) \ll x^{o(1)}$, and $\Phi(t)$ be a smooth function supported in $t \asymp 1$, with $\Phi^{(j)} \ll_j 1$ for $j \ge 0$. Then

$$\sum_{\substack{Q \le q \le Q' \\ (q,a)=1}} \sum_{r_1, r_2 \sim R} \sum_{\substack{s_1, s_2 \sim S \\ (r_1 s_1, a r_2 s_2) = 1 \\ (r_2 s_2, aqdr_1 s_1) = 1 \\ r_i s_i \le B_i}} \frac{\gamma_{r_1}\lambda_{s_1}\overline{\gamma_{r_2}\lambda_{s_2}}}{r_1 r_2 s_1 s_2 q} \sum_{\substack{n_1, n_2 \sim N \\ n_1 \equiv n_2 \pmod{qd} \\ (n_1, n_2 qdr_1 s_1) = 1 \\ (n_2, n_1 qdr_2 s_2) = 1 \\ (n_1 r_2 s_2; n_2) \in \mathscr{N} \\ |n_1 - n_2| \ge N/(\log x)^C}} \alpha_{n_1}\overline{\alpha_{n_2}}$$
$$\times \sum_{h \in \mathbb{Z}} e(h\omega)\, \Phi\left(\frac{h}{H}\right) e\left(\frac{ah(n_1 - n_2)\overline{n_2 r_1 s_1 dq}}{n_1 r_2 s_2}\right) \ll_{a,\varepsilon,C} \frac{N^2}{Qx^{\varepsilon}}.$$

Proof. We closely follow the proof of [May25b, Lemma 7.1], taking $Q \ll N$ without loss of generality (otherwise the sum over n_1, n_2 vanishes). After the substitution $fdq = n_1 - n_2$, a separation of variables and an application of Cauchy–Schwarz in $f, n_1, n_2, r_1, r_2, s_2$, we reach the sum

$$\mathscr{W}_4 := \sum_{\substack{b,c,f \\ (b,c)=1}} \Psi_0\left(\frac{b}{B}\right)\Psi_0\left(\frac{c}{C_0}\right)\Psi_0\left(\frac{f}{F_0}\right)\left|\sum_{s \sim S}\lambda'_s \sum_{h \in \mathbb{Z}} e(h\omega)\,\Phi\left(\frac{h_1}{H}\right) e\left(\frac{ahf\overline{bs}}{c}\right)\right|^2,$$

similar to [May25b, p. 23, third display]. Here we also inserted a smooth majorant in the f variable, where Ψ_0 is a compactly-supported nonnegative function satisfying $\Psi_0^{(j)} \ll_j 1$. As in [May25b, p. 23, second display], the ranges B, C_0, F_0 satisfy

$$B \ll NR, \qquad C_0 \ll NRS, \qquad F_0 \ll \frac{N}{Q}, \tag{4.27}$$

and as in [May25b, (7.4)], we need to show that $\mathscr{W}_4 \ll_{\varepsilon,a} x^{-6\varepsilon}N^2R^2S^3$. Normally at this stage, we would expand the square in $\mathscr{W}_4$, leading to a sum like in (4.11), and then complete Kloosterman sums. But as outlined in Section 4.2.3, to achieve good savings in the complementary divisor ($f \sim F$) aspect, we will need to 'move' f to the other entry of the resulting Kloosterman sums. Towards this goal, we split the sum according to the value of $d = (f, c)$:

$$\mathscr{W}_4 \le \sum_{1 \le d \ll x} \mathscr{W}_5(d), \tag{4.28}$$

where, after relaxing the constraint $(b,c)=1$ to $(b,c/d)=1$, substituting $(f,c)\leftarrow(fd,cd)$, and letting

$$C := \frac{C_0}{d}, \qquad F := \frac{F_0}{d}, \tag{4.29}$$

we have

$$\mathscr{W}_5 := \sum_{\substack{b,c,f\\(bf,c)=1}} \Psi_0\left(\frac{b}{B}\right)\Psi_0\left(\frac{c}{C}\right)\Psi_0\left(\frac{f}{F}\right)\left|\sum_{\substack{s\sim S\\(s,c)=1}}\lambda'_s\sum_{h\in\mathbb{Z}} e(h\omega)\,\Phi\left(\frac{h_1}{H}\right)e\left(\frac{ah\overline{fbs}}{c}\right)\right|^2.$$

Due to (4.28) and $\sum_{1\le d\ll x}\frac{1}{d}\ll\log x$, it is enough to show that

$$\mathscr{W}_5 \ll_{\varepsilon,a} x^{-7\varepsilon}\frac{N^2R^2S^3}{d}. \tag{4.30}$$

Now let

$$\mathscr{W}(x;c) := \sum_{\substack{s\sim S\\(s,c)=1}}\lambda'_s\sum_{h\in\mathbb{Z}} e(h\omega)\,\Phi\left(\frac{h_1}{H}\right)e\left(\frac{ah\overline{xs}}{c}\right),$$

for $x\in(\mathbb{Z}/c\mathbb{Z})^\times$, so we can write

$$\begin{aligned}
\mathscr{W}_5 &= \sum_c \Psi_0\left(\frac{c}{C}\right)\sum_{(f,c)=1}\Psi_0\left(\frac{f}{F}\right)\sum_{(b,c)=1}\Psi_0\left(\frac{b}{B}\right)|\mathscr{W}(b\overline{f};c)|^2\\
&= \sum_c \Psi_0\left(\frac{c}{C}\right)\sum_{(f,c)=1}\Psi_0\left(\frac{f}{F}\right)\sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times}|\mathscr{W}(x;c)|^2\sum_{b\equiv xf\pmod c}\Psi_0\left(\frac{b}{B}\right)\\
&\le \sum_c \Psi_0\left(\frac{c}{C}\right)\sum_{f}\Psi_0\left(\frac{f}{F}\right)\sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times}|\mathscr{W}(x;c)|^2\sum_{b\equiv xf\pmod c}\Psi_0\left(\frac{b}{B}\right),
\end{aligned}$$

where we dropped the restriction $(f,c)=1$ in the last line. Expanding the square and swapping sums, we get

$$\begin{aligned}
\mathscr{W}_5 \le \sum_{\substack{s_1,s_2\sim S\\(s_1s_2,a)=1}}\lambda'_{s_1}\overline{\lambda'_{s_2}}\sum_{h_1,h_2}\phi(h_1,h_2)\sum_f\Psi_0\left(\frac{f}{F}\right)\sum_{(c,s_1s_2)=1}\Psi_0\left(\frac{c}{C}\right)\\
\times\sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times}e\left(\frac{a\ell\overline{s_1s_2x}}{c}\right)\sum_{b\equiv xf\pmod c}\Psi_0\left(\frac{b}{B}\right),
\end{aligned}$$

where

$$\ell := h_1s_1-h_2s_2, \qquad \phi(h_1,h_2) := e((h_1-h_2)\omega)\,\Phi\left(\frac{h_1}{H}\right)\overline{\Phi\left(\frac{h_2}{H}\right)}. \tag{4.31}$$

Splitting the sum above into the terms with $\ell=0$ and $\ell\neq0$, we have

$$\mathscr{W}_5 \le \mathscr{W}_{\ell=0}+\mathscr{W}_{\ell\neq0}. \tag{4.32}$$

In light of (4.27) and (4.29), the diagonal terms contribute at most

$$\begin{aligned}\mathscr{W}_{\ell=0} &\ll \sum_{s_1,s_2\sim S}\sum_{\substack{h_1,h_2\asymp H\\ h_1s_1=h_2s_2}}\sum_{f\asymp F}\sum_{c\asymp C}\sum_{b\asymp B}\sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times}\mathbb{1}_{b\equiv xf \pmod c}\\ &\ll \sum_{s_1,s_2\sim S}\sum_{\substack{h_1,h_2\asymp H\\ h_1s_1=h_2s_2}}\sum_{f\asymp F}\sum_{c\asymp C}\sum_{b\asymp B}(b,c,f)\\ &\ll x^{o(1)}SHFCB\\ &\ll x^{o(1)}S\frac{QR^2S^2}{x/N}\frac{N}{dQ}\frac{NRS}{d}NR \ll \frac{N^4R^4S^4}{d^2x^{1-\varepsilon}},\end{aligned} \tag{4.33}$$

and this is acceptable in (4.30) provided that

$$N^2R^2S \ll_\varepsilon x^{1-8\varepsilon},$$

which we assumed in (4.26). For the off-diagonal terms, which roughly correspond to (4.11), we complete the inner sum over b via Lemma 2.2 to obtain

$$\mathscr{W}_{\ell\neq 0} \ll |\mathscr{W}_6| + O(x^{-90}) + x^{o(1)}\sup_{\substack{K\ll x^{o(1)}B^{-1}C\\ \Psi^{(k)}\ll_k 1\\ u\asymp 1}}|\mathscr{W}_7(K,u)|, \tag{4.34}$$

where Ψ is a smooth function supported in $(\frac12,2)$,

$$\begin{aligned}\mathscr{W}_6 := \sum_{\substack{s_1,s_2\sim S\\ (s_1s_2,a)=1}}\lambda'_{s_1}\overline{\lambda'_{s_2}}\sum_{h_1,h_2}\phi(h_1,h_2)\sum_f\Psi_0\left(\frac{f}{F}\right)\sum_{(c,s_1s_2)=1}\Psi_0\left(\frac{c}{C}\right)\\ \times\sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times}e\left(\frac{a\ell\overline{s_1s_2x}}{c}\right)\frac{B}{c}\widehat{\Psi_0}(0)\end{aligned}$$

is the contribution of the principal frequency, and

$$\begin{aligned}\mathscr{W}_7 := \sum_{\substack{s_1,s_2\sim S\\ (s_1s_2,a)=1}}\lambda'_{s_1}\overline{\lambda'_{s_2}}\sum_{\substack{h_1,h_2\\ \ell=h_1s_1-h_2s_2\neq 0}}\phi(h_1,h_2)\sum_f\Psi_0\left(\frac{f}{F}\right)\sum_{(c,s_1s_2)=1}\Psi_0\left(\frac{c}{C}\right)\\ \times\sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times}e\left(\frac{a\ell\overline{s_1s_2x}}{c}\right)\frac{B}{C}\Psi_0\left(\frac{uc}{C}\right)\sum_k\Psi\left(\frac{|k|}{K}\right)e\left(-k\frac{uB}{C}\right)e\left(\frac{xfk}{c}\right).\end{aligned}$$

We first bound $\mathscr{W}_6$ using the Ramanujan sum bound (see Lemma 2.1), (4.27) and (4.29):

$$\begin{aligned}\mathscr{W}_6 &= \sum_{\substack{s_1,s_2\sim S\\(s_1s_2,a)=1}} \lambda'_{s_1}\overline{\lambda'_{s_2}} \sum_{h_1,h_2}\phi(h_1,h_2)\sum_f \Psi_0\left(\frac{f}{F}\right)\sum_{(c,s_1s_2)=1}\Psi_0\left(\frac{c}{C}\right)S(0,a\ell;c)\frac{B}{c}\widehat{\Psi_0}(0)\\ &\ll \sum_{s_1,s_2\sim S}\sum_{h_1,h_2\asymp H}\sum_{f\asymp F}\sum_{c\asymp C}(a\ell,c)\frac{B}{c}\\ &\ll_a x^{o(1)}S^2H^2FB\\ &\ll x^{o(1)}S^2\left(\frac{QR^2S^2}{x/N}\right)^2\frac{N}{dQ}NR \ll_\varepsilon x^\varepsilon\frac{QR^5S^6N^4}{dx^2}.\end{aligned} \tag{4.35}$$

This is acceptable in (4.30) (i.e., $\ll_\varepsilon x^{-7\varepsilon}N^2R^2S^3/d$) provided that

$$N^2QR^3S^3 \ll x^{2-8\varepsilon},$$

which follows from $Q \ll N$ and the first and third assumptions in (4.26):

$$\begin{aligned}N^2QR^3S^3 &\ll N^3R^3S^3\\ &\le \left(N^2R^2S\right)^{4/3}\cdot\left(NR^2S^5\right)^{1/3} \le \left(x^{1-8\varepsilon}\right)^{4/3}\left(x^{2-32\varepsilon}\right)^{1/3} < x^{2-8\varepsilon}.\end{aligned}$$

We are left to consider $\mathscr{W}_7$, which roughly corresponds to the sum in (4.13), and can be rewritten as

$$\begin{aligned}\mathscr{W}_7 = \frac{B}{C}\sum_{\substack{s_1,s_2\sim S\\(s_1s_2,a)=1}}\lambda'_{s_1}\overline{\lambda'_{s_2}}\sum_{f,k}\Psi_0\left(\frac{f}{F}\right)\Psi\left(\frac{|k|}{K}\right)e\left(-k\frac{uB}{C}\right)\sum_{\substack{h_1,h_2\\ \ell=h_1s_1-h_2s_2\neq 0}}\phi(h_1,h_2)\\ \times\sum_{(c,s_1s_2)=1}\Psi_0\left(\frac{c}{C}\right)\Psi_0\left(\frac{uc}{C}\right)S(fk\overline{s_1s_2},a\ell;c).\end{aligned}$$

We can now apply Lemma 4.11 with the smooth weight

$$\Phi(x_1,x_2,z) := \Psi_0(x_1)\Psi(x_2)e\left(\mp x_2\frac{uKB}{C}\right)\Psi_0(z)\Psi_0(uz),$$

once for each choice of the $\pm$ sign (corresponding to the sign of k; note that we have $S(-fk\overline{s_1s_2},a\ell;c) = S(fk\overline{s_1s_2},-a\ell;c)$, so one can transfer the sign change to a without loss of generality). Then Φ is compactly supported and $\partial_{x_1}^{j_1}\partial_{x_2}^{j_2}\partial_z^\ell\Phi(x_1,x_2,z) \ll_{j_1,j_2,k,\ell} (KB/C)^{j_2} \ll x^{o(j_2)}$, where we used $K \ll x^{o(1)}B^{-1}C$ by (4.34). Since $\theta_{\max} \le \frac{7}{32} = \theta$, we can bound

$$\begin{aligned}\mathscr{W}_7 \ll_a x^{o(1)}\frac{B}{C}&\left(1+\frac{C\sqrt{F}}{S^2\sqrt{\frac{H^2}{H+S}}}\right)^\theta\\ &\times\sqrt{H^2S^3(H+S)FK}\left(\frac{C^2}{S}\left(FK+S^2\right)(H+S)+FKHS\right)^{1/2}.\end{aligned}$$

At this point we note that by (4.26),

$$H \le x^{o(1)}\frac{NQR^2S^2}{x} \ll x^{o(1)}\frac{N^2R^2S}{x}S \ll x^{o(1)}S.$$

Using this and the fact that $K \ll x^{o(1)}B^{-1}C$ from (4.34), our bound for $\mathscr{W}_7$ simplifies to

$$\begin{aligned}\mathscr{W}_7 &\ll_a x^{o(1)}\frac{B}{C}\left(1+\frac{C\sqrt{F}}{S^{3/2}H}\right)^{\theta}\sqrt{H^2S^4FC/B}\left(C^2\left(FC/B+S^2\right)+FCHS/B\right)^{1/2}\\ &\ll x^{o(1)}\left(1+\frac{C\sqrt{F}}{S^{3/2}H}\right)^{\theta}\sqrt{H^2S^4F}\left(C\left(FC+BS^2\right)+FHS\right)^{1/2}.\end{aligned}$$

Plugging in the bounds for B, C, F from (4.29) and (4.27), we are left with

$$\begin{aligned}\mathscr{W}_7 &\ll_a \frac{x^{o(1)}}{d}\left(1+\frac{C_0\sqrt{F_0}}{S^{3/2}H}\right)^{\theta}\sqrt{H^2S^4F}\left(C_0\left(FC+BS^2\right)+F_0HS\right)^{1/2}\\ &\ll_a \frac{x^{o(1)}}{d}\left(1+\frac{NRS\sqrt{\frac{N}{Q}}}{S^{3/2}H}\right)^{\theta}\sqrt{H^2S^4\frac{N}{Q}}\left(NRS\left(\frac{N}{Q}NRS+NRS^2\right)+\frac{N}{Q}HS\right)^{1/2}\\ &= \frac{x^{o(1)}}{d}\left(1+\frac{N^{3/2}R}{H\sqrt{SQ}}\right)^{\theta}\frac{HS^{5/2}N}{Q}\left(NR^2S\left(N+QS\right)+H\right)^{1/2}.\end{aligned}$$

Finally, noting that the right-hand side is increasing in $H \le x^{o(1)}NQR^2S^2/x$, we get

$$\begin{aligned}\mathscr{W}_7 &\ll_a \frac{x^{o(1)}}{d}\left(1+\frac{x\sqrt{N}}{RS^{5/2}Q^{3/2}}\right)^{\theta}\frac{N^2R^2S^{9/2}}{x}\left(N^2R^2S+NQR^2S^2+\frac{NQR^2S^2}{x}\right)^{1/2}\\ &\ll_a \frac{x^{o(1)}}{d}\left(1+\frac{x\sqrt{N}}{RS^{5/2}Q^{3/2}}\right)^{\theta}\frac{N^{5/2}R^3S^5}{x}\left(N+QS\right)^{1/2}\\ &\ll_{\varepsilon,a} \frac{N^3R^3S^5}{dx^{1-\varepsilon}}\left(1+\left(\frac{x\sqrt{N}}{RS^{5/2}}\right)^{\theta}\right)+\frac{N^{5/2}Q^{1/2}R^3S^{11/2}}{dx^{1-\varepsilon}}\left(1+\left(\frac{x\sqrt{N}}{RS^{5/2}Q^{3/2}}\right)^{\theta}\right),\end{aligned}$$

where we omitted a factor of $Q^{-3\theta/2}$ in the first term of the expansion. For this to be acceptable in (4.30) (i.e., $\ll_{\varepsilon,a} x^{-7\varepsilon}N^2R^2S^3/d$), we need the following restrictions:

$$\begin{aligned} NRS^2 &\ll x^{1-8\varepsilon}, & N^{2+\theta}R^{2-2\theta}S^{4-5\theta} &\ll x^{2-2\theta-16\varepsilon},\\ NQR^2S^5 &\ll x^{2-16\varepsilon}, & N^{1+\theta}Q^{1-3\theta}R^{2-2\theta}S^{5-5\theta} &\ll x^{2-2\theta-16\varepsilon}.\end{aligned}$$

All of these restrictions follow from (4.26), $Q \ll N$, and $\theta \le 1/2$, as shown below:

$$\begin{aligned}
NRS^2 &\le \big(N^2R^2S \cdot NR^2S^5\big)^{1/3} \le \big(x^{1-8\varepsilon} \cdot x^{2-32\varepsilon}\big)^{1/3} < x^{1-8\varepsilon},\\
NQR^2S^5 &\ll \left(\frac{N}{Q}\right)^{\frac{2\theta}{1-\theta}} NQR^2S^5 = N^{\frac{1+\theta}{1-\theta}}Q^{\frac{1-3\theta}{1-\theta}}R^2S^5 < x^{2-16\varepsilon},\\
N^{2+\theta}R^{2-2\theta}S^{4-5\theta} &= \left(N^{\frac{2+\theta}{2-2\theta}}RS^{\frac{4-5\theta}{2-2\theta}}\right)^{2-2\theta} \le (x^{1-16\varepsilon})^{2-2\theta} \le x^{2-2\theta-16\varepsilon},\\
N^{1+\theta}Q^{1-3\theta}R^{2-2\theta}S^{5-5\theta} &= \left(N^{\frac{1+\theta}{1-\theta}}Q^{\frac{1-3\theta}{1-\theta}}R^2S^5\right)^{1-\theta} \le (x^{2-32\varepsilon})^{1-\theta} \le x^{2-2\theta-16\varepsilon}.
\end{aligned}$$

In light of (4.32) to (4.35), this establishes (4.30) and completes our proof. □

Our next result is a convolution estimate corresponding to (4.6), which improves [May25b, Proposition 7.2]; to state it, we recall the Siegel–Walfisz condition from [May25b, Definition 3].

Definition 4.13 (Siegel–Walfisz sequences)**.** A complex sequence $(a_n)_{n\sim N}$ is said to obey the *Siegel–Walfisz condition* iff one has

$$\sum_{\substack{n\sim N\\(n,d)=1}} a_n\left(\mathbb{1}_{n\equiv a \pmod q} - \frac{\mathbb{1}_{(n,q)=1}}{\varphi(q)}\right) \ll_A \tau(d)^{O(1)}\frac{N}{(\log N)^A},$$

for all $d, q \in \mathbb{Z}_+$, $a \in \mathbb{Z}$ with $(a,q) = 1$, and all $A > 1$.

Proposition 4.14 (Triply-well-factorable convolution estimate)**.** *Let $a \in \mathbb{Z}\setminus\{0\}$, $A, \varepsilon > 0$, and $M, N, x, Q_1, Q_2, Q_3 \gg 1$ satisfy $MN \asymp x$ and*

$$\begin{aligned}
Q_1 &\le \frac{N}{x^{\varepsilon}},\\
N^2Q_2Q_3^2 &\le x^{1-15\varepsilon},\\
N^2Q_2^5Q_3^2 &\le x^{2-40\varepsilon}.
\end{aligned} \tag{4.36}$$

Let $(\alpha_n), (\beta_m)$ be 1-bounded sequences, such that (α_n) is supported on $P^-(n) \ge z_0 := x^{1/(\log\log x)^3}$ and satisfies the Siegel–Walfisz condition from Definition 4.13. Then for any 1-bounded complex sequences $(\gamma_{q_1}), (\lambda_{q_2}), (\nu_{q_3})$ supported on $(q_i, a) = 1$, one has

$$\begin{aligned}
\sum_{q_1\sim Q_1}\gamma_{q_1}\sum_{q_2\sim Q_2}\lambda_{q_2}\sum_{q_3\sim Q_3}\nu_{q_3}\sum_{n\sim N}\alpha_n\sum_{m\sim M}\beta_m\left(\mathbb{1}_{mn\equiv a \pmod q} - \frac{\mathbb{1}_{(mn,q)=1}}{\varphi(q)}\right)\\
\ll_{\varepsilon,A,a} \frac{x}{(\log x)^A}.
\end{aligned}$$

Proof. From (4.36) we can deduce the slightly-weaker system of inequalities (with $\theta := 7/32$)

$$\begin{aligned} Q_1 &\le \frac{N}{x^{\varepsilon}}, \\ N^2 Q_2 Q_3^2 &\le x^{1-9\varepsilon}, \\ N^{\frac{2+\theta}{2-2\theta}} Q_2^{\frac{4-5\theta}{2-2\theta}} Q_3 &\le x^{1-17\varepsilon}, \\ N^{\frac{1+\theta}{1-\theta}} Q_1^{\frac{1-3\theta}{1-\theta}} Q_2^5 Q_3^2 &\le x^{2-33\varepsilon}, \end{aligned} \tag{4.37}$$

which will be enough for this proof. Indeed, the third bound in (4.37) follows from (4.36) and $(2+\theta)/(2-2\theta) = 71/50$, $(4-5\theta)/(2-2\theta) = 93/50 < 187/100$, since

$$N^{\frac{2+\theta}{2-2\theta}} Q_2^{\frac{4-5\theta}{2-2\theta}} Q_3 \le (N^2 Q_2 Q_3^2)^{21/50} (N^2 Q_2^5 Q_3^2)^{29/100} \le x^{1-17.9\varepsilon}.$$

Similarly, the fourth bound in (4.37) follows from (4.36) since

$$N^{\frac{1+\theta}{1-\theta}} Q_1^{\frac{1-3\theta}{1-\theta}} Q_2^5 Q_3^2 \le N^{\frac{1+\theta}{1-\theta}} N^{\frac{1-3\theta}{1-\theta}} Q_2^5 Q_3^2 = N^2 Q_2^5 Q_3^2 \le x^{2-40\varepsilon}.$$

We now closely follow the proof of [May25b, Proposition 7.2], which begins by factoring out the z_0-smooth parts of q_2 and q_3 and applying [May25b, Proposition 5.8] (at that step we use $N \ge Q_1 x^{\varepsilon}$). With $y_0 := x^{1/\log\log x}$, $D \le y_0^2$ and $DR \asymp Q_2 Q_3$, it remains to bound $|\mathscr{E}_1| + |\mathscr{E}_2| \ll \frac{N^2}{DQ_1 y_0}$, where $\mathscr{E}_i$ are the exponential sums from [May25b, p. 26]. As in [May25b, p. 27], the contribution of $\mathscr{E}_1$ is acceptable provided that

$$N^{3/2} Q_2 Q_3 \le x^{1-2\varepsilon}, \qquad Q_1 Q_2 Q_3 \le x^{1-2\varepsilon},$$

both of which follow easily from (4.37). As in [May25b, p. 27], to handle $\mathscr{E}_2$ it suffices to bound another exponential sum $\mathscr{E}_3$ by $\mathscr{E}_3 \ll_{\varepsilon} N^2/(Q_1 x^{\varepsilon/10})$. We note that the range of the h variable can be extended to all $h \in \mathbb{Z} \setminus \{0\}$, since the contribution of $|h| > H_2 := (\log x)^5 QDR^2/M$ is negligible.

We apply a close variant of [May25b, Lemma 5.9] (which is [May25a, Lemma 14.5]) to $\mathscr{E}_3$. Specifically, in the proof of [May25a, Lemma 14.5], we omit the step of applying partial summation in the h variable; instead, we follow the proof of Lemma 2.2 and put $|h|$ in *smooth* dyadic ranges $\Psi(|h|/H')$, bound the contribution of $H' > H_2$ using the decay of ψ_0, separate the variables h and qdr_1r_2 variables via a Fourier integral, and fix the integration variable $u \asymp 1$. This produces a smooth factor $\widetilde{\psi}_0(uqdr_1r_2/(QDR^2))$, which we eliminate by partial summation in q, r_1, r_2, leading to the exponential sum $\mathscr{E}'$ below. But first, we need to verify the conditions [May25b, (5.1), (5.2)] from [May25b, Lemma 5.9],

$$Q_1 Q_2 Q_3 \le x^{2/3}, \qquad Q_1 (Q_2 Q_3)^2 \ll M x^{1-2\varepsilon},$$

both of which follow from (4.37). Indeed, recalling that $MN \asymp x$, we have

$$\begin{aligned} Q_1Q_2Q_3 &\leq \left(N^2Q_2Q_3^2\right)^{3/8}\left(N^{\frac{1+\theta}{1-\theta}}Q_1^{\frac{1-3\theta}{1-\theta}}Q_2^5Q_3^2\right)^{1/8} \leq \left(x^{1-9\varepsilon}\right)^{3/8}\left(x^{2-33\varepsilon}\right)^{1/8} < x^{5/8}, \\ Q_1(Q_2Q_3)^2 &\leq \frac{(NQ_2Q_3)^2}{N} \leq \frac{x^{2-18\varepsilon}}{N} \ll \frac{(N^2Q_2Q_3^2)^2}{N} \ll Mx^{1-18\varepsilon}. \end{aligned} \tag{4.38}$$

As in [May25b, p. 27], it remains to bound an exponential sum $\mathscr{E}'$ (roughly corresponding to (4.10), before passing to the complementary divisor of q) by

$$\mathscr{E}' \ll_\varepsilon \frac{N^2}{Q_1x^{\varepsilon/2}}, \tag{4.39}$$

but we now have

$$\mathscr{E}' := \sum_{\substack{Q_1\leq q\leq Q_1' \\ (q,a)=1}} \sum_{\substack{R\leq r_1\leq R_1 \\ R\leq r_2\leq R_2 \\ (r_1,ar_2)=1 \\ (r_2,aqdr_1)=1}} \frac{\lambda_{r_1}\overline{\lambda_{r_2}}}{qdr_1r_2} \sum_{\substack{n_1,n_2\sim N \\ n_1\equiv n_2 \pmod{qd} \\ (n_1,qdr_1n_2)=1 \\ (n_2,qdr_2n_1)=1 \\ (n_1r_2,n_2)\in\mathcal{N}}} \alpha_{n_1}\overline{\alpha_{n_2}} \sum_{h\in\mathbb{Z}} e(h\omega)\,\Psi\left(\frac{|h|}{H'}\right) \times e\left(\frac{ah\overline{n_2qdr_1}(n_1-n_2)}{n_1r_2}\right),$$

where $\Psi : (\frac{1}{2}, 2) \to \mathbb{C}$ is compactly-supported with $\Psi^{(j)} \ll_j 1$, $Q_1' \leq 2Q_1$, $R_1, R_2 \leq 2R$, $H' \leq H_2 = (\log x)^5QDR^2/M$, and

$$\omega := u\frac{M}{QDR^2} \ll x^{o(1)}H_2^{-1} \qquad \Rightarrow \qquad T_{H'}(\omega) \ll x^{o(1)}.$$

All that changed from the sum $\mathscr{E}'$ from [May25b, p. 27] is that h now lies in a smooth dyadic range, which is ultimately required by our large sieve inequality in Proposition 4.8. After expanding (λ_r) and fixing one of $x^{o(1)}$ choices of q_2'', q_3'', Lemma 4.12 gives the desired bound in (4.39) provided that

$$\begin{aligned} N^2{Q_3'}^2Q_2' &\leq x^{1-9\varepsilon}, \\ N^{\frac{2+\theta}{2-2\theta}}Q_3{Q_2'}^{\frac{4-5\theta}{2-2\theta}} &\leq x^{1-17\varepsilon}, \\ N^{\frac{1+\theta}{1-\theta}}Q_1^{\frac{1-3\theta}{1-\theta}}{Q_3'}^2{Q_2'}^5 &\leq x^{2-33\varepsilon}, \end{aligned}$$

for all $Q_2' \leq Q_2$ and $Q_3' \leq Q_3$. These bounds follow directly from (4.37), completing our proof. □

We are now ready to establish a Type II estimate for triply-well-factorable weights, improving [May25b, Proposition 4.1]; recall the triply-well-factorable condition from Definition 4.1.

Remark. The system of inequalities in (4.37) is significantly more flexible than the triply-well-factorable condition from Definition 4.1. In particular, in Section 4.5, we will need to use Proposition 4.14 directly rather than the result below.

Proposition 4.15 (Triply-well-factorable Type II estimate)**.** *Let $a \in \mathbb{Z} \setminus \{0\}$, $A, \varepsilon > 0$, and (λ_q) be triply-well-factorable of level $Q \le x^{5/8-100\varepsilon}$. Let $M, N, x \gg 1$ with $MN \asymp x$ and*

$$x^{\varepsilon} \le N \le x^{3/8}.$$

Let (α_n), (β_m) be divisor-bounded complex sequences, such that (α_n) is supported on $P^-(n) \ge z_0 := z_0 := x^{1/(\log\log x)^3}$ and satisfies the Siegel–Walfisz condition from Definition 4.13. Then for any interval $\mathcal{I} \subset [x, 2x]$, one has

$$\sum_{q \le Q} \lambda_q \sum_{\substack{m \sim M \\ n \sim N \\ mn \in \mathcal{I}}} \alpha_n \beta_m \left(\mathbb{1}_{mn \equiv a \pmod q} - \frac{\mathbb{1}_{(mn,q)=1}}{\varphi(q)} \right) \ll_{\varepsilon, A, a} \frac{x}{(\log x)^A}.$$

Proof. We follow the proof on [May25b, p. 28], reducing to the case of 1-bounded coefficients via [May25b, Lemma 5.1] and separating the variables mn (from the condition $mn \in \mathcal{I}$) via [May25b, Lemma 5.2]. We may assume that $x^{1/2-\varepsilon} \le Q \le x^{5/8-100\varepsilon}$, since the Bombieri–Vinogradov theorem yields the result for $Q \le x^{1/2-\varepsilon}$. The only difference is that we choose the ranges

$$Q_1 := \frac{N}{x^{\varepsilon}}, \qquad Q_2 := \frac{Q^2}{x^{1-21\varepsilon}}, \qquad Q_3 := \frac{x^{1-20\varepsilon}}{NQ},$$

where we note that $Q_i \ge 1$ since $x^{1/2-\varepsilon} \le Q \le x^{5/8-100\varepsilon}$ and $x^{\varepsilon} \le N \le x^{3/8}$.

We claim that any $Q_1' \le Q_1$, $Q_2' \le Q_2$, $Q_3' \le Q_3$ obey the constraints in (4.36); indeed, using $Q \le x^{5/8-100\varepsilon}$, we have

$$N^2 Q_2 Q_3^2 = x^{1-19\varepsilon} \le x^{1-15\varepsilon},$$

$$N^2 Q_2^5 Q_3^2 = \frac{Q^8}{x^{3-65\varepsilon}} \le x^{2-40\varepsilon}.$$

After decomposing the triply-well-factorable weights as in Definition 4.1 and putting q_i in dyadic ranges, Proposition 4.14 yields the result. □

Proof of Theorem 4.2.(i). This is completely analogous to the proof on [May25b, p. 10], decomposing the von Mangoldt function via the Heath-Brown identity [May25b, Lemma 4.3], and noting that $x^{1/3} \le x^{3/8}$. After Proposition 4.15 handles the critical ranges where some M_i or N_i lies in $[x^{\varepsilon}, x^{3/8}]$, [May25b, Lemma 4.4] handles the case of one large smooth factor $N_i > x^{3/8}$, while [May25b, Proposition 4.2] handles the case of two large smooth factors $N_i > x^{3/8}$. □

4.5 Primes with linear sieve weights

Here we work with the upper-bound and lower-bound linear sieve weights, using Proposition 4.14. We first recall some definitions from [May25b; Lic23].

Definition 4.16 (Linear sieve support)**.** For $D \geq 1$, consider the sets of positive integers

$$\begin{aligned}
\mathcal{D}^+(D) &:= \left\{p_1 \cdots p_r : p_1 \geq \cdots \geq p_r \text{ prime}, \, p_1 \cdots p_{j-1} p_j^3 \leq D \text{ for odd } j \leq r\right\}, \\
\mathcal{D}^-(D) &:= \left\{p_1 \cdots p_r : p_1 \geq \cdots \geq p_r \text{ prime}, \, p_1 \cdots p_{j-1} p_j^3 \leq D \text{ for } 2 \mid j \leq r;\, p_1^2 \leq D\right\}, \\
\mathcal{D}^{\mathrm{well}}(D) &:= \left\{p_1 \cdots p_r : p_1 \geq \cdots \geq p_r \text{ prime}, \, p_1 \cdots p_{j-1} p_j^2 \leq D \text{ for } j \leq r\right\},
\end{aligned} \tag{4.40}$$

which satisfy

$$\mathcal{D}^{\pm}(D) \subset \mathcal{D}^{\mathrm{well}}(D).$$

Similarly, for $r \in \mathbb{Z}_+$, we define the sets of vectors

$$\begin{aligned}
\mathbf{D}_r^+(D) &:= \left\{(P_1, \ldots, P_r) : P_1 \geq \cdots \geq P_r \geq 1, \, P_1 \cdots P_{j-1} P_j^3 \leq D \text{ for odd } j \leq r\right\}, \\
\mathbf{D}_r^-(D) &:= \left\{(P_1, \ldots, P_r) : P_1 \geq \cdots \geq P_r \geq 1, \, P_1 \cdots P_{j-1} P_j^3 \leq D \text{ for } 2 \mid j \leq r;\, P_1^2 \leq D\right\}, \\
\mathbf{D}_r^{\mathrm{well}}(D) &:= \left\{(P_1, \ldots, P_r) : P_1 \geq \cdots \geq P_r \geq 1, \, P_1 \cdots P_{j-1} P_j^2 \leq D \text{ for } j \leq r\right\},
\end{aligned} \tag{4.41}$$

which have

$$\mathbf{D}_r^{\pm}(D) \subset \mathbf{D}_r^{\mathrm{well}}(D).$$

The *standard* upper-bound $(+)$ and lower-bound $(-)$ linear sieve weights of level D are given by

$$\lambda_d^{\pm} := \mu(d) \cdot \mathbb{1}_{d \in \mathcal{D}^{\pm}(D)}.$$

There is also a *well-factorable* variant $\widetilde{\lambda}_d^{\pm}$ of these weights due to Iwaniec (see [FI10, Chapter 12.7], [Iwa80], [May25b, Chapter 8]), which produces results of essentially the same strength in sieve problems. Given small parameters $\upsilon, \eta > 0$, $\widetilde{\lambda}_d^{\pm} = \widetilde{\lambda}_d^{\pm}(\upsilon, \eta)$ is a sum of $O_{\upsilon,\eta}(1)$ sequences of the form

$$\widetilde{\lambda}_{d,\vec{P}}^{\pm} := \begin{cases} (-1)^r, & d = p_1 \cdots p_r,\ p_j \in (P_j, P_j^{1+\eta}] \text{ primes}, \\ 0, & \text{otherwise}, \end{cases}$$

where $\vec{P} = (P_1, \ldots, P_r) \in \mathbf{D}_r^{\pm}(D^{1/(1+\eta)})$ and P_i are part of the sequence $(D^{\upsilon(1+\eta)^j})_{j \geq 0}$. Each such sequence is supported on $d \in \mathcal{D}^{\pm}(D)$, and *well-factorable* in the sense that for *any* choice of $D_1 D_2 = D$ with $D_i \geq 1$, one can write

$$\widetilde{\lambda}_{d,\vec{P}}^{\pm} = \sum_{d_1 d_2 = d} \alpha_{d_1} \beta_{d_2},$$

for some 1-bounded sequences (α_{d_1}), (β_{d_2}) supported on $d_i \leq D_i$. This is inherited from the fact that every $d \in \mathcal{D}^{\mathrm{well}}(D)$ can be greedily factorized as $d = d_1 d_2$, for some positive integers $d_i \leq D_i$.

However, the weights $\widetilde{\lambda}_d^{\pm}$ are not *triply-well-factorable* in the sense of Definition 4.1, since not every $d \in \mathcal{D}^{\pm}(D)$ can be factorized as $d = d_1 d_2 d_3$ with $d_i \leq D_i$, given *any* choice of $D_1 D_2 D_3 = D$. Fortunately, to apply our Proposition 4.14, it will suffice to use a particular choice of D_1, D_2, D_3 which obey the system in (4.36). Specifically, following [May25b; Lic23], it will be enough to show that every modulus d of interest has a factorization $d = d_1 d_2 d_3$ into positive integers obeying the system

$$\begin{aligned} d_1 &\leq \frac{N}{x^{\delta}}, \\ N^2 d_2 d_3^2 &\leq x^{1-\delta}, \\ N^2 d_2^5 d_3^2 &\leq x^{2-\delta}, \end{aligned} \tag{4.42}$$

for some small $\delta > 0$ (compare this with (4.36)).

4.5.1 The upper-bound linear sieve weights

Here we deduce Theorem 4.2.(ii) for the upper-bound well-factorable linear sieve weights $\widetilde{\lambda}_d^{+}(\upsilon, \eta)$, where υ is chosen to be sufficiently small in terms of ε, and η is sufficiently small in terms of ε, υ (if one allows arbitrarily small values of υ, η, the implicit constant in (4.2) should also depend on υ, η). Our key factorization result is the following.

Proposition 4.17 (Factorization in the upper-bound linear sieve support)**.** *Let $0 < \delta < 10^{-5}$, $D = x^{3/5-50\delta}$, $x^{2\delta} \leq N \leq x^{1/3+\delta}$, and $d \in \mathcal{D}^{+}(D)$. Then there exists a factorization $d = d_1 d_2 d_3$ into positive integers obeying (4.42).*

Remark. The level $D = x^{3/5-o(1)}$ in Proposition 4.17 is optimal, as seen by taking $N = x^{1/5}$ and $p_1 \approx p_2 \approx x^{1/5}$. There are various other limiting cases, but essentially all situations where $D \leq x^{3/5-o(1)}$ can be handled by an interpolation of the ranges

$$d_1 \leq N x^{-o(1)}, \qquad d_2 \leq x^{1/5-o(1)}, \qquad d_3 \leq \frac{x^{2/5+o(1)}}{N},$$

and

$$d_1 \leq N x^{-o(1)}, \qquad d_2 \leq x^{4/15-o(1)}, \qquad d_3 \leq \frac{x^{1/3+o(1)}}{N},$$

both of which are acceptable in (4.42) (up to a good choice of the $o(1)$ exponents in terms of δ).

Proof of Theorem 4.2.(ii) assuming Proposition 4.17. This follows analogously as in [May25b, Chapter 8], using Proposition 4.14 instead of [May25b, Proposition 7.2], and Proposition 4.17 instead of [May25b, Proposition 8.1]. □

Our proof of Proposition 4.17 is structured differently from Maynard's computations in [May25b, Chapter 8], to accommodate a new range of limiting cases. We start with a preliminary result concerning the greatest 6 prime factors of the elements of $\mathcal{D}^+(D)$.

Lemma 4.18 (Placing the first 6 prime factors). *Let $0 < \delta < 10^{-5}$, $D = x^{3/5-50\delta}$, $x^{2\delta} \le N \le x^{1/3+\delta}$, and $d \in \mathcal{D}^+(D)$ have 6 prime factors, counting multiplicities. Then there exists a factorization $d = d_1d_2d_3$ into positive integers such that*

$$d_1 \le D_1 := Nx^{-\delta}, \qquad d_2 \le D_2 := x^{4/15-5\delta}, \qquad d_3 \le D_3 := \frac{x^{2/5+2\delta}}{N}. \tag{4.43}$$

Remark. The values of D_1, D_2, D_3 in (4.43) do *not* multiply up to D, and do not (yet) obey the conditions on Q_1, Q_2, Q_3 from Proposition 4.14. So Lemma 4.18 does not directly imply Proposition 4.17 (not even for integers with 6 primes factors), but it will be a crucial step in its proof.

Proof of Lemma 4.18. We first observe that $D_1, D_2, D_3 \ge x^\delta$ by the conditions on N and δ. Now let $d \in \mathcal{D}^+(D)$ have prime factorization $d = p_1 \cdots p_6$, with $p_1 \ge \cdots \ge p_6$; in particular, we have

$$p_1^3 \le D, \qquad p_1p_2p_3^3 \le D, \qquad p_1p_2p_3p_4p_5^3 \le D.$$

We claim that it is impossible for the following system of inequalities to hold true simultaneously:

$$\begin{aligned} p_1p_4p_5 &> D_2, \\ p_2p_3p_5 &> D_2, \\ p_1p_2p_3p_4p_5^4 &> D_1D_2D_3. \end{aligned}$$

Indeed, by multiplying all three inequalities, one would obtain

$$x^{6/5-14\delta} = D_1D_2^3D_3 < (p_1p_2p_3p_4p_5^3)^2 \le D^2,$$

which is a contradiction. Thus at least one of these inequalities fails, which leads us to three cases.

Case 1: $p_1p_4p_5 \le D_2$. Then, we fix $d_2 := p_1p_4p_5$. We will construct $d_1 \le D_1$ and $d_3 \le D_3$ such that $d_1d_3 = p_2p_3p_6$; for now, we set $d_1 = d_3 := 1$. Since

$$p_2^2 \le p_1^2 \le D^{2/3} \le x^{2/5} \le D_1D_3,$$

we must have $p_2 \le D_1$ or $p_2 \le D_3$; we set $d_1 \leftarrow p_2$ if $p_2 \le D_1$, and $d_3 \leftarrow p_2$ otherwise. We also have

$$p_2p_3^2 = \left(p_2p_3^4\right)^{1/2} p_2^{1/2} \le p_1^{1/2} \left(p_1p_2p_3^3\right)^{1/2} \le x^{1/10+3/10} \le D_1D_3,$$

so $p_3 \le \sqrt{D_1 D_3/(d_1 d_3)}$, which forces $p_3 \le D_1/d_1$ or $p_3 \le D_3/d_3$; we set $d_1 \leftarrow d_1 p_3$ if $p_3 \le D_1/d_1$, and $d_3 \leftarrow d_3 p_3$ otherwise. Finally, we note that

$$p_2 p_3 p_6^2 = p_2^{1/2} \left(p_2 p_3^2 p_6^4\right)^{1/2} \le p_1^{1/2} \left(p_1 p_2 p_3 p_4 p_5^3\right)^{1/2} \le x^{1/10+3/10} \le D_1 D_3,$$

so $p_6 \le \sqrt{D_1 D_3/(d_1 d_3)}$, which forces $p_6 \le D_1/d_1$ or $p_6 \le D_3/d_3$; we set $d_1 \leftarrow d_1 p_6$ if $p_6 \le D_1/d_1$, and $d_3 \leftarrow d_3 p_6$ otherwise. At this point, we have $d_1 d_2 d_3 = p_1 \cdots p_6$ and $d_i \le D_i$ for $i \in \{1,2,3\}$, as we wanted.

Case 2: $p_2 p_3 p_5 \le D_2$. Then, we fix $d_2 := p_2 p_3 p_5$, and construct $d_1 \le D_1, d_3 \le D_3$ such that $d_1 d_3 = p_1 p_4 p_6$. The process is completely analogous to the previous case (with p_1, p_4 taking the places of p_2, p_3), using the bounds

$$\begin{aligned} p_1 &\le D^{2/3} \le x^{2/5} \le D_1 D_3, \\ p_1 p_4^2 &= \left(p_1 p_4^4\right)^{1/2} p_1^{1/2} \le p_1^{1/2} \left(p_1 p_2 p_3^3\right)^{1/2} \le x^{1/10+3/10} \le D_1 D_3, \\ p_1 p_4 p_6^2 &= p_1^{1/2} \left(p_1 p_4^2 p_6^4\right)^{1/2} \le p_1^{1/2} \left(p_1 p_2 p_3 p_4 p_5^3\right)^{1/2} \le x^{1/10+3/10} \le D_1 D_3. \end{aligned} \tag{4.44}$$

Case 3: $p_1 p_2 p_3 p_4 p_5^4 \le D_1 D_2 D_3$. We can also assume without loss of generality that we are not in the previous case, so $p_2 p_3 p_5 > D_2$. Then, we fix $d_2 := p_2 p_3$, noting that

$$p_2 p_3 \le p_1^{1/3} (p_1 p_2 p_3^3)^{1/3} \le D^{4/9} \le x^{4/15-5\delta} = D_2.$$

We will construct $d_1 \le D_1$ and $d_3 \le D_3$ such that $d_1 d_3 = p_1 p_4 p_5 p_6$. We start by placing the primes p_1, p_4, p_6 into d_1 and d_3 exactly as in the previous case, using the bounds in (4.44).

At this point, we have $d_1 d_3 = p_1 p_4 p_6$, and it remains to place p_5. Since

$$p_1 p_4 p_6 p_5^2 \le p_1 p_4 p_5^3 = \frac{p_1 p_2 p_3 p_4 p_5^4}{p_2 p_3 p_5} \le \frac{D_1 D_2 D_3}{D_2} = D_1 D_3,$$

we have $p_5 \le \sqrt{D_1 D_3/(d_1 d_3)}$, which forces $p_5 \le D_1/d_1$ or $p_5 \le D_3/d_3$. Then we are done by setting $d_i \leftarrow d_i p_5$ for some $i \in \{1,3\}$. □

We can now prove Proposition 4.17, thus completing the proof of Theorem 4.2.

Proof of Proposition 4.17. Let $d \in \mathcal{D}^+(D)$ have prime factorization $d = p_1 \cdots p_r$ with $p_1 \ge p_2 \ge \ldots$. We recall from (4.40) that this implies

$$p_1 \cdots p_{2j} p_{2j+1}^3 \le D \qquad \text{and} \qquad p_1 \cdots p_{k-1} p_k^2 \le D,$$

for all $0 \le j < r/2$ and $1 \le k \le r$. We aim to place the primes $p_1, \ldots, p_r$ into d_1, d_2, d_3 such that the bounds in (4.42) hold.

We begin by setting $d_1 = d_2 = d_3 := 1$, and perform the following iterative greedy process. At step $1 \le j \le r$, we do the following:

$j.(i)$. If $j \le 6$, we place p_j in the corresponding factor d_i from the factorization in Lemma 4.18 (i.e., we set $d_i \leftarrow d_i p_j$). If $j \ge 7$, we act greedily and place p_j into any factor d_i such that after substituting $d_i \leftarrow d_i p_j$, we have the same system of inequalities

$$d_1 \le N x^{-\delta}, \qquad d_2 \le x^{4/15-5\delta}, \qquad d_3 \le \frac{x^{2/5+2\delta}}{N},$$

as in (4.43). We terminate unsuccessfully if this is impossible.

$j.(ii)$. Having placed p_j into some d_i, we check whether either of the lower bounds

$$d_2 > x^{1/5-5\delta}, \qquad \text{or} \qquad d_3 > \frac{x^{1/3+2\delta}}{N},$$

holds; if so, we terminate unsuccessfully. Otherwise, we continue with step $j+1$ (or terminate successfully if $j = r$).

Case 1: The process terminates successfully; this is the easier case.

We are left with a factorization $d = d_1 d_2 d_3$ satisfying

$$d_1 \le N x^{-\delta}, \qquad d_2 \le x^{1/5-5\delta}, \qquad d_3 \le \frac{x^{1/3+2\delta}}{N},$$

which actually forces $d \le x^{8/15-4\delta}$ (significantly smaller than $D = x^{9/15-50\delta}$). Then we can verify the conditions in (4.42), with room to spare:

$$N^2 d_2 d_3^2 \le x^{2(1/3+2\delta)+(1/5-5\delta)} < x^{1-\delta},$$

$$N^2 d_2^5 d_3^2 \le x^{2(1/3+2\delta)+(1-25\delta)} \le x^{2-\delta}.$$

Case 2: The process terminates unsuccessfully in substep $j.(i)$; we show that this cannot happen.

Indeed, we must have $j \ge 7$ since Lemma 4.18 handles all $j \le 6$. We are left with a factorization $p_1 \cdots p_{j-1} = d_1 d_2 d_3$, which must satisfy

$$d_2 p_j > x^{4/15-5\delta},$$

in order to terminate in substep $j.(i)$. Moreover, since we did not terminate in substep $(j-1).(ii)$, we must have

$$d_2 \le x^{1/5-5\delta}.$$

But since $j \ge 7$, we have

$$\frac{x^{4/15-5\delta}}{x^{1/5-5\delta}} < \frac{d_2 p_j}{d_2} \le p_7 \le \left(p_1 \cdots p_6 p_7^3\right)^{1/9} \le D^{1/9} < x^{1/15},$$

which gives a contradiction.

Case 3: The process terminates unsuccessfully in substep $j.(ii)$; this is the main case.

We are left with a factorization $p_1 \cdots p_j = d_1 d_2 d_3$ satisfying

$$d_1 \le N x^{-\delta}, \qquad d_2 \le x^{4/15-5\delta}, \qquad d_3 \le \frac{x^{2/5+2\delta}}{N}, \tag{4.45}$$

and *either* $d_2 > x^{1/5-5\delta}$ *or* $d_3 > x^{1/3+2\delta}/N$ (we cannot have both, since we should have terminated in a previous substep (ii) in that case; note that both of these bounds fail at the very beginning of the greedy process because $N \le x^{1/3+\delta}$, and that only one d_i gets updated in each substep (i)).

Case 3.1: One has $d_2 > x^{1/5-5\delta}$ and $d_3 \le x^{1/3+2\delta}/N$. In this case, we set $D_1 := Nx^{-\delta}$, $D_2 := d_2$, and $D_3 := x^{3/5-3\delta}/(Nd_2)$, which have $D_i \ge d_i$ (in light of (4.45)) and $D_1 D_2 D_3 \ge D$. We then run a greedy process to place the remaining primes p_k with $k \ge j+1$ into either d_1 or d_3, while preserving the inequalities $d_i \le D_i$. This works because at step k, before placing p_k, we have

$$p_1 \cdots p_{k-1} p_k^2 \le D \qquad \Rightarrow \qquad p_k^2 \le \frac{D}{d_1 d_2 d_3} \le \frac{D_1 D_3}{d_1 d_3},$$

so $p_k \le \max(D_1/d_1, D_3/d_3)$ (i.e., there is "enough room" for p_k in d_1 or d_3). In the end, we have $d = d_1 d_2 d_3$ with $d_i \le D_i$, and we can verify the bounds in (4.42) using $x^{1/5-5\delta} < d_2 \le x^{4/15-5\delta}$:

$$N^2 d_2 d_3^2 \le N^2 d_2 \left(\frac{x^{3/5-3\delta}}{Nd_2}\right)^2 = \frac{x^{6/5-6\delta}}{d_2} < x^{1-\delta},$$

$$N^2 d_2^5 d_3^2 \le N^2 d_2^5 \left(\frac{x^{3/5-3\delta}}{Nd_2}\right)^2 = d_2^3 x^{6/5-6\delta} \le x^{2-\delta}.$$

Case 3.2: One has $d_2 \le x^{1/5-5\delta}$ and $d_3 > x^{1/3+2\delta}/N$. Then we set $D_1 := Nx^{-\delta}$, $D_2 := x^{3/5-3\delta}/(Nd_3)$, and $D_3 := d_3$, which have $D_i \ge d_i$ (in light of (4.45)) and $D_1 D_2 D_3 \ge D$. We run a similar greedy process to place the remaining primes p_k with $k \ge j+1$ into either d_1 or d_2, while preserving the bounds $d_i \le D_i$. This works because at step k, before placing p_k, we have

$$p_1 \cdots p_{k-1} p_k^2 \le D \qquad \Rightarrow \qquad p_k^2 \le \frac{D}{d_1 d_2 d_3} \le \frac{D_1 D_2}{d_1 d_2},$$

so $p_k \le \max(\frac{D_1}{d_1}, \frac{D_2}{d_2})$. In the end, we have $d = d_1 d_2 d_3$ with $d_i \le D_i$, and we can verify the bounds in (4.42) using $x^{1/3+2\delta}/N < d_3 \le x^{2/5+2\delta}/N$:

$$N^2 d_2 d_3^2 \le N^2 \frac{x^{3/5-3\delta}}{Nd_3} d_3^2 = N d_3\, x^{3/5-3\delta} \le x^{1-\delta},$$

$$N^2 d_2^5 d_3^2 \le N^2 \left(\frac{x^{3/5-3\delta}}{Nd_3}\right)^5 d_3^2 = \frac{x^{3-15\delta}}{(Nd_3)^3} \le x^{2-\delta}.$$

We have now covered all cases. □

4.5.2 The linear sieve weights with special factors

From the work of Bombieri–Friedlander–Iwaniec [BFI86, Theorem 10] and our work in the previous subsection, we have exponents of distribution of $\frac{4}{7}-\varepsilon$ and $\frac{3}{5}-\varepsilon$ for the weights $\widetilde{\lambda}_d^-$ and $\widetilde{\lambda}_d^+$, respectively. Here we obtain a better level of distribution for $\widetilde{\lambda}_d^\pm$, up to $\frac{5}{8}-\varepsilon$, when more information about the factorization of d is available.

We adapt the computations from [Lic23, Section 6] using our Proposition 4.14 instead of [Lic23, Proposition 5.2]. More precisely, our Theorem 4.19, Propositions 4.20 and 4.24, and Lemmas 4.22 and 4.23 correspond respectively to [Lic23, Proposition 6.6, Proposition 6.1, Lemma 6.3, Lemma 6.4, Proposition 6.5]; additionally, we use Lemma 4.21 to fix a small error in the argument from [Lic23, Section 6]. For $t \geq 0$, we let

$$\vartheta(t) := \min\left(\frac{1+t}{2}, \frac{2-3t}{2}\right) \in \left[\frac{1}{2}, \frac{5}{8}\right], \tag{4.46}$$

which achieves its maximum (only) at $t = \frac{1}{4}$. For $\frac{1}{4} \geq t_1 \geq t_2 \geq t_3 \geq 0$ and $\delta > 0$, we also define

$$\begin{aligned}\vartheta(t_1,t_2,t_3) := \max\big\{&\vartheta(t_1), \vartheta(t_2), \vartheta(t_1+t_2), \vartheta(t_1+t_2+t_3), w(t_1,t_2,t_3), w(t_2,t_1,t_3),\\ &\psi(\vartheta(t_1+t_3), t_1+2t_2+t_3), \psi(\vartheta(t_2+t_3), 2t_1+t_2+t_3)\big\},\end{aligned} \tag{4.47}$$

where $\psi(x,y) := x\mathbb{1}_{x \geq y+2\delta}$ and

$$w(t_1,t_2,t_3) = \psi\left(\min\left\{\frac{5-3t_3}{8}, 1-2t_2-2\delta\right\}, \frac{1+t_1}{2}\right).$$

Our main result in this subsection, which will imply Corollary 4.3, is the following.

Theorem 4.19 (Primes in APs with linear sieve weights)**.** *Let $\varepsilon = 10\delta > 0$ be sufficiently small and $A > 0$. Let $(P_1, \dots, P_r) \in \mathbf{D}_r^{\mathrm{well}}(D)$ with $1 \leq D = x^{\vartheta_0}$ and $P_i = x^{t_i}$, where t_i are part of the sequence $(\varepsilon^2(1+\varepsilon^9)^j)_{j\geq 1}$. Then provided that $\vartheta_0 \leq \vartheta(t_1) - \varepsilon$, for any choice of the sign $\pm$, we have*

$$\sum_{\substack{b=p_1\cdots p_r\\ D_i<p_i\leq D_i^{1+\varepsilon^9}}} \sum_{\substack{d=bc\leq D\\ c\mid P(p_r)\\ (d,a)=1}} \widetilde{\lambda}^\pm(d)\left(\pi(x;d,a) - \frac{\pi(x)}{\varphi(d)}\right) \ll_{a,A,\varepsilon} \frac{x}{(\log x)^A}.$$

Moreover, if $t_1 \leq \frac{1}{4}$ and $r \geq 3$, then the same holds provided that $\vartheta_0 \leq \vartheta(t_1,t_2,t_3) - \varepsilon$.

Much like Theorem 4.2.(*ii*) stems from the factorization result in Proposition 4.17, Theorem 4.19 depends on the factorization result below.

Proposition 4.20 (Factorization in the well-factorable support)**.** *Let* $0 < \delta < 10^{-5}$, $1 \le D = x^{\vartheta_0}$, $x^{2\delta} \le N \le x^{1/3+\delta}$, *and* $d \in \mathcal{D}^{\mathrm{well}}(D)$. *Write* $d = p_1 \cdots p_r$ *where* $p_1 \ge \cdots \ge p_r$ *are primes with* $p_i = x^{t_i}$. *Then there exists a factorization* $d = d_1 d_2 d_3$ *into positive integers obeying* (4.42), *provided that* $\vartheta_0 \le \vartheta(t_1) - 2\delta$, *as in* (4.46).

Moreover, if $t_1 \le \frac{1}{4}$ *and* $r \ge 3$, *then it suffices that* $\vartheta_0 \le \vartheta(t_1, t_2, t_3) - 2\delta$, *as in* (4.47).

Proof of Theorem 4.19 assuming Proposition 4.20. This is almost identical to [Lic25, Proof of Proposition 5.4], using Proposition 4.14 instead of [Lic25, Theorem 2.5], and Proposition 4.20 instead of [Lic25, Proposition 3.3]. □

To prove Proposition 4.20, we will need a few lemmas.

Lemma 4.21 (Two-factor greedy algorithm)**.** *Let* $0 < \delta < 10^{-5}$, $\vartheta \in [\frac{1}{2}, \frac{5}{8}]$, $x^{2\delta} \le N \le x^{1/3+\delta}$, $1 \le D \le x^{\vartheta-2\delta}$, *and suppose that*

$$x^{\frac{5\vartheta-2}{3}-\delta} \le N.$$

Then, any $d \in \mathcal{D}^{\mathrm{well}}(D)$ *has a factorization* $d = d_1 d_2 d_3$ *into positive integers satisfying* (4.42).

Proof. Let $(D_1, D_2) := (Nx^{-\delta}, x^{\vartheta-\delta}/N)$, so that $D \le D_1 D_2$ and $D_i \ge 1$. Let $d_1 = d_2 = 1$ and $d = p_1 \cdots p_r$, where $p_1 \ge \cdots \ge p_r$ are primes. We will run a greedy algorithm to append each of these primes to one of d_1 or d_2, while preserving the bounds $d_1 \le D_1$, $d_2 \le D_2$. At step $j \ge 1$, the definition of $\mathcal{D}^{\mathrm{well}}(D)$ implies

$$p_j^2 = \frac{p_1 \cdots p_{j-1} p_j^2}{p_1 \cdots p_{j-1}} \le \frac{D}{p_1 \cdots p_{j-1}} \le \frac{D_1 D_2}{d_1 d_2},$$

so $p_j \le \max(D_1/d_1, D_2/d_2)$, and we can append p_j to one of d_1 and d_2. In the end, we take $d_3 = 1$, and obtain $d = d_1 d_2 d_3$ with $d_1 \le D_1$, $d_2 \le D_2$. To verify the system (4.42), we write

$$d_1 \le D_1 = \frac{N}{x^\delta},$$

$$N^2 d_2 d_3^2 \le N^2 D_2 = N x^{\vartheta-\delta} \le x^{1/3+\delta+2/3-2\delta} = x^{1-\delta},$$

and, using the hypothesis in the form $D_2 = x^{\vartheta-\delta}/N \le x^{(2-2\vartheta)/3}$,

$$N^2 d_2^5 d_3^2 \le (N D_2)^2 D_2^3 \le x^{2(\vartheta-\delta)} x^{2-2\vartheta} \le x^{2-\delta},$$

as required. □

Lemma 4.22 (General factorization criterion)**.** *Let* $0 < \delta < 10^{-5}$, $\vartheta \in [\frac{1}{2}, \frac{5}{8}]$, $x^{2\delta} \le N \le x^{1/3+\delta}$, $1 \le D \le x^{\vartheta-2\delta}$, *and*

$$v := 2\vartheta - 1 \quad < \quad \frac{1}{4} \quad < \quad u := \frac{2-2\vartheta}{3}. \tag{4.48}$$

Suppose $d \in \mathcal{D}^{\mathrm{well}}(D)$ *has a factorization* $d = d_1 d_2 d_3$ *into positive integers satisfying*

$$d_1 \le \frac{N}{x^\delta}, \qquad d_2 \in [x^v, x^u], \qquad d_3 \le \max\left(1, \frac{x^{\vartheta-\delta}}{d_2 N}\right).$$

Then, d *has a (potentially different) factorization obeying the system* (4.42).

Proof. First, if $\max(1, \frac{x^{\vartheta-\delta}}{d_2 N}) = 1$, then $x^{\vartheta-\delta} \le d_2 N \le x^u N$, and we can apply Lemma 4.21. Otherwise, we may assume $d_2 d_3 \le x^{\vartheta-\delta}/N$, and we will use the given factorization $d = d_1 d_2 d_3$.

The first bound in (4.42) is reiterated in the hypothesis here. For the second and third bounds, note that

$$N^2 d_2 d_3^2 = \frac{(N d_2 d_3)^2}{d_2} \le \frac{x^{2\vartheta-\delta}}{x^v} = x^{2\vartheta-\delta-2\vartheta+1} = x^{1-\delta},$$

$$N^2 d_2^5 d_3^2 = (N d_2 d_3)^2 d_2^3 \le x^{2\vartheta-\delta} x^{3u} = x^{2\vartheta-\delta+2-2\vartheta} = x^{2-\delta}.$$

□

Remark. When $\vartheta = \frac{3}{5}$, we have $v = \frac{1}{5}$ and $u = \frac{4}{15}$, which gave a relevant interval for the construction of d_2 in Section 4.5.1.

Lemma 4.23 (Three-factor greedy algorithm)**.** *Let* $\delta, \vartheta, u, v, N$ *be as in Lemma 4.22,* $r \ge 3$, $1 \le D \le x^{\vartheta-2\delta}$, *and* $d \in \mathcal{D}^{\mathrm{well}}(D)$. *Write* $d = p_1 \cdots p_r$ *where* $p_1 \ge \cdots \ge p_r$ *are primes. Suppose that* $p_3 \le x^{u-v}$. *Also, assume that* ($p_1 \le x^v$, $p_2^2 \le x^{1-\vartheta-2\delta}$) *or* ($p_2 \le x^v$, $p_1^2 \le x^{1-\vartheta-2\delta}$). *Then there exists a factorization* $d = d_1 d_2 d_3$ *into positive integers satisfying* (4.42).

Proof. Let $(D_1, D_2, D_3) := (Nx^{-\delta}, x^v, x^{1-\vartheta-\delta}/N)$, so that $D \le D_1 D_2 D_3$ and $D_i \ge 1$. Note that any tuple (d_1, d_2, d_3) with $d_i \le D_i$ satisfies (4.42), since

$$N^2 D_2 D_3^2 = x^{v+2-2\vartheta-2\delta} = x^{1-2\delta},$$

and using $\vartheta \le \frac{5}{8}$,

$$N^2 D_2^5 D_3^2 = x^{5v+2-2\vartheta-2\delta} = x^{8\vartheta-3-2\delta} \le x^{2-2\delta}.$$

By assumption, we have ($p_1 \le D_2$ and $p_2^2 \le D_1 D_3$) or ($p_2 \le D_2$ and $p_1^2 \le D_1 D_3$), so for some choice $\{d_1, d_2, d_3\} = \{1, p_1, p_2\}$ we must have $d_i \le D_i$ for all i. We keep

this choice, and run a greedy algorithm to append the primes p_j, for $j \geq 3$, to one of d_1, d_2, d_3 (i.e., $d_i \leftarrow d_i p_j$), while preserving the bounds $d_i \leq D_i$. If this algorithm terminates after appending all primes $p_3, \dots, p_r$, then we obtain a factorization $d = d_1 d_2 d_3$ which satisfies $d_i \leq D_i$, and thus also (4.42).

Otherwise, there must be some index $3 \leq j \leq r$ such that the prime p_j cannot be appended to any d_i, where $d_1 d_2 d_3 = p_1 \cdots p_{j-1}$; thus we have $d_i p_j > D_i$ for all i. By our assumption, we thus have

$$x^v = D_2 < d_2 p_j \leq D_2 p_3 \leq x^v x^{u-v} = x^u,$$

so $d_2' := d_2 p_j \in [x^v, x^u]$, and

$$D_1 < d_1 p_j = \frac{d_1 d_2 d_3 p_j^2}{d_2 d_3 p_j} \leq \frac{D}{d_2 d_3 p_j} \leq \frac{D_1 D_2 D_3}{d_2' d_3},$$

so $D_3' := D_2 D_3 / d_2' \geq d_3$. By the definition of $\mathcal{D}^{\mathrm{well}}(D)$ we have that for each $k > j$,

$$p_k^2 \leq \frac{D}{p_1 \cdots p_{k-1}} \leq \frac{D_1 D_2 D_3}{d_1 d_2 d_3 p_j \cdots p_{k-1}} = \frac{D_1 D_3'}{d_1 d_3 p_{j+1} \cdots p_{k-1}}.$$

Using this bound, we can greedily construct a factorization $d_1' d_3' = d_1 d_3 p_{j+1} \cdots p_r$ (starting from $d_1' = d_1$, $d_3' = d_3$ and appending each p_k at a time) such that $d_1' \leq D_1$ and $d_3' \leq D_3'$. Therefore, we have $d_1' d_2' d_3' = d$ and

$$d_1' \leq D_1 = \frac{N}{x^\delta}, \qquad d_2' = d_2 p_j \in [x^v, x^u], \qquad d_2' d_3' \leq d_2' D_3' = D_2 D_3 = \frac{x^{\vartheta - \delta}}{N}.$$

By Lemma 4.22, we conclude that d_1', d_2', d_3' satisfy (4.42). □

Proposition 4.24 (Factorization depending on the anatomy)**.** *Let $\delta, \vartheta, u, v, N$ be as in Lemma 4.22, $1 \leq D \leq x^{\vartheta - 2\delta}$, and $d \in \mathcal{D}^{\mathrm{well}}(D)$. Write $d = p_1 \cdots p_r$ where $p_1 \geq \cdots \geq p_r$ are primes. Assume that $p_1 \leq x^u$, and that one of the following holds (statements involving p_j implicitly assume $r \geq j$):*

(i). $d_2 \in [x^v, x^u]$ for some $d_2 \in \{p_1, p_2, p_1 p_2, p_1 p_2 p_3\}$;

(ii). $d_2 := p_1 p_3 \in [x^v, x^u]$ and $p_2^2 \leq x^{\vartheta - 2\delta}/d_2$;

(iii). $d_2 := p_2 p_3 \in [x^v, x^u]$ and $p_1^2 \leq x^{\vartheta - 2\delta}/d_2$;

(iv). $p_3 \leq x^{u-v}$, and ($p_1 \leq x^v$, $p_2^2 \leq x^{1-\vartheta-2\delta}$) or ($p_2 \leq x^v$, $p_1^2 \leq x^{1-\vartheta-2\delta}$).

Then there exists a factorization $d = d_1 d_2 d_3$ into positive integers satisfying (4.42).

Proof. Assuming (iv), the conclusion follows immediately from Lemma 4.23. So let us assume that one of (i), (ii), (iii) holds. Let $D_1 := Nx^{-\delta}$, d_2 be the corresponding value from (i), (ii), or (iii) (say, the first of these that holds), and

$$D_3 := \max\left(1, \frac{x^{\vartheta-\delta}}{Nd_2}\right).$$

Note that $D \le D_1 d_2 D_3$ and $D_1, D_3 \ge 1$. If we can find a factorization $d = d_1 d_2 d_3$ with $d_1 \le D_1$ and $d_3 \le D_3$, then Lemma 4.22 will complete the proof.

Suppose for a start that $d_2 = p_1 \cdots p_i \in [x^v, x^u]$ for some $i \in \{1, 2, 3\}$ as in (i). For each $j \in \{i+1, \cdots, r\}$, by the definition of $\mathcal{D}^{\mathrm{well}}$ we have

$$p_j^2 \le \frac{D}{p_1 \cdots p_{j-1}} \le \frac{D_1 D_3}{p_{i+1} \cdots p_{j-1}}.$$

Using this bound, we can greedily construct a factorization $d_1 d_3 = p_{i+1} \cdots p_r$ (starting from $d_1 = 1$, $d_3 = 1$ and appending each p_j at a time) such that $d_1 \le D_1$ and $d_3 \le D_3$, so we are done.

Otherwise, we have $p_1, p_1p_2, p_1p_2p_3 \notin [x^v, x^u]$; in particular, $p_1 \le x^u$ and $p_1 \notin [x^v, x^u]$ imply $p_1 < x^v$, so $p_2 < x^v$ as well; thus (i) cannot hold.

- If (ii) holds, so $d_2 := p_1 p_3 \in [x^v, x^u]$ and $p_2^2 \le x^{\vartheta-2\delta}/d_2 \le D_1 D_3$, then we have $p_2 \le D_1$ or $p_2 \le D_3$.
- If (iii) holds, so $d_2 := p_2 p_3 \in [x^v, x^u]$ and $p_1^2 \le x^{\vartheta-2\delta}/d_2 \le D_1 D_3$, then we have $p_1 \le D_1$ or $p_1 \le D_3$.

In either case, we can factor $p_1p_2p_3 = d_1d_2d_3$ where $d_1 \le D_1$ and $d_3 \le D_3$. Then for each $j \in \{4, \dots, r\}$ (if any), we have

$$p_j^2 \le \frac{D}{p_1 \cdots p_{j-1}} \le \frac{D_1 d_2 D_3}{p_1 \cdots p_{j-1}} \le \frac{(D_1/d_1)(D_3/d_3)}{p_4 \cdots p_{j-1}},$$

so we can greedily append p_j to one of d_1 and d_3 until we have $d = d_1 d_2 d_3$ with $d_1 \le D_1$, $d_3 \le D_3$. □

Proof of Proposition 4.20. Let $d = p_1 \cdots p_r$ where $p_1 \ge \cdots \ge p_r$ are primes with $p_i = x^{t_i}$. We want to show that d has a factorization obeying (4.42), under one of the following assumptions:

(a). $d \in \mathcal{D}^{\mathrm{well}}(x^{\vartheta(t_1)-2\delta})$ where $\vartheta(t_1)$ is as in (4.46), or

(b). $t_1 \le \frac{1}{4}$, $r \ge 3$, and $d \in \mathcal{D}^{\mathrm{well}}(x^{\vartheta(t_1,t_2,t_3)-2\delta})$, where $\vartheta(t_1, t_2, t_3)$ is as in (4.47).

Applying Proposition 4.24 for some $\vartheta \in [\frac{1}{2}, \frac{5}{8}]$ and letting $u = u(\vartheta)$, $v = v(\vartheta)$ be as in (4.48), we deduce that d has a factorization obeying (4.42) provided that $d \in \mathcal{D}^{\mathrm{well}}(x^{\vartheta-2\delta})$, $t_1 \le u$, and that one of the following holds:

(*i*). $t \in [v, u]$ for some $t \in \{t_1, t_2, t_1 + t_2, t_1 + t_2 + t_3\}$;

(*ii*). $t_1 + t_3 \in [v, u]$ and $t_1 + 2t_2 + t_3 \le \vartheta - 2\delta$;

(*iii*). $t_2 + t_3 \in [v, u]$ and $2t_1 + t_2 + t_3 \le \vartheta - 2\delta$;

(*iv*). $t_3 \le u - v$ and $t_1 \le v$, $2t_2 \le 1 - \vartheta - 2\delta$;

(*v*). $t_3 \le u - v$ and $t_2 \le v$, $2t_1 \le 1 - \vartheta - 2\delta$.

Note that from (4.48), (4.46) and a short computation, we have the equivalence

$$t \in [v, u] = \left[2\vartheta - 1, \frac{2 - 2\vartheta}{3}\right] \qquad \Longleftrightarrow \qquad \vartheta \le \vartheta(t) = \min\left(\frac{1 + t}{2}, \frac{2 - 3t}{2}\right).$$

Now suppose assumption (*a*) holds. Then we can use $\vartheta = \vartheta(t_1)$, which implies $t_1 \in [v, u]$. So $d \in \mathcal{D}^{\mathrm{well}}(x^{\vartheta-2\delta})$, $t_1 \le u$, and (*i*) holds for $t = t_1$, and thus d has a factorization as required.

Next, suppose assumption (*b*) holds. Then we can use $\vartheta = \vartheta(t_1, t_2, t_3)$, which also lies in $[\frac{1}{2}, \frac{5}{8}]$. Moreover, we have $t_1 \le \frac{1}{4} \le \frac{2-2\vartheta}{3} = u$. So $d \in \mathcal{D}^{\mathrm{well}}(x^{\vartheta-2\delta})$, $t_1 \le u$, and it suffices to verify one of conditions (*i*)-(*v*) above; we split into cases based on the maximum from (4.47):

- If $\vartheta \in \{\vartheta(t_1), \vartheta(t_2), \vartheta(t_1 + t_2), \vartheta(t_1 + t_2 + t_3)\}$, then (*i*) holds;
- If $\vartheta = \psi(\vartheta(t_1 + t_3), t_1 + 2t_2 + t_3)$, so $\vartheta = \vartheta(t_1 + t_3)$ and $t_1 + 2t_2 + t_3 + 2\delta \le \vartheta(t_1 + t_3)$, then (*ii*) holds;
- If $\vartheta = \psi(\vartheta(t_2 + t_3), 2t_1 + t_2 + t_3)$, so $\vartheta = \vartheta(t_2 + t_3)$ and $2t_1 + t_2 + t_3 + 2\delta \le \vartheta(t_2 + t_3)$, then (*iii*) holds;
- If $\vartheta = w(t_1, t_2, t_3)$, so $\vartheta = \min\left\{\frac{5-3t_3}{8}, 1 - 2t_2 - 2\delta\right\}$ and $\frac{1+t_1}{2} + 2\delta \le \vartheta$, then (*iv*) holds (noting that $u - v = \frac{5-8\vartheta}{3}$);
- If $\vartheta = w(t_2, t_1, t_3)$, so $\vartheta = \min\left\{\frac{5-3t_3}{8}, 1 - 2t_1 - 2\delta\right\}$ and $\frac{1+t_2}{2} + 2\delta \le \vartheta$, then (*v*) holds.

This completes our proof. □

Proof of Corollary 4.3. We very closely follow the sieve computations in [Lic23, Sections 7.1 and 7.2], using our Theorem 4.19 instead of [Lic23, Proposition 6.6]. By comparing the exponents $\vartheta(t)$, $\vartheta(t_1,t_2,t_3)$ from (4.46) and (4.47) with [Lic23, (6.2) and (6.4), with $\alpha=0$], this simply amounts to taking $\theta=0$ rather than $\theta=7/32$, and correcting the typo $w(t_1,t_3,t_2)\to w(t_2,t_1,t_3)$ in [Lic23, (6.4)]. Adapting the Mathematica file 'PrimeAPTwinTheta.nb' from [Lic23] with these quick changes, we obtain adjusted values for the sieve integrals on [Lic23, p. 30] as below (to be compared with the table on [Lic23, p. 32]).

n	G_n	n	G_n
1	38.8989	5	1.84027
2	-5.88606	6	0.628688
3	-4.13106	7	0.420003
4	-5.20164	8	0.913626

This results in an improvement of [Lic23, (7.13)] to

$$\{p\le x : p,p+2 \text{ are prime}\}\le 3.20254\ \Pi_2(x),$$

as we claimed. Note that we have omitted various parameter optimizations for simplicity. □

4.6 Smooth numbers with arbitrary weights

Here we prove Theorem 4.4, building on the arguments of Drappeau [Dra15]. As in Section 4.4, we will work in reverse compared to the outline in Section 4.2, gradually building up to a triple convolution estimate in Proposition 4.27.

We start with a bound for multilinear forms of incomplete Kloosterman sums as in (4.11), which follows from Propositions 4.8 and 4.9, and plays a similar role to Lemma 4.11.

Lemma 4.25. *Let $\varepsilon>0$, $1\ll N,T,H,K,L\ll x$ with $TH\ll N$, $a,d\in\mathbb{Z}\setminus\{0\}$ with $1\le|a|\le x^{\varepsilon}$, $1\le d\le x^{2\varepsilon}$, $\Phi_i(t)$ be smooth functions supported in $t\asymp 1$ with $\Phi_j^{(j)}\ll_j 1$, and*

$$\phi(h_1,h_2):=\Phi_1\left(\frac{h_1}{H}\right)\Phi_2\left(\frac{h_2}{H}\right)e(h_1\alpha_1+h_2\alpha_2),$$

where $\alpha_i\in\mathbb{R}/\mathbb{Z}$ have $\min_i T_H(\alpha_i)\ll x^{2\varepsilon}$ (recall (4.17)). Then for any smooth function $\Phi(x_1,x_2,z)$ supported in $x_i,z\asymp 1$, satisfying $\partial_{x_1}^{j_1}\partial_{x_2}^{j_2}\partial_z^{\ell}\Phi(x_1,x_2,z)\ll_{j_1,j_2,\ell,\varepsilon}1$, one

has

$$\sum_{n,n'\sim N}\left|\sum_{\substack{1\le|t|\le T\\(t,nn')=1\\t|n-n'}}\sum_{\substack{h,h'\\e=at(n'h-nh')\neq 0}}\phi(h,h')\sum_{\substack{k,\ell\\(k,dnn'\ell)=1}}\Phi\left(\frac{\ell}{L},\frac{k}{K}\right)e\left(\frac{e\overline{dnn'\ell}}{k}\right)\right| \tag{4.49}$$
$$\ll_\varepsilon x^{6\varepsilon}THN\left(L^2THN^3+\left(1+\frac{K^2}{N^3TH^2}\right)^{\theta_{\max}}K\left(K+LN^2\right)N^2\right)^{1/2}.$$

Proof. Let $\mathcal{K}$ denote the sum in question; we begin by splitting

$$\mathcal{K}=\mathcal{K}(n=n')+\mathcal{K}(n\neq n'), \tag{4.50}$$

where after a rescaling of the e variable,

$$\mathcal{K}(n=n'):=\sum_{n\sim N}\left|\sum_{\substack{1\le|t|\le T\\(t,n)=1}}\sum_{\substack{h,h'\\e=at(h-h')\neq 0}}\phi(h,h')\sum_{\substack{k,\ell\\(k,dn\ell)=1}}\Phi\left(\frac{\ell}{L},\frac{k}{K}\right)e\left(\frac{e\overline{dn\ell}}{k}\right)\right|.$$

The dominant contribution will come from $\mathcal{K}(n\neq n')$, but let us first bound the simpler sum $\mathcal{K}(n=n')$. Setting $e\leftarrow|e|$, putting e and $q=dn$ in dyadic ranges and denoting

$$a_{e,q}:=\mathbb{1}_{d|q}\sum_{\substack{1\le|t|\le T\\(t,q/d)=1\\h,h'\in\mathbb{Z}\\\pm at(h-h')=e}}\phi(h,h'),$$

we get

$$\mathcal{K}(n=n')\ll x^{o(1)}\sup_{\substack{E\ll|a|TH\\Q\asymp dN}}\mathcal{K}_1(E,Q), \tag{4.51}$$

where

$$\mathcal{K}_1=\sum_{q\sim Q}\left|\sum_{e\sim E}a_{e,q}\sum_{\substack{k,\ell\\(k,q\ell)=1}}\Phi\left(\frac{\ell}{L},\frac{k}{K}\right)e\left(\frac{\pm e\overline{q\ell}}{k}\right)\right|.$$

We recall that by Proposition 4.7, the tuple $(q,E,1,(a_{e,q})_{e\sim E},\|(a_{e,q})_{e\sim E}\|_2,1)$ satisfies Assumption 4.6. So by Proposition 4.9 with $S=1$ (which uses none of our new large sieve technology in this instance), we have

$$\mathcal{K}_1\ll x^{o(1)}\|(a_{e,q})_{e\sim E,q\sim Q}\|_2\left(L^2EQ+\left(1+\frac{K^2}{Q^2}\right)^{\theta_{\max}}K(K+LQ)(Q+E)\right)^{1/2}.$$

Recalling that $\phi(h,h')$ is supported on $h,h' \asymp H$, we can bound $a_{e,q} \ll x^{o(1)}H$ by the divisor bound, and thus $\|a_{e,q}\|_2 \ll x^{o(1)}\sqrt{EQ}H$. The resulting bound for $\mathcal{K}_1$ is non-decreasing in E,Q, so we can plug this into (4.51) to bound

$$\mathcal{K}(n=n') \ll_\varepsilon x^{6\varepsilon}H\sqrt{THN}\left(L^2THN + \left(1+\frac{K^2}{N^2}\right)^{\theta_{\max}} K(K+LN)(N+TH)\right)^{1/2}$$
$$\ll x^{6\varepsilon}THN\left(\frac{L^2HN}{T} + \frac{1}{T^2N}\left(1+\frac{K^2}{N^2}\right)^{\theta_{\max}} K(K+LN)N^2\right)^{1/2},$$

where in the second line we multiplied and divided by T, then used the assumption $TH \ll N$. Since $\theta_{\max} \le 1/3$, we have

$$\frac{1}{T^2N}\left(1+\frac{K^2}{N^2}\right)^{\theta_{\max}} \ll \left(1+\frac{K^2}{N^5T}\right)^{\theta_{\max}} \ll \left(1+\frac{K^2}{N^3TH^2}\right)^{\theta_{\max}},$$

so the contribution of $\mathcal{K}(n=n')$ is acceptable in (4.49).

To bound $\mathcal{K}(n\neq n')$, we let $n_0 := (n,n')$, substitute $(n,n',e) \leftarrow (n_0n, n_0n', n_0e)$, and use the triangle inequality in t to obtain

$$\mathcal{K}(n\neq n') \ll$$
$$\sum_{\substack{n_0\le 2N\\ n,n'\sim N/n_0\\ (n,n')=1}} \sum_{\substack{1\le|t|\le T\\ t|n-n'\neq 0}} \left| \sum_{\substack{h,h'\\ e=at(n'h-nh')\neq 0}} \phi(h,h') \sum_{\substack{k,\ell\\ (k,dn_0nn'\ell)=1}} \Phi\left(\frac{\ell}{L},\frac{k}{K}\right) e\left(\frac{e\overline{dn_0nn'\ell}}{k}\right)\right|.$$

We then put $n_0, e \leftarrow |e|$, and $q = dn_0nn'$ in dyadic ranges, and use the divisor bound to write

$$\mathcal{K}(n\neq n') \ll x^{o(1)} \sup_{\substack{N_0\ll N\\ E\ll |a|THN/N_0\\ Q\asymp dN^2/N_0}} \mathcal{K}_2(N_0,E,Q), \tag{4.52}$$

where

$$\mathcal{K}_2 := \sum_{q\sim Q} \max_{\substack{n_0\le 2N\\ n,n'\sim N/n_0\\ (n,n')=1\\ 1\le|t|\le T}} \left| \sum_{e\sim E} \sum_{\substack{h,h'\\ at(n'h-nh')=\pm e}} \phi(h,h') \sum_{\substack{k,\ell\\ (k,q\ell)=1}} \Phi_0\left(\frac{\ell}{L}\right)\Phi_0\left(\frac{k}{K}\right) e\left(\frac{\pm e\overline{q\ell}}{k}\right)\right|$$
$$= \sum_{q\sim Q}\left| \sum_{e\sim E} a_{e,q} \sum_{\substack{k,\ell\\ (k,q\ell)=1}} \Phi_0\left(\frac{\ell}{L}\right)\Phi_0\left(\frac{k}{K}\right) e\left(\frac{\pm e\overline{q\ell}}{k}\right)\right|.$$

Above, we denoted

$$a_{e,q} := \sum_{\substack{h,h'\in\mathbb{Z}\\ \pm at(q)(n'(q)h-n(q)h')=e}} \phi(h,h')$$

if the maximum on the first line is attained at some choice of $n(q), n'(q), t(q)$; if the maximum is empty, we let $a_{e,q} = 0$. Then by Proposition 4.8, we know that $(q, E, x, (a_{e,q})_{e\sim E}, A_q, Y)$ satisfies Assumption 4.6, where

$$Y := \frac{EH}{|a|(H + N/N_0)(N/N_0)\min_i T_H(\alpha_i)},$$
$$A_q := \left(\sum_{e\sim E} |a_{e,q}|^2\right)^{1/2} + \sqrt{TE}\sqrt{\frac{HN_0}{N} + \frac{H^2N_0^2}{N^2}}.$$

Since $\min_i T_H(\alpha_i) \ll x^{2\varepsilon}$, we further have

$$Y \gg_\varepsilon x^{-2\varepsilon}\frac{EHN_0}{|a|(H+N)N}.$$

From Proposition 4.9, we conclude that

$$\mathcal{K}_2 \ll_{\varepsilon,a} x^{2\varepsilon}\|A_q\|_2\left(L^2EQ + \left(1 + \frac{K^2}{Q^2\frac{EHN_0}{(H+N)N}}\right)^{\theta_{\max}} K(K+LQ)(Q+E)\right)^{1/2}.$$

Now by the same computation as in (4.25) (incorporating a sum over $1 \le |t| \le T$, $t \mid e$), we have

$$\|A_q\|_2^2 \ll_\varepsilon x^{2\varepsilon}TE(HN + H^2N_0),$$

so that (using $|a| \le x^\varepsilon$)

$$\mathcal{K}_2 \ll_\varepsilon x^{4\varepsilon}\sqrt{TE(HN+H^2N_0)}$$
$$\times\left(L^2EQ + \left(1 + \frac{K^2(H+N)N}{Q^2EHN_0}\right)^{\theta_{\max}} K(K+LQ)(Q+E)\right)^{1/2}.$$

Since this right-hand side is non-decreasing in E (due to $\theta_{\max} \le 1$), we may use the bounds $E \ll |a|THN/N_0$, $Q \asymp dN^2/N_0$ from (4.52), and $d \le x^{2\varepsilon}$ to obtain

$$\mathcal{K}_2 \ll_\varepsilon x^{5\varepsilon}\sqrt{T\frac{THN}{N_0}(HN+H^2N_0)} \times$$
$$\left(L^2\frac{THN}{N_0}\frac{N^2}{N_0} + \left(1 + \frac{K^2(H+N)N}{\left(\frac{N^2}{N_0}\right)^2\frac{THN}{N_0}HN_0}\right)^{\theta_{\max}} K\left(K + L\frac{N^2}{N_0}\right)\left(\frac{N^2}{N_0} + \frac{THN}{N_0}\right)\right)^{1/2}.$$

Since $\theta_{\max} \le 1/2$, this bound is seen to be non-increasing in the $N_0 \gg 1$ parameter; plugging this into (4.52) and using the assumption $TH \ll N$, we conclude that

$$\mathcal{K}(n \ne n') \ll_\varepsilon x^{6\varepsilon}THN\left(L^2THN^3 + \left(1 + \frac{K^2}{N^3TH^2}\right)^{\theta_{\max}} K\left(K + LN^2\right)N^2\right)^{1/2},$$

which gives the right-hand side of (4.49). □

We now deduce a power-saving bound for an exponential sum as in (4.10) (before passing to the complementary divisor), which improves the first set of conditions in [Dra15, Proposition 1].

Lemma 4.26 (Exponential sum bound for convolutions)**.** *Let $\varepsilon > 0$ be small enough, $a \in \mathbb{Z} \setminus \{0\}$, $v, d_1, d_2 \in \mathbb{Z}_+$, $\theta := 7/32$, and $1 \ll M, K, N, L, H, R \ll x$ satisfy*

$$|avd_1d_2| \ll x^{\varepsilon}, \quad NL \ll x^{\varepsilon}K, \quad R \ll K \ll \min(x^{-3\varepsilon}MN, LN^2), \quad H \ll x^{\varepsilon}\frac{R}{M},$$
$$K \ll x^{-25\varepsilon}\sqrt{MNR}, \qquad K^{3+\theta}N^{2-3\theta} \ll x^{-200\varepsilon}M^{2-2\theta}R^{2+\theta}L. \tag{4.53}$$

Let $(u_k)_{K<k\le 4K}, (\beta_n)_{n\sim N}, (\lambda_\ell)_{\ell\sim L}$ be complex sequences such that $|u_k| \le \tau(k)$, $|\beta_n| \le 1$, $|\lambda_\ell| \le 1$, and

$$(k, vd_1d_2) > 1 \;\Rightarrow\; u_k = 0, \qquad (n\ell, vd_1) > 1 \;\Rightarrow\; \beta_n\lambda_\ell = 0.$$

Then for any smooth functions $\Phi(t)$, $\Psi(t)$ supported in $t \asymp 1$ with $\Phi^{(j)}, \Psi^{(j)} \ll_j 1$, and any $\omega \in \mathbb{R}/\mathbb{Z}$ with $T_H(\omega) \ll x^{\varepsilon}$, one has

$$\sum_{\substack{r\sim R\\(r, avd_1d_2)=1}} \frac{M}{r}\Psi\left(\frac{r}{R}\right) \sum_{\substack{k,n,\ell\\ d_1k\equiv d_2n\ell \pmod r\\ (d_1k, d_2n\ell)=1}} u_k\beta_n\lambda_\ell \sum_{h\in\mathbb{Z}} e(h\omega)\,\Phi\left(\frac{h}{H}\right) e\left(\frac{-ha\overline{vd_1d_2k}}{r}\right)$$
$$\ll_{\varepsilon} x^{-10\varepsilon}\frac{KMNL}{R}. \tag{4.54}$$

Remark. As is common for exponential sum estimates with a variable $h \sim H$ coming from Poisson summation, Lemma 4.26 needs to win a factor of H (times an extra x^{ε}) over the trivial bound; the same was true for Lemma 4.12.

Proof of Lemma 4.26. We closely follow the proof of [Dra15, Proposition 1]. We denote the exponential sum considered in (4.54) by $\mathcal{R}$; it is essentially identical to the sum in [Dra15, Section 3.5], except that h lies in a smooth dyadic range. As in [Dra15, Section 3.5], we denote

$$\nu := vd_1d_2 \ll x^{\varepsilon} \qquad \text{and} \qquad T := \frac{\max(d_1K, d_2NL)}{R} \ll x^{2\varepsilon}\frac{K}{R}.$$

Following through the computations in [Dra15, p. 844–846] with minor changes, we obtain

$$\mathcal{R} \ll_{\varepsilon} x^{5\varepsilon}KM^{-1} + \max_{\substack{\sigma\mid a\\ w \pmod v}} \left(x^{5\varepsilon}MR^{-1}(KLT)^{1/2}\mathcal{B}^{1/2}\right), \tag{4.55}$$

where

$$\mathcal{B} := \sum_{n,n'\sim N} \Bigg| \sum_{\substack{1\le|t|\le T\\(t,nn')=1\\t|n-n'}} \sum_{\ell} \Phi_0\left(\frac{\ell}{L}\right) \sum_{(k,\nu d_2 nn'\ell)=1} \Phi_0\left(\frac{k}{K}\right) \sum_{h,h'\in\mathbb{Z}} \phi(h,h') \times e\left(at(n'h-nh')\frac{\overline{\nu d_2 nn'\ell}}{k}\right)\Bigg|,$$

with

$$\phi(h,h') := \Phi\left(\frac{h}{H}\right)\overline{\Phi\left(\frac{h'}{H}\right)} e\left((h-h')\omega'\right), \qquad \omega' := \omega + \frac{a\overline{w}}{v}.$$

This corresponds to the sum on top of [Dra15, p. 847]; note that we broke up the coefficients $\beta(n,h)$ in [Dra15, p. 846] and ignored the phases in n, n' via absolute values. We note at this point that by (4.17),

$$\begin{aligned} T_H(\omega') &\le \min_{t\in\mathbb{Z}_+}\left(tv + H\|tv\omega'\|\right)\\ &= \min_{t\in\mathbb{Z}_+}\left(tv + H\|tv\omega\|\right)\\ &\le \min_{t\in\mathbb{Z}_+} v\left(t + H\|t\omega\|\right) = vT_H(\omega) \ll x^{2\varepsilon}. \end{aligned}$$

Letting $e := at(n'h - nh')$, the contribution of $e = 0$ is bounded by

$$\mathcal{B}(e=0) \ll_\varepsilon x^\varepsilon KLNHT,$$

just as in [Dra15, (3.24)]. Since by (4.53),

$$TH \ll x^{3\varepsilon}\frac{K}{R}\frac{R}{M} \ll x^{3\varepsilon}\frac{K}{M} \ll N,$$

Lemma 4.25 applies directly to the contribution of $e \neq 0$, giving

$$\mathcal{B}(e\neq 0) \ll_\varepsilon x^{6\varepsilon}THN\left(L^2THN^3 + \left(1+\frac{K^2}{N^3TH^2}\right)^\theta K\left(K+LN^2\right)N^2\right)^{1/2}.$$

Plugging these bounds and $K \le LN^2$ (from (4.53)) into (4.55), we obtain

$$\begin{aligned} \mathcal{R} &\ll_\varepsilon x^{5\varepsilon}KM^{-1} + x^{10\varepsilon}MR^{-1}\sqrt{KLT}\\ &\times\left(\sqrt{KLNHT} + \sqrt{THN}\left(L^2THN^3 + \left(1+\frac{K^2}{N^3TH^2}\right)^\theta KLN^4\right)^{1/4}\right). \end{aligned}$$

Combining $TH \ll N$ with $L \le NL \ll x^\varepsilon K$ (from (4.53)), we see that

$$L^2THN^3 \ll L^2N^4 \ll x^\varepsilon KLN^4,$$

so

$$\mathcal{R} \ll_\varepsilon x^{5\varepsilon} KM^{-1} + x^{11\varepsilon} MR^{-1}\sqrt{KLT}\left(\sqrt{KLNHT} + \sqrt{THN}\left(1+\frac{K^2}{N^3TH^2}\right)^{\theta/4}\left(KLN^4\right)^{1/4}\right).$$

Since this bound is non-decreasing in H and T, we can plug in $H \le x^\varepsilon R/M$ (from (4.53)) and $T \le x^{2\varepsilon}K/R$ to obtain

$$\begin{aligned}\mathcal{R} &\ll_\varepsilon x^{5\varepsilon} KM^{-1} + x^{15\varepsilon}\frac{M}{R}\sqrt{\frac{K^2L}{R}}\left(K\sqrt{\frac{LN}{M}} + \sqrt{\frac{KN}{M}}\left(1+\frac{KM^2}{N^3R}\right)^{\theta/4}\left(KLN^4\right)^{1/4}\right)\\ &= x^{5\varepsilon} KM^{-1} + x^{15\varepsilon}\frac{K^2L\sqrt{MN}}{R^{3/2}} + x^{15\varepsilon}\frac{K^{7/4}L^{3/4}M^{1/2}N^{3/2}}{R^{3/2}}\left(1+\frac{KM^2}{N^3R}\right)^{\theta/4}.\end{aligned}$$

This is acceptable in (4.54) (i.e., $\ll_\varepsilon x^{-10\varepsilon}KMNL/R$) provided that

$$\begin{aligned} R &\ll x^{-15\varepsilon}M^2NL, & K &\ll x^{-25\varepsilon}\sqrt{MNR},\\ K^3N^2 &\ll x^{-100\varepsilon}M^2R^2L, & K^{3+\theta}N^{2-3\theta} &\ll x^{-100\varepsilon}M^{2-2\theta}R^{2+\theta}L.\end{aligned}$$

The first of these conditions follows easily from $R \ll K \ll x^{-25\varepsilon}\sqrt{MNR}$, while the second and fourth conditions are part of (4.53). It remains to verify the third condition which can be deduced from (4.53) as follows:

$$\begin{aligned} K^3N^2 = \left(K^{3-3\theta}N^{2-2\theta}\right)^{\frac{1}{1-\theta}} &\ll \left(\left(\frac{K}{R}\right)^{3\theta}\left(\frac{x^\varepsilon K}{NL}\right)^{\theta}K^{3-3\theta}N^{2-2\theta}\right)^{\frac{1}{1-\theta}}\\ &= \left(\frac{x^{\theta\varepsilon}}{R^{3\theta}L^\theta}K^{3+\theta}N^{2-3\theta}\right)^{\frac{1}{1-\theta}}\\ &\ll \left(\frac{1}{R^{3\theta}L^\theta}x^{(\theta-200)\varepsilon}M^{2-2\theta}R^{2+\theta}L\right)^{\frac{1}{1-\theta}}\\ &\ll x^{-100\varepsilon}\left(M^{2-2\theta}R^{2-2\theta}L^{1-\theta}\right)^{\frac{1}{1-\theta}} = x^{-100\varepsilon}M^2R^2L.\end{aligned}$$

This completes our proof. □

We can now deduce an estimate on the equidistribution in arithmetic progressions of convolutions of three sequences, corresponding to (4.7) and improving [Dra15, Théorème 3]. For $r \in \mathbb{Z}_+$ and $k \pmod r$, we recall Drappeau's notation

$$\omega_\varepsilon(k;r) := \sum_{\substack{\chi \text{ primitive}\\ \mathrm{cond}(\chi)\le x^\varepsilon\\ \mathrm{cond}(\chi)\mid r}} \chi(k) \tag{4.56}$$

from [Dra15, (3.1)]. Separating all the Dirichlet characters of conductors $\le x^\varepsilon$ was crucial to obtaining power-saving convolution estimates in [Dra15]. In a certain sense,

$\frac{\mathbb{1}_{(k,r)=1}}{\varphi(r)}\omega_\varepsilon(k;r)$ gives a better approximation to the function $\mathbb{1}_{k\equiv 1 \pmod r}$ than the crude $\frac{\mathbb{1}_{(k,r)=1}}{\varphi(r)}$; indeed, $\frac{\mathbb{1}_{(k,r)=1}}{\varphi(r)}\omega_\varepsilon(k;r)$ interpolates between the latter two quantities as ε varies in $[0,\infty)$.

The relevant constraints on the ranges of the convolved sequences are gathered in (4.57). We note that the conditions on the top row of (4.57) also appear in [Dra15, Théorème 3].

Proposition 4.27 (Triple convolution estimate)**.** *For any small enough $\varepsilon > 0$, there exists $\delta > 0$ such that the following holds. Let $M, N, L \gg 1$, $x := MNL$, $a_1, a_2 \in \mathbb{Z} \setminus \{0\}$ satisfy $|a_1 a_2| \le x^\delta$, $(a_1, a_2) = 1$, and $(\alpha_m)_{m\sim M}, (\beta_n)_{n\sim N}, (\gamma_\ell)_{\ell\sim L}$ be 1-bounded complex sequences. Suppose that with $\theta := 7/32$, one has*

$$x^\varepsilon \le N, \quad NL \le x^{2/3-5\varepsilon}, \quad L \le x^{-\varepsilon}M, \quad M \le R \le x^{-\varepsilon}NL, \quad N^2L^3 \le x^{1-\varepsilon}R,$$
$$N^{7-4\theta}L^{4-\theta} \le x^{2-2\theta-\varepsilon}R^{2+\theta}. \tag{4.57}$$

Then one has

$$\sum_{\substack{r\sim R\\(r,a_1a_2)=1}} \left| \sum_{\substack{m\sim M\\n\sim N\\\ell\sim L}} \alpha_m\beta_n\gamma_\ell \left(\mathbb{1}_{mn\ell\equiv a_1\overline{a_2} \pmod r} - \frac{\mathbb{1}_{(mn\ell,r)=1}}{\varphi(r)}\omega_\varepsilon(mn\ell\overline{a_1}a_2;r)\right)\right| \ll_{\varepsilon,a_1,a_2} x^{1-\delta}.$$

Remark. The above inequalities $R \le x^{-o(1)}NL$, $N^2L^3 \le x^{1-o(1)}R$, and $N^{7-4\theta}L^{4-\theta} \le x^{2-2\theta-o(1)}R^{2+\theta}$ imply $R \le x^{5/8-o(1)}$, which corresponds to the level from Theorem 4.4. We note that for $R = x^{-o(1)}NL$, the last inequality in (4.57) is equivalent to $N^7L^4 \le x^{-o(1)}R^2$, which explains why our final exponent of distribution does not depend on the θ parameter.

Proof of Proposition 4.27. We closely follow the proof of [Dra15, Théroème 3], which applies Cauchy–Schwarz in r, m (and inserts a smooth majorant $f(m) = \Phi(m/M)$) to obtain three dispersion sums [Dra15, Section 3.1]. We change nothing in the treatment of the second and third dispersion sums from [Dra15, Sections 3.2, 3.3], noting that the conditions on the top row of (4.57) are sufficient here.

We also begin treating the first dispersion sum similarly as in [Dra15, Section 3.4], with the technical change that we Poisson complete via Lemma 2.2 rather than [Dra15, Lemme 2]. Instead of [Dra15, (3.12)], we thus obtain

$$\mathcal{S}_1 = \widehat{f}(0)X_1 + \sum_{\substack{v,d_1,e_1,e_2\le x^\eta\\ d_1\mid v^\infty, e_1\mid a_2^\infty, e_2\mid a_2}} \int_{0.1}^{10} \sum_{\substack{H_j=2^j\\1\le H_j\le H}} R_{j,u}(v;d_1,e_1,e_2)\,\frac{du}{u} + O_\eta(x^{1-\eta/4}KR^{-1}),$$

where X_1 is the main term from [Dra15, (3.12)], $H = x^\eta RM^{-1}$, Ψ_j are as in Lemma 2.2, and

$$R_{j,u}(v;d_1,e_1,e_2) := \sum_{\substack{r\sim R\\(r,a_1a_2v)=1}} \frac{M}{r}\widetilde{\Phi}\left(\frac{ur}{R}\right) \sum_{(k_1,k_2)\in\mathcal{K}} u_{k_1}\overline{u_{k_2}} \sum_{h\in\mathbb{Z}} e(h\omega)\,\Psi_j\left(\frac{|h|}{H_j}\right) \times e\left(\frac{-ha_1\overline{a_2k_1}}{r}\right),$$

where $u_k := \sum_{n\ell=k}\beta_n\gamma_\ell$, $\mathcal{K}$ is as in [Dra15, p. 841], and

$$\omega := \frac{uM}{R} \ll H^{-1}x^\eta \qquad \Rightarrow \qquad T_H(\omega) \ll x^\eta.$$

We then develop and bound $R_{j,u}$ as in [Dra15, p. 843], with the only major change that we use our Lemma 4.26 instead of [Dra15, Proposition 1]. To apply Lemma 4.26 (with $\eta > 0$ in place of ε), we need to verify the conditions in (4.53); thus instead of the third-to-last display on [Dra15, p. 843], we require that $|a_1a_2| \le x^{\eta/10}$ and

$$R \ll x^{-100\eta}NL, \qquad L \ll x^{-100\eta}M, \qquad 1 \ll x^{-100\eta}N,$$
$$\sqrt{N}L \ll x^{-200\eta}\sqrt{MR}, \qquad N^{5-2\theta}L^{2+\theta} \ll x^{-300\eta}M^{2-2\theta}R^{2+\theta}.$$

Here, we implicitly used that in Drappeau's computations near [Dra15, bottom of p. 843], one has $vd_1d_2 \ll x^{5\eta}$, $H = x^\eta RM^{-1}$, and $x^{10\eta}NL \ll K \ll x^{-10\eta}NL$. Since $MNL \asymp x$, these conditions follow from (4.57) provided η is chosen sufficiently small in terms of ε. □

Finally, we prove a direct generalization of Theorem 4.4, in a form analogous to [Dra15, Théorème 1]. We recall the notation specific to smooth numbers,

$$u := \frac{\log x}{\log y}, \qquad H(u) := \exp\left(\frac{u}{(\log(u+1))^2}\right),$$

from [Dra15], as well as the definitions of $\Psi_q(x,y)$ and $\Psi(x,y;a,q)$ from (4.3).

Theorem 4.28 (Smooth numbers in APs to large moduli, refined)**.** *For any $\varepsilon > 0$, there exist $\delta, C > 0$ such that the following holds. Let $x \ge 2$ and $a_1, a_2 \in \mathbb{Z}\setminus\{0\}$ satisfy $(a_1,a_2) = 1$ and $|a_1a_2| \le x^\delta$. Then for any $y \in [(\log x)^C, x^{1/C}]$ and $A \ge 0$, one has*

$$\sum_{\substack{q\le x^{5/8-\varepsilon}\\(q,a_1a_2)=1}} \left|\Psi(x,y;a_1\overline{a_2},q) - \frac{\Psi_q(x,y)}{\varphi(q)}\right| \ll_{\varepsilon,A} \Psi(x,y)\left(H(u)^{-\delta}(\log x)^{-A} + y^{-\delta}\right).$$

The implicit constant is effective if $A < 1$.

Proof. We assume without loss of generality that $\varepsilon > 0$ is small enough, and choose η to be a small multiple of ε. As in [Dra15, p. 855–856], Harper's result [Dra15, Lemme 5] (see also [Har12]) handles the contribution of Dirichlet characters with conductors $\leq x^\eta$, so it suffices to prove the bound

$$\sum_{\substack{q \leq x^{5/8-\varepsilon} \\ (q, a_1 a_2)=1}} \left| \sum_{\substack{n \leq x \\ P^+(n) \leq y}} \left(\mathbb{1}_{n \equiv a_1 \overline{a_2} \pmod q} - \frac{\mathbb{1}_{(n,q)=1}}{\varphi(q)} \omega_\eta(n; a_1 \overline{a_2}) \right) \right| \ll_\varepsilon x^{1-\delta/2}, \tag{4.58}$$

for some $\delta = \delta(\varepsilon) > 0$. Such a power-saving is enough up to a final rescaling of δ, due to the bound $x^{1-\delta/2} \ll_\delta \Psi(x,y) y^{-\delta/4}$ for sufficiently large C (see [Dra15, p. 856]).

The proof of (4.58) is completely analogous to that of [Dra15, Proposition 2], except that we use our triple convolution estimate from Proposition 4.27 instead of [Dra15, Théorème 3]. The key point is that the indicator function of smooth numbers can be approximated by convolutions of three sequences with pre-specified ranges, due to their flexible factorization. Specifically, we rescale $\varepsilon \leftarrow 100\varepsilon$, take $C = \varepsilon^{-1}$ so that $y \leq x^{1/C} \leq x^\varepsilon$, and put $q \leftarrow r$ in dyadic ranges $r \sim R$. Then instead of the parameters on the bottom of [Dra15, p. 852], we pick

$$M_0 := \frac{x^{1-10\varepsilon}}{R}, \qquad N_0 := \frac{R^2}{x^{1-40\varepsilon}}, \qquad L_0 := \frac{x^{1-30\varepsilon}}{R},$$

in the range $x^{(1/2)-(\varepsilon/10)} \leq R \leq x^{5/8-100\varepsilon}$ (smaller values of R are covered by previous results [Dra15]). Any resulting values of M, N, L with

$$M_0 \leq M \leq y\frac{M_0}{2}, \qquad L_0 \leq L \leq y\frac{L_0}{2}, \qquad y^{-2} N_0 \leq N \leq N_0$$

are seen to satisfy the conditions in (4.57). In particular, for the last two conditions in (4.57), we note that

$$\frac{xR}{N_0^2 L_0^3} = x^{10\varepsilon} \qquad \text{and} \qquad \frac{x^{2-2\theta} R^{2+\theta}}{N_0^{7-4\theta} L_0^{4-\theta}} = \frac{x^{5(1-\theta)-(160-130\theta)\varepsilon}}{R^{8(1-\theta)}} \geq x^{100\varepsilon},$$

since $R \leq x^{5/8-100\varepsilon}$; this gives enough $x^{o(1)}$ room when replacing M_0, N_0, L_0 by M, N, L, since $y \leq x^\varepsilon$. Following through the combinatorial decompositions and separations of variables in [Dra15, p. 852–854], we can apply Proposition 4.27 for the sequences $(\alpha_m)^{(j)}, (\beta_m)^{(j)}, (\lambda_\ell)^{(j)}$ on [Dra15, p. 854], which recovers the desired bound. □

4.7 Smooth numbers with weights on smooth moduli

Here we quickly prove a variant (and in fact, a generalization) of Theorem 4.4 when the sum over q is restricted to smooth moduli, which improves the first exponent of

distribution in [BD20, Théorème 2.1] from $\frac{3}{5}-\varepsilon$ to $\frac{5}{8}-\varepsilon$. Recall again the notation from (4.3).

Theorem 4.29 (Smooth numbers in APs to smooth moduli)**.** *For any $\varepsilon, A > 0$ and $k \geq 1$, there exist $\delta, C > 0$ such that the following holds. Let $x \geq 2$ and $a_1, a_2 \in \mathbb{Z} \setminus \{0\}$ satisfy $(a_1, a_2) = 1$ and $|a_1 a_2| \leq x^\delta$. Then for any $y_1 \in [(\log x)^C, x^{1/C}]$, $y_2 \in [(\log x)^C, x]$, $Q \leq x^{5/8-\varepsilon}$, and $q_0 \in \mathbb{Z}_+$ with $q_0 \leq x^\delta$, $(q_0, a_1 a_2) = 1$, $P^+(q_0) \leq y_2$, one has*

$$\sum_{\substack{q\sim Q\\ P^+(q)\leq y_2\\ (q,a_1a_2)=1}} \tau_k(q)\left|\Psi(x,y_1;a_1\overline{a_2},q_0q)-\frac{\Psi_{q_0q}(x,y_1)}{\varphi(q_0q)}\right| \ll_{\varepsilon,A,k} \frac{\Psi(x,y_1)}{(\log x)^A}\frac{\Psi(Q,y_2)}{\varphi(q_0)Q}e^{O_k(u_2)}, \tag{4.59}$$

where $u_2 := (\log x)/\log y_2$.

Proof. Again, we assume without loss of generality that $\varepsilon > 0$ is small enough, and we will pick η, δ to be small enough in terms of ε, A, k. It suffices to prove our claim with δ replaced by $\delta/10$.

Let $\mathcal{S}$ denote the left-hand side of (4.59). Recalling the notation in (4.56), we separate the contribution of Dirichlet characters of small conductors by writing

$$|\mathcal{S}| \leq \mathcal{S}_{\text{small}} + \mathcal{S}_{\text{large}},$$

where

$$\mathcal{S}_{\text{large}} := \sum_{\substack{q\sim Q\\ P^+(q)\leq y_2\\ (q,a_1a_2)=1}} \tau_k(q)\left|\sum_{\substack{n\leq x\\ P^+(n)\leq y_1}}\left(\mathbb{1}_{n\equiv a_1\overline{a_2} \pmod{q_0q}}-\frac{\mathbb{1}_{(n,q_0q)=1}}{\varphi(q_0q)}\omega_\eta(n;q_0q)\right)\right|,$$

$$\mathcal{S}_{\text{small}} := \sum_{\substack{q\sim Q\\ P^+(q)\leq y_2\\ (q,a_1a_2)=1}} \frac{\tau_k(q)}{\varphi(q_0q)}\left|\sum_{\substack{n\leq x\\ P^+(n)\leq y_1}}\sum_{\substack{\chi \pmod{q_0q}\\ 1<\mathrm{cond}(\chi)\leq x^\eta}}\chi(n)\right|.$$

For $\mathcal{S}_{\text{small}}$, we use the triangle inequality for the sum over χ, and then proceed identically as in [BD20, after (2.3)]; this gives the desired bound when δ is sufficiently small and C is sufficiently large.

For $\mathcal{S}_{\text{large}}$, we drop the smoothness condition on q, use the pointwise divisor bound $\tau_k(q) \ll_k q^{o(1)}$, group q_0q into a new variable, and drop its divisibility constraint by q_0. Combined with (4.58) (which followed from Proposition 4.27), this gives

$$\mathcal{S}_{\text{large}} \ll_{\varepsilon,k} x^{1-\delta/3},$$

provided δ is sufficiently small and C is sufficiently large in terms of ε. This is acceptable once C is chosen to be large enough in terms of δ, A, due to the bounds $x^{1-\delta/10} \ll_\delta \Psi(x, y_1)y_1^{-\delta/20}$, $Q^{1-\delta/10} \ll_\delta \Psi(Q, y_2)$, and $y_1 \geq (\log x)^C$, $q_0 \leq x^{\delta/10}$. □

Corollary 4.30 (Smooth values of factorable quadratic polynomials)**.** *For any $\varepsilon > 0$, there exist $C, \delta > 0$ such that the following holds. Let $x \geq 2$ and $a, b, c, d \in \mathbb{Z}$ satisfy $(a, c) = 1$, $ad - bc \neq 0$, and $|a|, |b|, |c|, |d| \leq x^\delta$. Then for any $(\log x)^C \leq y_1 \leq y_2 \leq x$ with $y_2 \leq y_1^C$, one has*

$$\#\{n \leq x : P^+(an + b) \leq y_1,\ P^+(cn + d) \leq y_2\} \ll_\varepsilon \Psi(x, y_1)\varrho(u_2)^{5/8-\varepsilon},$$

where $u_2 := (\log x)/\log y_2$.

Proof. This is identical to the proof of [BD20, Théorème 4.1], using Theorem 4.29 instead of [BD20, Théorème 2.1]. When applying Theorem 4.29, q will be a divisor of $cn + d$ coming from an upper-bound sieve [BD20, Proposition 3.1], while $q_0 = a$, $a_1 = -(ad - bc)$, and $a_2 = c$; note that

$$q \mid cn + d \qquad \Longleftrightarrow \qquad an + b \equiv a_1\overline{a_2} \pmod{q_0 q},$$

and $P^+(an + b) \leq y_1$, $P^+(q) \leq y_2$. □

Proof of Corollary 4.5. Take $(a, b, c, d) = (1, 0, 1, 1)$, $y_1 = y_2$ in Corollary 4.30, and use $\Psi(x, y_1) = x\varrho(u)e^{O(u)}$ where $u := (\log x)/\log y$ (see [BD20, (1.7)] and [Hil86, (2.6) and (2.7)]). □

Chapter 5

Density theorems for GL_n via Rankin–Selberg L-functions

(based on joint work with Jared Duker Lichtman)

5.1 Introduction

Several results in analytic number theory [Top18; BD20; DPR23; Lic23; Wu23] depend on the best progress towards the Ramanujan–Petersson conjecture and its Archimedean counterpart, Selberg's eigenvalue conjecture. These concern the sizes of the Hecke and Laplacian eigenvalues of automorphic forms for congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$, corresponding to the local parameters of GL_2 automorphic representations. While the full conjectures seem out of the reach of current methods, it is desirable for such applications to obtain partial results about the (conjecturally inexistent) *exceptional forms*, which disobey the Ramanujan and Selberg bounds. In some cases, such substitutes can even match the best conditional results [Wat95; ABL21].

The GL_n setting [LRS99; BB11; Blo23; AB24] has also attracted significant interest, in part because it leads to bounds for GL_2 via symmetric power lifts [LRS95; Kim03]. The generalized Ramanujan conjecture (GRC) from Conjecture 2.13, one of the most important unsolved problems in number theory [BB13], asserts that the local components of cuspidal automorphic representations of GL_n over a number field are tempered. For concreteness, let π be a cuspidal automorphic representation of $\mathrm{GL}_n(\mathbb{A}_{\mathbb{Q}})$ with unitary central character; such π has a generic, unitary, and irreducible local component at each place v, which is parametrized by Langlands parameters $\{\mu_{\pi,j}(v)\}_{j=1}^n$ (see Section 2.3.1). At the unramified places v, GRC predicts that

$$\operatorname{Re} \mu_{\pi,j}(v) = 0.$$

What motivates the present work is a discrepancy between the two main ways to make progress towards the generalized Ramanujan conjecture. On the one hand, the best pointwise, uniform bounds (i.e., valid for any given representation π) rely on properties of L-functions. Without further assumptions, the record at unramified places v of $\mathbb{Q}$ is

$$|\mathrm{Re}\,\mu_{\pi,j}(v)| \leq \begin{cases} \frac{1}{2} - \frac{1}{n^2+1}, & \text{if } n \geq 5, \\ \frac{1}{2} - \frac{1}{\frac{n(n+1)}{2}+1}, & \text{if } n \in \{3,4\}, \\ \frac{7}{64}, & \text{if } n = 2, \end{cases} \tag{5.1}$$

due to Luo–Rudnick–Sarnak [LRS95] when $n \geq 5$ (see also [Ser81; RS96]), and to Kim–Sarnak [Kim03, Appendix 2] otherwise; see [LRS99; BB11] for results on general number fields.

On the other hand we have results that hold on average, stating that there are few representations in a given family which fail the Ramanujan conjecture by too much. The best such results have come from the spectral theory of automorphic forms [Iwa97; Hum18; Blo23; AB24], and relied, through Kuznetsov-type trace formulae [DI82c; AB24], on bounds for Kloosterman sums. Sarnak's density conjecture [Sar91; SX91] gives a prediction for families of representations induced by forms on a congruence subgroup, based on a linear interpolation between two extreme cases. One of these cases is the non-cuspidal trivial representation, so one should expect Sarnak's conjecture to be suboptimal for cuspidal forms, and the following result of Blomer [Blo23, Theorem 1] gives an improvement for the subgroup $\Gamma_0^{(n)}(q) \subset \mathrm{SL}_n(\mathbb{Z})$.

Theorem 5.1 (Blomer [Blo23]). *Let $n \in \mathbb{Z}_+$, v be a place of $\mathbb{Q}$, $q \neq v$ be a prime, $I \subset [0,\infty)$ be a fixed compact interval, and $\mathcal{S}_I(q)$ be the family of cuspidal automorphic representations of $\mathrm{GL}_n(\mathbb{A}_\mathbb{Q})$ induced by $\Gamma_0^{(n)}(q)$-invariant Maass forms, with Laplacian eigenvalues $\lambda \in I$. Then for any $\varepsilon > 0$ and $\theta \in [0, \frac{1}{2})$, one has*

$$\#\left\{\pi \in \mathcal{S}_I(q) : \max_j |\mathrm{Re}\,\mu_{\pi,j}(v)| \geq \theta\right\} \ll_{n,v,I,\varepsilon} q^{n-1-4\theta+\varepsilon}.$$

On GL_2, this result is due to Iwaniec [Iwa90]; see also the density theorems of Humphries [Hum18].

A consequence of the two different methods leading to these pointwise and on-average results is that density theorems like Theorem 5.1 fail to see[1] the sharp cutoffs in (5.1). In particular, the upper bound in Theorem 5.1 is always $q^{n-O(1)}$, although (5.1) implies that the set is empty for $\theta > \frac{1}{2} - \frac{1}{n^2+1}$. Given the great difficulty of improving the pointwise bounds, one may hope for a density bound that decays more smoothly as

[1] Blomer [Blo23] also remarks that it is not clear how to combine their spectral methods with the L-function techniques of Luo–Rudnick–Sarnak [LRS99].

one approaches $\theta \approx \frac{1}{2}$ (naturally, this is the most relevant range for applications where the uniform bound (5.1) is currently used).

In this chapter, we prove density theorems for the local parameters of cuspidal automorphic representations of $\mathrm{GL}_n(\mathbb{A}_{\mathbb{Q}})$, using the L-function techniques that were previously applied to the pointwise bounds; to achieve this, we study averages over π and π' of coefficients of Rankin–Selberg L-functions $L(s, \pi \times \widetilde{\pi}')$. Such an approach is likely to achieve a smoother decay towards the thresholds in (5.1), and thus to improve results like Theorem 5.1 for large values of θ. We state our main result using the notation from Section 2.3 (for the base field $F = \mathbb{Q}$), which normalizes the central characters of automorphic representations to avoid duplicates by Archimedean twists.

Theorem 5.2. *Let $n \in \mathbb{Z}_+$, v be a place of $\mathbb{Q}$, $\mathcal{S} \subset \mathfrak{F}_n$ be a finite subset of cuspidal automorphic representations unramified at v, and $\mathfrak{C} := \max_{\pi \in \mathcal{S}} \mathfrak{C}_\pi$ (these are the total conductors, defined as in (2.50)). Then for any $\varepsilon > 0$ and any $\theta \in (0, \frac{1}{2})$, one has*

$$\# \left\{ \pi \in \mathcal{S} : \max_j |\mathrm{Re}\, \mu_{\pi,j}(v)| \geq \theta \right\} \ll_{n,v,\theta,\varepsilon} \mathfrak{C}^{n\frac{1-2\theta}{2\theta}+\varepsilon}.$$

In fact, one can replace the base $\mathfrak{C}^n$ in the right-hand side with $\max_{\pi,\pi' \in \mathcal{S}} \mathfrak{C}_{\pi \times \widetilde{\pi}'}^{1/2}$.

Remark. The more explicit dependency on $\max_{\pi,\pi' \in \mathcal{S}} \mathfrak{C}_{\pi \times \widetilde{\pi}'}^{1/2}$ in Theorem 5.2 is advantageous for 'close-knit' families (in the sense of [PY23, §1.5]), which have comparatively-small Rankin–Selberg conductors. This is the case for families of representations with the same arithmetic conductor, and perhaps also the same central character (as in Theorem 5.1); see the improved Bushnell–Henniart bounds in [BTZ22, Appendix B].

A smaller close-knit family is $\mathcal{S} := \{\pi \otimes \chi : \chi \in \Xi\}$, where $\pi \in \mathfrak{F}_n$ is fixed and Ξ contains all the primitive even Dirichlet characters of a large prime conductor $q \nmid \mathfrak{q}_\pi$. Then $|\mathcal{S}| \asymp q$, and applying Theorem 5.2 with $\theta := \max_j |\mathrm{Re}\, \mu_{\pi,j}(v)|$ implies that $q \ll_{n,v,\theta,\varepsilon} (\mathfrak{C}_\pi^n q^{n^2/2})^{\frac{1-2\theta}{2\theta}+\varepsilon}$. Letting $q \to \infty$, this gives a contradiction unless

$$\theta \leq \frac{1}{2} - \frac{1}{n^2+2}.$$

In fact, in running the proof of Theorem 5.2 for this particular family, one can apply Deligne's bound for hyper-Kloosterman sums as in [DI90; LRS95; Kim03] to get an additional square-root cancellation over the Dirichlet characters; this would ultimately recover the pointwise bound $\theta_n \leq \frac{1}{2} - \frac{1}{n^2+1}$ of Luo–Rudnick–Sarnak [LRS95; LRS99] from (2.44). See Section 5.4.3 for further discussion on the role of character twists.

Remark. When $v = p < \infty$, Theorem 5.2 holds, with the same proof, for the Iwaniec–Sarnak notion of analytic conductors [IS00] (i.e., with (2.48) instead of (2.50)). In Corollary 6.2 from Chapter 6, we will remove the constraint that π_p is unramified, and compare our bound to the size of subfamilies of $\mathfrak{F}_n$ ordered by analytic conductor.

Finally, let us compare our result to previous density theorems. We will focus on Blomer's Theorem 5.1, but a similar comparison holds, e.g., with Jana's density result for $\mathrm{PGL}_n(\mathbb{Z})$ [Jan21, Theorem 3] (notably, the latter matches Sarnak's density hypothesis, while Blomer's Theorem 5.1 goes beyond it).

It can be deduced from Sections 2.3.1 and 2.3.2 that the representations considered in Theorem 5.1 have total conductors of size $\mathfrak{C}_\pi \ll_I q$, so $\mathcal{S}_I(q) \subset \mathcal{S}(\mathfrak{C})$ for some $\mathfrak{C} \asymp_I q$. Therefore, Theorem 5.2 gains over Theorem 5.1 in two key aspects:

(1). The upper bound in Theorem 5.2 is better for large values of θ (near $\frac{1}{2}$) and n. In particular, at the pointwise threshold of $\theta = \frac{1}{2} - \frac{1}{n^2+1}$ from (5.1), the upper bounds in Theorem 5.1, respectively Theorem 5.2, have the shape

$$q^{O(n)}, \qquad \text{respectively} \qquad q^{O(1/n)}.$$

(2). Theorem 5.2 includes all conductors $\mathfrak{C}_\pi \leq \mathfrak{C}$ (corresponding to all levels $q \leq Q$), and makes meaningful use of the average over different conductors. Relatedly, Theorem 5.2 allows for arbitrary central characters, i.e., for including all nebentypen $\chi \pmod d$, $d \mid q$.

Let us detail this analysis; suppose first that $n \geq 3$. Given $Q \geq 1$ and $\theta \in (0, 1/2)$, consider how many Hecke Maass forms for GL_n, with levels up to Q and Laplacian eigenvalues in a fixed interval, have $\max_j |\mathrm{Re}\, \mu_{\pi,j}(v)| \geq \theta$. Summing Theorem 5.1 over $q \leq Q$ gives a bound of $Q^{n-4\theta+o(1)}$. In this context, Theorem 5.2 (with $\mathfrak{C} \asymp_I Q$) beats Theorem 5.1 whenever

$$n\frac{1-2\theta}{2\theta} < n - 4\theta \qquad \Longleftrightarrow \qquad \theta > \frac{n - \sqrt{(n-2)n}}{4},$$

which approaches the range $\theta \geq \frac{1}{4}$ as $n \to \infty$. The barrier at $\frac{1}{4}(n - \sqrt{(n-2)n})$ is always below the pointwise threshold from (5.1), so our Theorem 5.2 gives a new result for $n \geq 3$.

When $n = 2$, one should first take the symmetric fourth power lifts [Kim03] of the GL_2 representations considered, and then apply Theorem 5.2 for the resulting family of GL_5 representations. Currently, this fails to beat the existing density theorems [Hum18] for θ below the 7/64 threshold of Kim–Sarnak, even when summing over all levels $q \leq Q$ and all nebentypen $\chi \pmod d$, $d \mid q$. However, it is likely that Theorem 5.2 may be improved using similar ideas, and it would be interesting to obtain new results in the GL_2 setting this way. For applications, it would be most relevant to obtain a nontrivial density theorem with only one form per level, as below; such results seem inaccessible to spectral methods, which rely on a trace formula for a single congruence subgroup.

Open Problem 1. *Let v be a place of $\mathbb{Q}$. For $q \in \mathbb{Z}_+$, let $\theta_q := \max_{\pi,j} |\operatorname{Re} \mu_{\pi,j}(v)|$, where π ranges over cuspidal automorphic representations of $\mathrm{GL}_2(\mathbb{A}_\mathbb{Q})$, generated by $\Gamma_0(q)$-invariant Maass forms. Show that, for all $Q \geq 1$ and some explicit $\varepsilon, \delta > 0$,*

$$\# \left\{ q \leq Q : \theta_q > \frac{7}{64} - \delta \right\} \ll_v Q^{1-\varepsilon}.$$

In particular, this means beating the Kim–Sarnak bounds [Kim03, Appendix 2] for almost all levels q.

Indeed, such a density theorem with averaging over the levels could be combined with fixed-level *large sieve inequalities* for exceptional Maass forms, which arise in the dispersion method [DI82c; Dra17] as in Chapters 3 and 4 (these are similar to density theorems, but incorporate additional information about the orthogonality of coefficients, which is useful in bounding multilinear forms of Kloosterman sums)[2]. This would automatically improve several results which currently rely on the pointwise bounds of Kim–Sarnak [Top18; BD20; DPR23; Lic23; Wu23]. We note that the dependency of the implied constant on v can be important in applications.

5.2 Outline

This section gives a brief and informal outline of our method. We refer the reader to Section 2.3 for background on automorphic and Rankin–Selberg L-functions.

Let v be a fixed place of $\mathbb{Q}$. First, consider the problem of bounding the local parameters $\mu_{\pi,j} = \mu_{\pi,j}(v)$ of a given cuspidal automorphic representation π for $\mathrm{GL}_n(\mathbb{A}_\mathbb{Q})$, unramified at v, as in (5.1). Recall that these parameters appear in the local factors $L_v(s, \pi)$ and $L_v(s, \pi \times \widetilde{\pi})$.

The "trivial" bound $|\operatorname{Re} \mu_{\pi,j}| \leq \frac{1}{2}$ follows from the fact that the local factors $L_v(s, \pi \times \widetilde{\pi})$ have no poles in $\operatorname{Re} s > 1$ [JS81]. But there is another simple, global argument for this bound (which is somewhat redundant in this setting, but generalizes well): for any $\ell \geq 1$, one has

$$\lambda_{\pi \times \widetilde{\pi}}(\ell) \leq \sum_{m \sim \ell/2} \lambda_{\pi \times \widetilde{\pi}}(m) \ll \ell^{1+o(1)}, \tag{5.2}$$

by the nonnegativity of the Rankin–Selberg coefficients $\lambda_{\pi \times \widetilde{\pi}}(m)$ [RS96] and the absolute convergence of the Dirichlet series of $L(s, \pi \times \widetilde{\pi})$ in $\operatorname{Re} s > 1$. When $v = p < \infty$, taking $\ell = p^k$, the coefficient $\lambda_{\pi \times \widetilde{\pi}}(\ell)$ grows (at least on a subsequence of k's) like

[2] There exist large sieve inequalities for exceptional Maass forms which use averaging over levels (see [DI82c, Theorems 6, 7], [Wat95]), but the same techniques do not seem to apply to proving density theorems with averaging over levels.

$\ell^{2\max_j |\mathrm{Re}\,\mu_{\pi,j}(p)|}$; so letting $k \to \infty$, this gives a contradiction unless $|\mathrm{Re}\,\mu_{\pi,j}(p)| \le \frac{1}{2}$. Similarly, when $v = \infty$, one can instead bound

$$\ell^\beta \lambda_{\pi\times\widetilde{\pi}}(1) \le \sum_{m\le\ell} \lambda_{\pi\times\widetilde{\pi}}(m) \left(\frac{\ell}{m}\right)^\beta \ll \ell^{1+o(1)}, \tag{5.3}$$

for $\beta \in \mathbb{R}$ such that $L(\beta, \pi\times\widetilde{\pi}) = 0$, by a contour-shifting argument to $\mathrm{Re}\,s = 1+\varepsilon$ (namely, the zero of $L(s,\pi\times\widetilde{\pi})$ at $s=\beta$ cancels the pole of a Mellin-transformed smooth majorant; see Lemma 5.6). Taking $\beta = \mu_{\pi,j} + \overline{\mu}_{\pi,j}$, which must be a zero of $L(s,\pi\times\widetilde{\pi})$ to cancel the corresponding pole of $L(s,\pi_\infty\times\widetilde{\pi}_\infty)$ (see (2.52)), and letting $\ell\to\infty$, one recovers the bound $|\mathrm{Re}\,\mu_{\pi,j}(\infty)| \le \frac{1}{2}$.

One can improve this to $|\mathrm{Re}\,\mu_{\pi,j}| \le \frac{1}{2} - \frac{1}{n^2+1}$, as anticipated in (5.1), by considering a family of twisted L-functions. Indeed, the classical method of Landau–Serre [Ser81] essentially uses twists by Archimedean characters to achieve this bound at the finite places of $\mathbb{Q}$. Luo–Rudnick–Sarnak [LRS95] used twists by Dirichlet characters to obtain results of the same strength at the infinite place, and their method extends to general number fields [LRS99]. In the latter setting, one restricts the sums from (5.2) and (5.3) to residue classes $m \equiv \pm\ell \pmod q$ (resp., $m \equiv \pm 1 \pmod q$) for a large prime q, and detects the congruences by Dirichlet characters $\chi \pmod q$; one then employs the properties of the twisted L-functions $L(s, (\pi\otimes\chi)\times\widetilde{\pi})$, along with Deligne's bounds for hyper-Kloosterman sums.

When $n \le 4$, one can work with the symmetric square $\mathrm{Sym}^2\pi$ instead of $\pi\times\widetilde{\pi}$, which improves the bound to $|\mathrm{Re}\,\mu_{\pi,j}| \le \frac{1}{2} - \frac{1}{n(n+1)/2+1}$, using related ideas of Duke–Iwaniec [DI90]. When $n=2$, one can also combine such results with symmetric power lifts [Kim03, Appendix 2]. This explains (5.1).

Our Theorem 5.2 uses neither twists by Dirichlet characters nor symmetric squares; see Section 5.4.3 for a discussion of the potential role of character twists in our estimates. Rather, we insert averaging over representations $\pi, \pi' \in \mathcal{S}$ in the simpler argument from (5.2) and (5.3), working with the Rankin–Selberg L-functions $L(s,\pi\times\widetilde{\pi}')$. The *diagonal terms* with $\pi = \pi'$ can be treated as before, but constitute only a $|\mathcal{S}|^{-1}$-fraction of the total sum. In the *off-diagonal terms* with $\pi \ne \pi'$, we obtain savings from the fact that $L(s,\pi\times\widetilde{\pi}')$ has no poles – so essentially, from the orthogonality of the coefficients of $L(s,\pi)$ and $L(s,\pi')$. This is akin to the proof of mean-value estimates in [DK00].

However, the pointwise argument in (5.2) and (5.3) depended on the fact that the coefficients $\lambda_{\pi\times\widetilde{\pi}}(m)$ are nonnegative[3]. Our argument uses, as a key input, a generalization of this fact: the Rankin–Selberg coefficients $\lambda_{\pi\times\widetilde{\pi}'}(m)$ form a *positive semi-definite* matrix in $\pi,\pi' \in \mathcal{S}$. In other words, for any weights $w_\pi \in \mathbb{C}$ and any positive

[3]Even when using $\mathrm{Sym}^2\pi$ instead of $\pi\times\widetilde{\pi}$, one deduces the absolute convergence of the Dirichlet series of $L(s,\mathrm{Sym}^2\pi)$ in $\mathrm{Re}\,s > 1$ from that of $L(s,\pi\times\widetilde{\pi})$.

integer m, one has

$$\sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} \lambda_{\pi \times \widetilde{\pi}'}(m) \geq 0, \tag{5.4}$$

as we show Proposition 5.5 (and Section 5.5). This allows one to bound

$$\left| \sum_{m \leq M} u_m \sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} \lambda_{\pi \times \widetilde{\pi}'}(m) \right| \leq \|u\|_\infty \sum_{m=1}^{\infty} \Phi\left(\frac{m}{M}\right) \sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} \lambda_{\pi \times \widetilde{\pi}'}(m),$$

for any complex sequence (u_m) and a suitable smooth majorant Φ; the inner sums in the right-hand side can then be expressed in terms of the L-functions $L(s, \pi \times \widetilde{\pi}')$. We consider such trilinear sums over m, π, π', with additional weights of $(M/m)^{\beta_\pi + \overline{\beta}_{\pi'}}$, in Proposition 5.7. We will ultimately apply this for the (very sparse!) sequences $u_m = \mathbb{1}_{m=\ell}$, respectively $u_m = \mathbb{1}_{m=1}$ (with $M = \ell$).

This argument leads to better bounds for averages like $\frac{1}{|\mathcal{S}|^2} \sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} \lambda_{\pi \times \widetilde{\pi}'}(\ell)$, which can be exploited to produce density theorems: rather than seeking a contradiction when $\max_j |\operatorname{Re} \mu_{\pi,j}|$ is too large, we seek an upper bound for $|\mathcal{S}|$ in terms of the smallest local parameter $\min_{\pi \in \mathcal{S}} \max_j |\operatorname{Re} \mu_{\pi,j}|$ and the largest total conductor $\max_{\pi \in \mathcal{S}} \mathfrak{C}_\pi$. The conductor aspect is crucial in such results (unlike in the pointwise bounds from [LRS95; RS96; Kim03]), so we need to make all dependencies on it explicit; the convexity bounds of Li [Li10] are helpful here. Another difficulty is that we cannot simply let various parameters tend to ∞ as in [LRS95; RS96; Kim03]; we will need to carefully optimize such parameters. Thus for instance, in our results at finite places, concluding the argument requires a more explicit lower bound in (5.4) when $m = p^k$, together with Turán's lower bounds for power sums.

5.2.1 Structure

For the rest of this chapter, the reader should be familiar with the L-function notation from Section 2.3, in the case $F = \mathbb{Q}$. We identify the nonzero ideals of $\mathcal{O}_\mathbb{Q} = \mathbb{Z}$ with the positive integers, the finite places of $\mathbb{Q}$ with the primes, and use '∞' to denote the unique Archimedean place.

In Section 5.3, we study sums over $\pi, \pi' \in \mathcal{S}$ of Rankin–Selberg coefficients; in particular, we establish the aforementioned Propositions 5.5 and 5.7, which may be of independent interest. In Sections 5.4.1 and 5.4.2, we carry out the argument described above, with averaging over $\pi, \pi' \in \mathcal{S}$, to prove density theorems for the local parameters of GL_n automorphic representations at $v = p < \infty$ and then $v = \infty$. We discuss some limitations and potential improvements in Section 5.4.3. A technical proof of the positive semi-definite property of Rankin–Selberg L-functions at ramified places is left to Section 5.5.

5.3 Families of Rankin–Selberg L-functions

5.3.1 Positive semi-definite coefficients of L-functions

As discussed in Section 5.2, it is often a helpful property that the Dirichlet coefficients of a given L-function are nonnegative. When dealing with families of L-functions, it is desirable to generalize this property to produce nonnegative averages of coefficients. A natural way to proceed is to consider positive semi-definite matrices.

Recall that a Hermitian matrix $M = (M_{i,j}) \in \mathbb{C}^{N\times N}$ is *positive semi-definite* iff all eigenvalues of M are nonnegative, equivalently $\vec{v}^* M \vec{v} \in \mathbb{R}_{\geq 0}$ for all vectors $\vec{v} \in \mathbb{C}^N$.

Definition 5.3 (Positive semi-definite families)**.** Let $\mathcal{I}$ be a finite ordered sequence. For $i, j \in \mathcal{I}$, let $L_{i,j}(s) = \sum_{m=1}^{\infty} \lambda_{i,j}(m)\, m^{-s}$ be a formal Dirichlet series with complex coefficients. We say that the family $(L_{i,j}(s))_{i,j\in\mathcal{I}}$ is *positive semi-definite* iff for any $m \geq 1$, the matrix $M \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ with entries

$$M_{i,j} := \lambda_{i,j}(m)$$

is (Hermitian and) positive semi-definite. Note that this is independent of the ordering of the sequence $\mathcal{I}$. When applied to complex L-functions, this definition refers to their Dirichlet expansions in $\operatorname{Re} s > \sigma$, for large enough σ.

Recall that a matrix is positive semi-definite iff it may be written as a positive linear combination of rank-1 matrices of the form $\vec{w}\vec{w}^*$, e.g., via its eigendecomposition (here $\vec{w}^*$ denotes the conjugate transpose of a complex vector $\vec{w}$). It follows that $(L_{i,j}(s))_{i,j\in\mathcal{I}}$ is a positive semi-definite family of L-functions iff $L_{i,j}(s)$ is a positive linear combination of terms

$$w_i \overline{w}_j\, m^{-s}, \qquad \text{where } \vec{w} = (w_i) \in \mathbb{C}^{\mathcal{I}},\ m \in \mathbb{Z}_+. \tag{5.5}$$

Lemma 5.4 (L-function operations preserve positive semi-definiteness)**.** *Let $\mathcal{I}$ be a finite ordered sequence and $c \geq 0$; let $(L_{i,j}(s))_{i,j\in\mathcal{I}}$ and $(L_{i,j}^{(k)}(s))_{i,j\in\mathcal{I}}$ be positive semi-definite families of L-functions, for $k \geq 1$. Then the families*

$$(cL_{i,j}(s))_{i,j\in\mathcal{I}}, \qquad \left(L_{i,j}^{(1)}(s) + L_{i,j}^{(2)}(s)\right)_{i,j\in\mathcal{I}}, \qquad \text{and} \qquad \left(L_{i,j}^{(1)}(s) \cdot L_{i,j}^{(2)}(s)\right)_{i,j\in\mathcal{I}}$$

are positive semi-definite. Moreoever, if there exists a family of formal Dirichlet series $(L_{i,j}^{(\infty)})_{i,j\in\mathcal{I}}$ such that $L_{i,j}^{(k)} \to L_{i,j}^{(\infty)}$ (in the sense of pointwise convergence of Dirichlet coefficients), then the limit family $(L_{i,j}^{(\infty)})_{i,j\in\mathcal{I}}$ is also positive semi-definite. In particular, if well-defined, then the families

$$\left(\sum_{k=1}^{\infty} L_{i,j}^{k}(s)\right)_{i,j\in\mathcal{I}}, \qquad \left(\prod_{k=1}^{\infty} L_{i,j}^{k}(s)\right)_{i,j\in\mathcal{I}}, \qquad \text{and} \qquad (\exp(L_{i,j}(s)))_{i,j\in\mathcal{I}}.$$

are positive semi-definite.

Remark. The last fact in Lemma 5.4 can be rephrased as follows: to show that the L-functions $L_{i,j}$ form a positive semi-definite family, it suffices to show the same for their formal logarithms $\log L_{i,j}$.

Proof. The facts that a positive scaling of a positive semi-definite family, a sum of two positive semi-definite families, and the limit of a sequence of positive semi-definite families are positive semi-definite follow immediately from the corresponding matrix properties. For the product of two L-functions $L_{i,j}^{(k)} = \sum_{m\geq 1} \lambda_{i,j}^k(m)\, m^{-s}$, we note that the Dirichlet coefficients of $L_{i,j}^{(1)}(s) \cdot L_{i,j}^{(2)}(s)$ are given by

$$\lambda_{i,j}(m) = \sum_{d|m} \lambda_{i,j}^{(1)}(d)\, \lambda_{i,j}^{(2)}\left(\frac{m}{d}\right),$$

and the claim follows from Schur's product theorem that the Hadamard product of two positive semi-definite matrices is positive semi-definite. Equivalently, the fact that $(L_{i,j}^{(1)}(s) \cdot L_{i,j}^{(2)}(s))_{i,j\in\mathcal{I}}$ is positive semi-definite is apparent from the characterization in (5.5).

The last claim (about infinite sums, products, and exponentials) follows from the previous properties, noting that $\exp L(s) = \sum_{k\geq 0} \frac{1}{k!} L(s)^k$. □

We apply this notion to families of Rankin–Selberg L-functions. The following result is closely related to the computations of Brumley in [ST19, Appendix].

Proposition 5.5 (Rankin–Selberg L-functions are positive semi-definite)**.** *For any finite set $\mathcal{S} \subset \mathfrak{F}_n$, the family $(\log L(s, \pi \times \widetilde{\pi}'))_{\pi,\pi'\in\mathcal{S}}$ is positive semi-definite. In particular, so is $(L(s, \pi \times \widetilde{\pi}'))_{\pi,\pi'\in\mathcal{S}}$; thus for any $m \in \mathbb{Z}_+$ and $w_\pi \in \mathbb{C}$, one has*

$$\sum_{\pi,\pi'\in\mathcal{S}} w_\pi \overline{w}_{\pi'}\, \lambda_{\pi\times\widetilde{\pi}'}(m) \geq 0.$$

In fact, if $m = p^k$ is a prime power, where $k \geq 1$ and p is an unramified prime for all $\pi \in \mathcal{S}$, one has

$$\sum_{\pi,\pi'\in\mathcal{S}} w_\pi \overline{w}_{\pi'}\, \lambda_{\pi\times\widetilde{\pi}'}(p^k) \geq \frac{1}{k} \left| \sum_{\pi\in\mathcal{S}} w_\pi \sum_{j=1}^{n} \alpha_{\pi,j}(p)^k \right|^2. \tag{5.6}$$

Proof. In what follows we work with the Dirichlet expansions of Rankin–Selberg L-functions in $\operatorname{Re} s > 1$; one may alternatively treat these as formal Dirichlet series. By Lemma 5.4, positive semi-definiteness can be verified locally: i.e., it suffices to show that each local factor in $\log L(s, \pi \times \widetilde{\pi}') = \sum_p \log L(s, \pi_p \times \widetilde{\pi}'_p)$ forms a positive semi-definite family in π, π'.

If p is unramified for both π and π', by [RS96, (2.18)] the (formal) logarithm of local factor has the form

$$\begin{aligned}\log L(s, \pi_p \times \widetilde{\pi}'_p) &= \log \prod_{j,j'=1}^{n} \left(1 - \frac{\alpha_{\pi,j}(p)\,\overline{\alpha}_{\pi',j'}(p)}{p^s}\right)^{-1} \\ &= -\sum_{j,j'=1}^{n} \log\left(1 - \frac{\alpha_{\pi,j}(p)\,\overline{\alpha}_{\pi',j'}(p)}{p^s}\right) \\ &= \sum_{j,j'=1}^{n} \sum_{q=1}^{\infty} \frac{\left(\alpha_{\pi,j}(p)\,\overline{\alpha}_{\pi',j'}(p)\right)^q}{qp^{qs}} \\ &= \sum_{q=1}^{\infty} \frac{\left(\sum_{j=1}^{n} \alpha_{\pi,j}(p)^q\right)\overline{\left(\sum_{j=1}^{n} \alpha_{\pi',j}(p)^q\right)}}{qp^{qs}},\end{aligned}$$

which is clearly positive semi-definite in π, π'. In fact, if p is unramified for all $\pi \in \mathcal{S}$, it follows that

$$\begin{aligned}&\sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} \sum_{k=0}^{\infty} \frac{\lambda_{\pi \times \widetilde{\pi}'}(p^k)}{p^{ks}} \\ &= \sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} \exp\left(\sum_{q=1}^{\infty} \frac{\left(\sum_{j=1}^{n} \alpha_{\pi,j}(p)^q\right)\overline{\left(\sum_{j=1}^{n} \alpha_{\pi',j}(p)^q\right)}}{qp^{qs}}\right) \\ &= \sum_{\ell=1}^{\infty} \frac{1}{\ell!} \sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} \left(\sum_{q=1}^{\infty} \frac{\left(\sum_{j=1}^{n} \alpha_{\pi,j}(p)^q\right)\overline{\left(\sum_{j=1}^{n} \alpha_{\pi',j}(p)^q\right)}}{qp^{qs}}\right)^{\ell}.\end{aligned}$$

Each of the inner Dirichlet series is positive semi-definite in π, π' in the sense of Definition 5.3. Thus for $k \geq 1$, identifying coefficients of p^{-ks} and dropping all terms except for $\ell = 1$ by nonnegativity, we obtain

$$\begin{aligned}\sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'}\, \lambda_{\pi \times \widetilde{\pi}'}(p^k) &\geq \sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} \frac{\left(\sum_{j=1}^{n} \alpha_{\pi,j}(p)^k\right)\overline{\left(\sum_{j=1}^{n} \alpha_{\pi',j}(p)^k\right)}}{k} \\ &= \frac{1}{k}\left|\sum_{\pi \in \mathcal{S}} w_\pi \sum_{j=1}^{n} \alpha_{\pi,j}(p)^k\right|^2.\end{aligned}$$

This proves (5.6). It remains to show that $\log L(s, \pi \times \widetilde{\pi}')$ forms a positive semi-definite family in π, π' when p is an arbitrary prime, which may be ramified for some $\pi \in \mathcal{S}$. This follows almost immediately from [ST19, Formula (A.8)], after explicitating J_a, z_j and K_b, z'_k as functions of π and π'; but for completeness, we include a proof in Section 5.5. □

Remark. Taking $|\mathcal{S}| = 1$ in Proposition 5.5 recovers the fact that the 'diagonal' Rankin–Selberg L-functions $L(s, \pi \times \widetilde{\pi})$ have nonnegative Dirichlet coefficients [RS96], i.e. for all $m \geq 1$,

$$\lambda_{\pi \times \widetilde{\pi}}(m) \geq 0.$$

Taking $|\mathcal{S}| = 2$, say $\mathcal{S} = \{\pi, \pi'\}$, shows that the matrix

$$\begin{pmatrix} \lambda_{\pi \times \widetilde{\pi}}(m) & \lambda_{\pi \times \widetilde{\pi}'}(m) \\ \lambda_{\pi' \times \widetilde{\pi}}(m) & \lambda_{\pi' \times \widetilde{\pi}'}(m) \end{pmatrix}$$

has nonnegative eigenvalues (in particular, nonnegative determinant), so

$$|\lambda_{\pi \times \widetilde{\pi}'}(m)| \leq \sqrt{\lambda_{\pi \times \widetilde{\pi}}(m)\, \lambda_{\pi' \times \widetilde{\pi}'}(m)}.$$

Applying the same argument for the Dirichlet coefficients of $\log L(s, \pi \times \widetilde{\pi}')$ recovers [ST19, Lemma 2.2, first part]. But in this work, we will only use Proposition 5.5 for large families $\mathcal{S}$ of representations.

Remark. Writing $\log L(s, \pi \times \widetilde{\pi}') = \sum_{m=1}^{\infty} \frac{b_{\pi \times \widetilde{\pi}'}}{m^s}$ in $\operatorname{Re} s > 1$, the positive semi-definiteness property in Proposition 5.5 states that $\sum_{\pi, \pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} b_{\pi \times \widetilde{\pi}'} \geq 0$, for any complex weights (w_π). If one was only interested in the case $w_\pi \equiv 1$, this would follow by considering the isobaric sum of all $\pi \in \mathcal{S}$ (and the Rankin–Selberg convolution with its contragredient). Morally, our situation corresponds to a "weighted isobaric sum" with complex weights.

5.3.2 Triple sums of Rankin–Selberg coefficients

In the previous subsection, we looked at sums over π and π' of the Rankin–Selberg coefficients $\lambda_{\pi \times \widetilde{\pi}'}(m)$. Here we insert an additional sum over m, searching for upper bounds; we begin with the following lemma. We recall that n is fixed, and the notation $\mathfrak{F}_n$ from Section 2.3 for the base field $F = \mathbb{Q}$.

Lemma 5.6 (Duality + convexity bound)**.** *Let $\pi, \pi' \in \mathfrak{F}_n$. Let*[4] *$\Phi(x) := x^\beta e^{-x^2}$, for some $\beta \in \mathbb{C}$ such that one of the following is true:*

(i). $0 < \operatorname{Re} \beta \asymp 1$ and $\operatorname{Im} \beta \ll 1$, or

(ii). π, π' are unramified at ∞, and $-\beta = \mu_{\pi \times \widetilde{\pi}', j, j'}(\infty)$ for some $1 \leq j, j' \leq n$.

Then for any $M \gg 1$, one has

$$\sum_{m=1}^{\infty} \Phi\left(\frac{m}{M}\right) \lambda_{\pi \times \widetilde{\pi}'}(m) \ll (M \mathfrak{C}_{\pi \times \widetilde{\pi}'})^{o(1)} \cdot \begin{cases} M, & \text{if } \pi = \pi' \\ \min(M, \sqrt{\mathfrak{C}_{\pi \times \widetilde{\pi}'}}), & \text{otherwise.} \end{cases} \tag{5.7}$$

[4] We use this explicit choice of Φ to have fine control over its Mellin transform.

Remark. Condition (i) in Lemma 5.6 will ultimately be relevant for our estimates at the finite places, while condition (ii) will be relevant for the infinite place.

Proof of Lemma 5.6. We first note that condition (ii) implies $|\operatorname{Re}\beta| \leq 1$; so under either of conditions (i) and (ii), we have $-1 \leq \operatorname{Re}\beta \ll 1$ and

$$\operatorname{Im}\beta \ll 1 + \max_{1\leq j,j'\leq n} |\mu_{\pi\times\widetilde{\pi}',j,j'}(\infty)|. \tag{5.8}$$

Let $\varepsilon \in (0, \frac{1}{2})$; if (i) holds, we also take $\varepsilon < \operatorname{Re}\beta$. Moreover, we can assume that $|\operatorname{Re}\beta - \varepsilon| \geq \varepsilon/2$ by substituting $\varepsilon \leftarrow \varepsilon/10$ if necessary. By replacing π with $\pi[-iu] = \pi \otimes |\cdot|^{-iu}$ and β with $\beta - iu$ where $u = \operatorname{Im}\beta$, we may also assume that

$$\operatorname{Im}\beta = 0.$$

Indeed, these substitutions make no impact on conditions (i) and (ii), and affect the left-hand side of (5.7) only by a factor of $M^{i\operatorname{Im}\beta}$, and the right hand side of (5.7) only by a constant. For the last claim, note that (2.59) and (5.8) imply[5]

$$\mathfrak{C}_{\pi[-i\operatorname{Im}\beta]\times\widetilde{\pi}'} \ll \mathfrak{C}_{\pi\times\widetilde{\pi}}.$$

We now use Mellin inversion, as in (2.4), to expand

$$\sum_{m=1}^{\infty} \Phi\left(\frac{m}{M}\right) \lambda_{\pi\times\widetilde{\pi}'}(m) = \frac{1}{2\pi i}\int_{(2)} M^s L(s, \pi\times\widetilde{\pi}')\,\widetilde{\Phi}(s)\,ds. \tag{5.9}$$

Here we can explicitly compute (in $\operatorname{Re}(s+\beta) > 0$, and by meromorphic continuation elsewhere)

$$\begin{aligned}\widetilde{\Phi}(s) &= \int_0^\infty x^{s-1+\beta} e^{-x^2} dx \\ &= \frac{1}{2}\int_0^\infty y^{(s-1+\beta)/2} e^{-y} y^{-1/2} dy = \frac{1}{2}\Gamma\left(\frac{s+\beta}{2}\right).\end{aligned}$$

In particular, (2.5) and $\operatorname{Re}\beta \ll 1, \operatorname{Im}\beta = 0$ imply that for $\sigma \ll 1$ and $\min_{m\in 2\mathbb{Z}_{\leq 0}} |\sigma + it + \beta - m| \geq \varepsilon/2$, one has

$$\widetilde{\Phi}(\sigma+it) \asymp_\varepsilon \left(1+\frac{|t|}{2}\right)^{\frac{\sigma+\operatorname{Re}\beta-1}{2}} e^{-\frac{\pi}{4}|t|} \ll e^{-\frac{1}{2}|t|}. \tag{5.10}$$

Since $L(s, \pi\times\widetilde{\pi}')$ has moderate vertical growth, we may shift contours. We first shift to $\operatorname{Re} s = 1+\varepsilon$ to obtain

$$\begin{aligned}\sum_{m=1}^{\infty} \Phi\left(\frac{m}{M}\right) \lambda_{\pi\times\widetilde{\pi}'}(m) &= \frac{1}{2\pi i}\int_{(1+\varepsilon)} M^s L(s, \pi\times\widetilde{\pi}')\,\widetilde{\Phi}(s)\,ds \\ &\ll_\alpha M^{1+\varepsilon} \max_{t\in\mathbb{R}} \frac{L(1+\varepsilon+it, \pi\times\widetilde{\pi}')}{1+|t|}.\end{aligned}$$

[5]This is the critical step where we need total conductors as in (2.59) rather than analytic conductors as in (2.57), and this only really affects our results towards GRC at the Archimedean place.

But by the convexity bound of Li from Lemma 2.14 (noting $C_{\pi\times\widetilde{\pi}'} \leq \mathfrak{C}_{\pi\times\widetilde{\pi}'}$), we have

$$L(1+\varepsilon+it, \pi\times\widetilde{\pi}') = L(1+\varepsilon, \pi[it]\times\widetilde{\pi}') \ll_\varepsilon ((1+|t|)\,\mathfrak{C}_{\pi\times\widetilde{\pi}'})^\varepsilon, \tag{5.11}$$

where we recall the notation $\pi[z] = \pi\otimes|\cdot|^z$. Thus we always have

$$\sum_{m=1}^{\infty} \Phi\left(\frac{m}{M}\right)\lambda_{\pi\times\widetilde{\pi}'}(m) \ll_\varepsilon M^{1+\varepsilon}\mathfrak{C}_{\pi\times\widetilde{\pi}'}^{2\varepsilon}.$$

Now suppose that the original representations in $\mathfrak{F}_n$ are distinct, so we have $\pi[it] \neq \pi'$ for all $t\in\mathbb{R}$ (even after potentially twisting π by $|\cdot|^{-iu}$). Then $\Lambda(s, \pi\times\widetilde{\pi}')$ is entire, so $L(s, \pi\times\widetilde{\pi}')$ is also entire. If (ii) holds, then $L(s, \pi\times\widetilde{\pi}')$ has a zero at $s=-\beta$ to cancel the corresponding pole of $\Gamma_{\mathbb{R}}(s-\mu_{\pi\times\widetilde{\pi}',j,j'})$ inside $L_\infty(s, \pi\times\widetilde{\pi}')$.

Then we can shift the contour in (5.9) to $\operatorname{Re} s = -\varepsilon$, picking up no residues in the process. Indeed, the simple pole of $\widetilde{\Phi}(s)$ at $s=-\beta$ is either outside the contour integral (if (i) holds), or cancelled by the zero of $L(s, \pi\times\widetilde{\pi}')$ (if (ii) holds). The other poles at $s+\beta\in 2\mathbb{Z}_{<0}$ are also outside the contour since $-\varepsilon+\operatorname{Re}\beta > -\frac{1}{2}-1 > -2$. Thus

$$\begin{aligned}\sum_{m=1}^{\infty} \Phi\left(\frac{m}{M}\right)\lambda_{\pi\times\widetilde{\pi}'}(m) &= \frac{1}{2\pi i}\int_{(-\varepsilon)} M^s L(s, \pi\times\widetilde{\pi}')\,\widetilde{\Phi}(s)\,ds \\ &= \frac{1}{2\pi i}\int_{(-\varepsilon)} M^s L(1-s, \widetilde{\pi}\times\pi')\frac{L(s, \pi\times\widetilde{\pi}')}{L(1-s, \widetilde{\pi}\times\pi')}\,\widetilde{\Phi}(s)\,ds.\end{aligned}$$

Plugging in (5.11), (2.58) (noting again that analytic conductors are not larger than total conductors), and (5.10), the triangle inequality gives

$$\sum_{m} \Phi\left(\frac{m}{M}\right)\lambda_{\pi\times\widetilde{\pi}'}(m) \ll_\varepsilon \mathfrak{C}_{\pi\times\widetilde{\pi}'}^{\frac{1}{2}+2\varepsilon},$$

which completes our proof. □

Finally, using Lemma 5.6 and Proposition 5.5, we can prove our key estimate for exploiting the averaging over automorphic representations. This result may be of independent interest to the reader.

Proposition 5.7 (Triple sums of Rankin–Selberg coefficients)**.** *Let $\mathcal{S}\subset\mathfrak{F}_n$ be a finite set. Let $\mathfrak{C}_{RS} := \max_{\pi,\pi'\in\mathcal{S}}\mathfrak{C}_{\pi\times\widetilde{\pi}'}$ and $M\gg 1$. Let $(u_m)_{m\leq M}$ and $(\beta_\pi)_{\pi\in\mathcal{S}}$ be complex sequences such that one of the following holds.*

(i). For all $\pi\in\mathcal{S}$, $\beta_\pi \ll 1$, and (u_m) is supported on $m\asymp M$, or

(ii). All $\pi\in\mathcal{S}$ are unramified at ∞ and satisfy $\beta_\pi = \mu_{\pi,j}(\infty)$ for some $1\leq j\leq n$.

Then for any complex weights $(w_{\pi,\pi'})_{\pi,\pi'\in\mathcal{S}}$ forming a positive semi-definite matrix in $\mathbb{C}^{\mathcal{S}\times\mathcal{S}}$, one has

$$\sum_{\pi,\pi'\in\mathcal{S}} w_{\pi,\pi'} \sum_{m\leq M} u_m\, \lambda_{\pi\times\tilde{\pi}'}(m) \left(\frac{M}{m}\right)^{\beta_\pi+\overline{\beta}_{\pi'}} \ll (M\mathfrak{C}_{RS})^{o(1)}\, \|w\|_\infty |\mathcal{S}|^2\, \|u\|_\infty\, M \times \left(|\mathcal{S}|^{-1} + \left(1+\frac{M}{\sqrt{\mathfrak{C}_{RS}}}\right)^{-1}\right). \tag{5.12}$$

Remark. The first line from the right-hand side of (5.12) contains the 'trivial' bound, which can be achieved without the averaging over $\pi,\pi'\in\mathcal{S}$; the second line contains two saving factors: one from the diagonal terms $\pi=\pi'$, and one from the off-diagonal terms $\pi\neq\pi'$. Relatedly, if $w_{\pi,\pi'}$ are arbitrary complex weights, an application of Cauchy–Schwarz in π,π' combined with the argument below produces a similar bound as in (5.12), with the diagonal saving $|\mathcal{S}|^{-1}$ replaced by $|\mathcal{S}|^{-1/2}$.

Proof of Proposition 5.7. Let $\mathcal{B}$ denote the sum in the left-hand side of (5.12). By Proposition 5.5 and Schur's product theorem, the matrix $M\in\mathbb{C}^{\mathcal{S}\times\mathcal{S}}$ with entries

$$M_{\pi,\pi'} := w_{\pi,\pi'}\, \lambda_{\pi\times\tilde{\pi}'}(m)$$

is also positive semi-definite. Letting $\vec{v}\in\mathbb{C}^{\mathcal{S}}$ be the (column) vector with entries

$$v_\pi := \left(\frac{M}{m}\right)^{\overline{\beta}_\pi},$$

we thus have

$$\sum_{\pi,\pi'\in\mathcal{S}} w_{\pi,\pi'}\, \lambda_{\pi\times\tilde{\pi}'}(m) \left(\frac{M}{m}\right)^{\beta_\pi+\overline{\beta}_{\pi'}} = \vec{v}^{\,*} M \vec{v} \in \mathbb{R}_{\geq 0}.$$

By the triangle inequality, it follows that

$$\begin{aligned} |\mathcal{B}| &= \left| \sum_{m\leq M} u_m \sum_{\pi,\pi'\in\mathcal{S}} w_{\pi,\pi'}\, \lambda_{\pi\times\tilde{\pi}'}(m) \left(\frac{M}{m}\right)^{\beta_\pi+\overline{\beta}_{\pi'}} \right| \\ &\ll \|u\|_\infty \sum_{m=1}^{\infty} \Phi\left(\frac{m}{M}\right) \sum_{\pi,\pi'\in\mathcal{S}} w_{\pi,\pi'}\, \lambda_{\pi\times\tilde{\pi}'}(m) \left(\frac{M}{m}\right)^{\beta_\pi+\overline{\beta}_{\pi'}}, \end{aligned}$$

where we inserted a majorant given by the nonnegative smooth function

$$\Phi(x) := x^B e^{-\sqrt{x}}, \qquad B := \begin{cases} 1+2\max_{\pi\in\mathcal{S}} |\mathrm{Re}\,\beta_\pi|, & \text{if } (i) \text{ holds.} \\ 0, & \text{otherwise.} \end{cases}.$$

Note that having $B>0$ is acceptable if condition (i) in our assumption holds, since then (u_m) is supported in $m\asymp M$. Denoting $\Phi_{\pi,\pi'}(x) := \Phi(x)\, x^{-\beta_\pi-\overline{\beta}_{\pi'}}$, we thus have

$$|\mathcal{B}| \leq \|u\|_\infty \sum_{\pi,\pi'\in\mathcal{S}} w_{\pi,\pi'} \sum_{m=1}^{\infty} \Phi_{\pi,\pi'}\left(\frac{m}{M}\right) \lambda_{\pi\times\tilde{\pi}'}(m).$$

Now the inner sum over m satisfies the assumptions in Lemma 5.6, with $\beta = B - \beta_\pi - \overline{\beta_{\pi'}}$; indeed, if condition (i) or (ii) of Proposition 5.7 holds, then the corresponding condition of Lemma 5.6 holds. So applying Lemma 5.6 yields

$$\begin{aligned}|\mathcal{B}| &\ll (M\mathfrak{C}_{RS})^{o(1)}\|u\|_\infty \sum_{\pi,\pi'\in\mathcal{S}} |w_{\pi,\pi'}| \left(M\mathbb{1}_{\pi=\pi'} + \min(M, \mathfrak{C}_{RS}^{1/2})\right) \\ &= (M\mathfrak{C}_{RS})^{o(1)}\|u\|_\infty M \left(\sum_{\pi\in\mathcal{S}} |w_{\pi,\pi}| + \min\left(1, \frac{\sqrt{\mathfrak{C}_{RS}}}{M}\right) \sum_{\pi,\pi'\in\mathcal{S}} |w_{\pi,\pi'}|\right).\end{aligned}$$

Trivially bounding $\frac{1}{|\mathcal{S}|}\sum_{\pi\in\mathcal{S}}|w_{\pi,\pi}|$ and $\frac{1}{|\mathcal{S}|^2}\sum_{\pi,\pi'\in\mathcal{S}}|w_{\pi,\pi'}|$ by the ℓ^∞ norm completes our proof. □

5.4 The density theorems

5.4.1 The non-Archimedean case

Fix $n \geq 2$ and a place $v = p < \infty$. We aim for a density theorem for the local parameters of cuspidal automorphic representations at p, which we access through the Dirichlet coefficients at powers of p. The following power sum bound due to Turán [Mon94, Ch. 5] will be helpful in this process.

Lemma 5.8 (Turán's second theorem for power sums [Mon94])**.** *For any positive integers M, N and any complex numbers $z_1, \dots, z_N$ with $\max_j |z_j| \geq 1$, one has*

$$\max_{M+1\leq k\leq M+N} \left|\sum_{j=1}^N z_j^k\right| \gg_N \frac{1}{M^N} \max_j |z_j|^M.$$

Theorem 5.9 (Density at finite places)**.** *Let $\mathcal{S} \subset \mathfrak{F}_n$ be a finite set of cuspidal automorphic representations unramified at $v = p$, satisfying*

$$\forall \pi \in \mathcal{S}, \qquad \max_j |\mathrm{Re}\,\mu_{\pi,j}(p)| \geq \theta,$$

for some $\theta \in (0, \frac{1}{2})$. Then for $\mathfrak{C}_{RS} := \max_{\pi,\pi'\in\mathcal{S}} \mathfrak{C}_{\pi\times\widetilde{\pi}'}$, one has

$$|\mathcal{S}| \ll_{n,p,\theta} \mathfrak{C}_{RS}^{\frac{1-2\theta}{4\theta}+o(1)}.$$

Proof. Let ℓ be a positive integer and consider the sum

$$\mathcal{S} := \frac{1}{|\mathcal{S}|^2} \sum_{\pi,\pi'\in\mathcal{S}} w_\pi \overline{w}_{\pi'} \lambda_{\pi\times\widetilde{\pi}'}(\ell), \tag{5.13}$$

where w_π are complex numbers of absolute value 1 to be chosen shortly. By Proposition 5.7 (specifically, (5.12)) for $M = \ell$ and the sequence

$$u_m := \mathbb{1}_{m=\ell},$$

we can upper bound

$$\mathcal{S} \ll (\ell \mathfrak{C}_{RS})^{o(1)} \left(\frac{\ell}{|\mathcal{S}|} + \sqrt{\mathfrak{C}_{RS}} \right).$$

Now take $\ell = p^k$, for some $k \geq 1$. By Proposition 5.5, we can lower bound

$$\mathcal{S} \geq \frac{1}{k|\mathcal{S}|^2} \left| \sum_{\pi \in \mathcal{S}} w_\pi \sum_{j=1}^{n} \alpha_{\pi,j}(p)^k \right|^2.$$

Picking w_π to achieve absolute values around the inner sum above, it follows that

$$\frac{1}{\sqrt{k}|\mathcal{S}|} \sum_{\pi \in \mathcal{S}} \left| \sum_{j=1}^{n} \alpha_{\pi,j}(p)^k \right| \ll (p^k \mathfrak{C}_{RS})^{o(1)} \sqrt{\frac{p^k}{|\mathcal{S}|} + \sqrt{\mathfrak{C}_{RS}}}.$$

This holds for any $k \geq 1$. Summing over k from $k_0 + 1$ to $k_0 + n$, for some $k_0 \geq 1$ to be chosen shortly, we obtain

$$\frac{1}{\sqrt{k_0}|\mathcal{S}|} \sum_{\pi \in \mathcal{S}} \sum_{k=k_0+1}^{k_0+n} \left| \sum_{j=1}^{n} \alpha_{\pi,j}(p)^k \right| \ll_p (p^{k_0} \mathfrak{C}_{RS})^{o(1)} \sqrt{\frac{p^{k_0}}{|\mathcal{S}|} + \sqrt{\mathfrak{C}_{RS}}},$$

where we implicitly used that n is fixed. Applying Lemma 5.8 with $M = k_0$, we reach

$$\frac{1}{k_0^{n+1/2}|\mathcal{S}|} \sum_{\pi \in \mathcal{S}} \max_j |\alpha_{\pi,j}(p)|^{k_0} \ll_p (p^{k_0} \mathfrak{C}_{RS})^{o(1)} \sqrt{\frac{p^{k_0}}{|\mathcal{S}|} + \sqrt{\mathfrak{C}_{RS}}}. \tag{5.14}$$

But by our assumption on $\mathcal{S}$ and unitarity ($\{\mu_{\pi,j}\} = \{-\overline{\mu}_{\pi,j}\}$), we have

$$\max_j |\alpha_{\pi,j}(p)|^{k_0} = \max_j p^{k_0 \mathrm{Re}\, \mu_{\pi,j}(p)} \geq p^{k_0 \theta}, \qquad \forall \pi \in \mathcal{S}.$$

Plugging this into (5.14) and squaring, we reach

$$\frac{1}{k_0^{2n+1}} p^{2k_0\theta} \ll_p (p^{k_0} \mathfrak{C}_{RS})^{o(1)} \left(\frac{p^{k_0}}{|\mathcal{S}|} + \sqrt{\mathfrak{C}_{RS}} \right) \tag{5.15}$$

To optimize, we pick k_0 such that $p^{k_0} \asymp_p |\mathcal{S}|\sqrt{\mathfrak{C}_{RS}}$; in particular, the factor of k_0^{2n+1} grows like $(|\mathcal{S}|\mathfrak{C}_{RS})^{o(1)}$. We conclude that

$$(|\mathcal{S}|\sqrt{\mathfrak{C}_{RS}})^{2\theta} \ll_p (|\mathcal{S}|\mathfrak{C}_{RS})^{o(1)} \sqrt{\mathfrak{C}_{RS}},$$

which rearranges to the desired bound. □

In particular, Theorem 5.9 establishes Theorem 5.2 at the finite places (using (2.61) to bound $\mathfrak{C}_{RS}$).

5.4.2 The Archimedean case

Fix $n \geq 2$ and $v = \infty$. Here we modify the argument in Section 5.4.1 to prove density theorems for the local parameters of cuspidal automorphic representations at ∞. Following Luo–Rudnick–Sarnak [LRS95], we access these parameters through the vanishing of $L(s, \pi \times \widetilde{\pi}')$ at $s = \mu_{\pi,j}(\infty) + \overline{\mu}_{\pi,j'}(\infty)$; this was used in Lemma 5.6 and Proposition 5.7, conditions (ii).

Theorem 5.10 (Density at the infinite place)**.** *Let $\mathcal{S} \subset \mathfrak{F}_n$ be a finite set of cuspidal automorphic representations of $\mathrm{GL}_n(\mathbb{A}_{\mathbb{Q}})$ unramified at $v = \infty$, satisfying*

$$\forall \pi \in \mathcal{S}, \qquad \max_j |\mathrm{Re}\, \mu_{\pi,j}(\infty)| \geq \theta,$$

for some $\theta \in (0, \frac{1}{2})$. Then with $\mathfrak{C}_{RS} := \max_{\pi,\pi' \in \mathcal{S}} \mathfrak{C}_{\pi \times \widetilde{\pi}'}$, one has

$$|\mathcal{S}| \ll_{n,\theta} \mathfrak{C}_{RS}^{\frac{1-2\theta}{4\theta}+o(1)}.$$

Proof. For $\pi \in \mathcal{S}$, let

$$\beta_\pi := \max_j \mathrm{Re}\, \mu_{\pi,j}(\infty).$$

In particular, by unitarity we have $\beta_\pi = \max_j |\mathrm{Re}\, \mu_{\pi,j}(\infty)| \geq \theta$. In analogy with (5.13), we consider the sum

$$\begin{aligned} \mathcal{S} &:= \frac{1}{|\mathcal{S}|^2} \sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} \ell^{\beta_\pi + \overline{\beta}_{\pi'}} \\ &= \frac{1}{|\mathcal{S}|^2} \sum_{\pi,\pi' \in \mathcal{S}} w_\pi \overline{w}_{\pi'} \lambda_{\pi \times \widetilde{\pi}'}(1) \left(\frac{\ell}{1}\right)^{\beta_\pi + \overline{\beta}_{\pi'}}, \end{aligned}$$

where ℓ is a positive integer and w_π are 1-bounded weights to be chosen shortly. Using Proposition 5.7 (specifically, (5.12)) for $M = \ell$ and the sequence

$$u_m := \mathbb{1}_{m=1},$$

we can again upper bound

$$\mathcal{S} \ll (\ell \mathfrak{C}_{RS})^{o(1)} \left(\frac{\ell}{|\mathcal{S}|} + \sqrt{\mathfrak{C}_{RS}}\right).$$

On the other hand, picking $w_\pi := |\ell^{\beta_\pi}|/\ell^{\beta_\pi}$, we have the lower bound

$$\ell^{2\theta} \leq \frac{1}{|\mathcal{S}|^2} \sum_{\pi,\pi' \in \mathcal{S}} \left|\ell^{\beta_\pi} \ell^{\overline{\beta}_{\pi'}}\right| = \mathcal{S},$$

Putting these two bounds together gives

$$\ell^{2\theta} \ll (\ell \mathfrak{C}_{RS})^{o(1)} \left(\frac{\ell}{|\mathcal{S}|} + \sqrt{\mathfrak{C}_{RS}}\right).$$

We pick $\ell \asymp |\mathcal{S}|\sqrt{\mathfrak{C}_{RS}}$ to optimize, and conclude that

$$\left(|\mathcal{S}|\sqrt{\mathfrak{C}_{RS}}\right)^{2\theta-o(1)} \ll \mathfrak{C}_{RS}^{\frac{1}{2}+o(1)},$$

which rearranges to the desired bound, as before. □

This also completes the proof of Theorem 5.2, in light of Theorem 5.9 and (2.61).

5.4.3 Remarks on character twists

Recall that one can apply our Theorem 5.2, as a black-box, to the family of twists of a given cuspidal automorphic representation by Dirichlet characters, to almost recover the pointwise bounds of Luo–Rudnick–Sarnak [LRS95]. If one unpacks the proof of Theorem 5.2 for this particular family, the resulting argument resembles that in [LRS95]. A more optimistic goal would be to use twists by Dirichlet characters to obtain a better density theorem, i.e., to enhance the strength of Theorem 5.2 for a general family of cuspidal automorphic representations (say, inequivalent by Hecke character twists). Indeed, such twists have led to the threshold $\theta = \frac{1}{2} - \frac{1}{n^2+1}$ in the pointwise bounds [LRS95; RS96; LRS99], so one could hope for a density theorem with an exponent that vanishes at $\theta = \frac{1}{2} - \frac{1}{n^2+1}$; currently, in the right-hand side of Theorem 5.2, the exponent only vanishes at $\theta = \frac{1}{2}$. Below, we explain why this approach doesn't quite work.

There are two essentially-equivalent ways to enhance our bounds via Dirichlet character twists; unfortunately, their impact is limited when $\theta < \frac{1}{2} - \frac{1}{n^2+1}$. Indeed, let $\mathcal{S}$ be a finite family of unitary cuspidal automorphic representations of $\mathrm{GL}_n(\mathbb{A}_{\mathbb{Q}})$, unramified at a fixed place v, with $\max_j |\mathrm{Re}\,\mu_{\pi,j}(v)| \geq \theta$, no two equivalent by a Hecke character twist. Let $\mathfrak{C}_{RS} := \max_{\pi,\pi'\in\mathcal{S}} \mathfrak{C}_{\pi\times\widetilde{\pi}'}$.

The first approach is to boost the size of $\mathcal{S}$ by considering the family

$$\mathcal{S}' := \{\pi \otimes \chi : \pi \in \mathcal{S}, \chi \in \Xi\},$$

where Ξ contains all the primitive even Dirichlet characters of a well-chosen prime conductor q. This increases the size of $|\mathcal{S}|$ by $q^{1-o(1)}$, and the maximal Rankin–Selberg conductor by q^{n^2} (provided that $q \nmid \mathfrak{q}_\pi$ for all $\pi \in \mathcal{S}$). Applying Theorem 5.2 as a black-box to $\mathcal{S}$ leads to the bound

$$q|\mathcal{S}| \ll_\theta \left(\mathfrak{C}_{RS}\, q^{n^2}\right)^{\frac{1-2\theta}{4\theta}+o(1)}. \tag{5.16}$$

The second approach is to adapt the proofs of Theorems 5.9 and 5.10 by boosting the contribution of the terms $m = \ell$, respectively $m = 1$. More specifically, in these proofs, one can majorize the indicator function $\mathbb{1}_{m=\ell}$ (respectively, $\mathbb{1}_{m=1}$)

by $\mathbb{1}_{m\equiv\pm\ell \pmod q}$ (respectively, $\mathbb{1}_{m\equiv\pm 1 \pmod q}$), then detect the congruences via even[6] Dirichlet characters, and ultimately apply Lemma 5.6 to the twisted Rankin–Selberg L-function $L(s,(\pi\otimes\chi)\times\widetilde{\pi}')$, via (2.64). This would lead to the bound

$$\ell^{2\theta} \ll (\ell q \mathfrak{C}_{RS})^{o(1)} \left(\frac{\ell}{q|\mathcal{S}|} + q^{\frac{n^2}{2}} \sqrt{\mathfrak{C}_{RS}} \right), \tag{5.17}$$

which recovers (5.16) after optimizing in ℓ.

In both approaches, a more careful analysis of the off-diagonal terms (to which Lemma 5.6 applies the functional equation) would extract an additional square-root cancellation over Dirichlet characters mod q, stemming from Deligne's bounds for hyper-Kloosterman sums (as in [LRS95; Kim03]). This would produce the slightly-better bound

$$\ell^{2\theta} \ll (\ell q \mathfrak{C}_{RS})^{o(1)} \left(\frac{\ell}{q|\mathcal{S}|} + \sqrt{\mathfrak{C}_{RS}\, q^{n^2-1}} \right), \tag{5.18}$$

and ultimately

$$q|\mathcal{S}| \ll_\theta \left(\mathfrak{C}_{RS}\, q^{n^2-1} \right)^{\frac{1-2\theta}{4\theta}+o(1)}. \tag{5.19}$$

The issue is that at this point, if $\theta < \frac{1}{2} - \frac{1}{n^2+1}$ (as is always the case in a nontrivial density theorem), the power of q is smaller in the left-hand side of (5.19) than in the right-hand side, so it is optimal to pick $q = 1$. In other words, the gamble of character twists only pays off in the range where it produces a contradiction by letting $q \to \infty$, which already gave the pointwise bounds.

There are, however, two caveats which could make twists by Dirichlet characters useful in the setting of density theorems. On the one hand, when $v = p < \infty$, the dependency of the implied constant in (5.19) on p may be important for applications. This dependency arises mainly through the optimization in $\ell = p^k$ (which only allows for multiplicative jumps of size p), and can be improved when the parameter q is also available for optimization.

On the other hand, it might be possible to run a similar argument with character twists mod q, where q is *not* relatively prime to the arithmetic conductors $\mathfrak{q}_\pi$. Suppose that all $\pi \in \mathcal{S}$ have the same arithmetic conductor $\mathfrak{q}_\pi = \mathfrak{q}_0$, which is a prime, and that the Langlands parameters at $v = \infty$ of all $\pi \in \mathcal{S}$ are $O(1)$; this is the case in Theorem 5.1. Then for a large enough exponent $k \approx 2n$, one would expect $\mathfrak{C}_{(\pi\otimes\chi)\times\widetilde{\pi}'} \approx q^{n^2}$ for all primitive even Dirichlet characters mod $q := \mathfrak{q}_0^k$, as in [Cor19]. The analogues of (5.18) and (5.19) in this context would be

$$\ell^{2\theta} \ll (\ell \mathfrak{q}_0)^{o(1)} \left(\frac{\ell}{q|\mathcal{S}|} + \sqrt{q^{n^2-1}} \right) \qquad \rightsquigarrow \qquad q|\mathcal{S}| \ll_\theta \left(q^{n^2-1} \right)^{\frac{1-2\theta}{4\theta}+o(1)}.$$

[6] Working with *even* Dirichlet characters is only crucial when $v = \infty$.

If successful, this strategy would therefore imply a bound of the shape

$$|\mathcal{S}| \ll_\theta \mathfrak{q}_0^{O_n\left(\frac{1}{2}-\frac{1}{n^2+1}-\theta\right)+o(1)},$$

which would improve Theorem 5.2 for θ close to $\frac{1}{2}-\frac{1}{n^2+1}$, with an exponent of $o(1)$ when $\theta = \frac{1}{2}-\frac{1}{n^2+1}$. Note that this would automatically beat the pointwise threshold of $\frac{1}{2}-\frac{1}{n^2+1}$ for almost all forms of a given prime conductor; the downside of this approach is that it cannot leverage any averaging over the arithmetic conductor (i.e., over the level of the underlying Maass forms), as required in Problem 1.

5.5 Positive semi-definiteness at ramified primes

Here we complete the proof of Proposition 5.5, by showing that $(\log L(s, \pi_p \times \widetilde{\pi}'_p))_{\pi,\pi' \in \mathcal{S}}$ is positive semi-definite for an arbitrary prime p (which may be ramified for some $\pi \in \mathcal{S}$). We assume the setup of Proposition 5.5, fix p, and follow the computations of Rudnick–Sarnak [RS96, §5]; we also point the reader again to the closely-related computations of Brumley in [ST19, Appendix].

As in [RS96, (5.1)], we can write the local component π_p of $\pi \in \mathcal{S}$ as a Langlands quotient, the unique irreducible quotient of the induced representation

$$\mathrm{Ind}\left(\mathrm{GL}_n, P_\pi; (\sigma_{\pi,j}[t_{\pi,j}])_{j=1}^{J_\pi}\right),$$

where P_π is a standard parabolic subgroup of type $(n_{\pi,j})_{j=1}^{J_\pi}$, $\sigma_{\pi,j}$ are unitary *tempered* representations of $\mathrm{GL}_{n_{\pi,j}}$, and $t_{\pi,j} \in \mathbb{R}$ are the Langlands parameters. Here we recall the notation $\sigma[t] := \sigma \otimes |\cdot|^t$; since π is unitary, we also have $\{\sigma_{\pi,j}[t_{\pi,j}]\} = \{\widetilde{\sigma}_{\pi,j}[-t_{\pi,j}]\}$.

Furthermore, as in [RS96, (5.3)], each tempered $\sigma_{\pi,j}$ is given as an induced representation

$$\mathrm{Ind}\left(\mathrm{GL}_{n_{\pi,j}}, P_{\pi,j}; (\tau_{\pi,j,k})_{k=1}^{K_{\pi,j}}\right),$$

where $P_{\pi,j}$ a standard parabolic subgroup of type $(n_{\pi,j,k})_{k=1}^{K_{\pi,j}}$, and $\tau_{\pi,j,k}$ are unitary *square-integrable* representations of $\mathrm{GL}_{n_{\pi,j,k}}$. In turn, as in [RS96, §5.2], each square-integrable $\tau_{\pi,j,k}$ is the unique square-integrable constituent of the induced representation

$$\mathrm{Ind}\left(\mathrm{GL}_{n_{\pi,j,k}}, P_{\pi,j,k}; (\rho_{\pi,j,k}[\ell - \tfrac{L_{\pi,j,k}+1}{2}])_{\ell=1}^{L_{\pi,j,k}}\right),$$

for some $L_{\pi,j,k} \mid n_{\pi,j,k}$, and $P_{\pi,j,k}$ is a standard parabolic subgroup of type $(d, \dots, d)$ with $d = n_{\pi,j,k}/L_{\pi,j,k}$, and $\rho_{\pi,j,k}$ is a unitary *super-cuspidal* representation of $\mathrm{GL}_{d_{\pi,j,k}}$. We can write $\tau_{\pi,j,k} = \Delta(L_{\pi,j,k}, \rho_{\pi,j,k})$, where the contragredient of such a representation $\Delta(L, \rho)$ is $\widetilde{\Delta}(L, \rho) = \Delta(L, \widetilde{\rho})$.

Then as in [RS96, §5.2], the local factor at p of $L(s,\pi)$ splits as a product

$$\begin{aligned} L(s,\pi_p) &= \prod_{j=1}^{J_\pi} L(s+t_{\pi,j},\sigma_{\pi,j}) = \prod_{j=1}^{J_\pi}\prod_{k=1}^{K_\pi} L(s+t_{\pi,j},\tau_{\pi,j,k}) \\ &= \prod_{j=1}^{J_\pi}\prod_{k=1}^{K_\pi} L(s+t_{\pi,j},\Delta(L_{\pi,j,k},\rho_{\pi,j,k})) \\ &= \prod_{j=1}^{J_\pi}\prod_{k=1}^{K_\pi} L\left(s+t_{\pi,j}+\tfrac{L_{\pi,j,k}-1}{2},\rho_{\pi,j,k}\right). \end{aligned} \tag{5.20}$$

For unitary supercuspidal representations, we have

$$L(s,\rho_{\pi,j,k}) = \begin{cases} \left(1-p^{-(s+t)}\right)^{-1} & \text{if } d_{\pi,j,k}=1,\ \rho=|\cdot|^{it},\ t\in\mathbb{R}, \\ 1 & \text{else.} \end{cases}$$

Similarly, the local factor of the Rankin–Selberg L-function of $\pi,\pi'\in\mathcal{S}$ splits as a product

$$\begin{aligned} &L\left(s,\pi_p\times\widetilde{\pi}'_p\right) \\ &= \prod_{j=1}^{J_\pi}\prod_{j'=1}^{J_{\pi'}} L(s+t_{\pi,j}-t_{\pi',j'},\sigma_{\pi,j}\times\widetilde{\sigma}_{\pi',j'}) \\ &= \prod_{j=1}^{J_\pi}\prod_{k=1}^{K_{\pi,j}}\prod_{j'=1}^{J_{\pi'}}\prod_{k'=1}^{K_{\pi',j'}} L\left(s+t_{\pi,j}-t_{\pi',j'},\tau_{\pi,j,k}\times\widetilde{\tau}_{\pi',j',k'}\right) \\ &= \prod_{j=1}^{J_\pi}\prod_{k=1}^{K_{\pi,j}}\prod_{j'=1}^{J_{\pi'}}\prod_{k'=1}^{K_{\pi',j'}} L\left(s+t_{\pi,j}-t_{\pi',j'},\Delta(L_{\pi,j,k},\rho_{\pi,j,k})\times\Delta(L_{\pi',j',k'},\widetilde{\rho}_{\pi',j',k'})\right) \\ &= \prod_{\substack{j\le J_\pi\\ k\le K_{\pi,j}\\ j'\le J_{\pi'}\\ k'\le K_{\pi',j'}}} \prod_{\ell=1}^{\min(L_{\pi,j,k},L_{\pi',j',k'})} L\left(s+t_{\pi,j}-t_{\pi',j'}+\frac{L_{\pi,j,k}+L_{\pi',j',k'}}{2}-\ell,\rho_{\pi,j,k}\times\widetilde{\rho}_{\pi',j',k'}\right), \end{aligned} \tag{5.21}$$

where we expanded the Rankin–Selberg factor $L(s,\Delta(L,\rho)\times\Delta(L',\rho'))$ for square-integrable representations as in [RS96, (5.5)]. For unitary supercuspidal representations, we have

$$L(s,\rho\times\widetilde{\rho}') = \begin{cases} \left(1-p^{-r(s+iu)}\right)^{-1} & \text{if } \rho'\simeq\rho[iu],\ u\in\mathbb{R}, \\ 1 & \text{else,} \end{cases} \tag{5.22}$$

where in the first case, $r=r(\rho)$ is the order of the cyclic group of characters $|\det(\cdot)|^{iu}$ such that $\rho\simeq\rho[iu]$; note that this r only depends on the twist class of ρ among unitary supercuspidals,

$$[\rho] := \{\rho[iu] : u\in\mathbb{R}\},$$

so we may write $r = r_{[\rho]}$.

In particular, the product in (5.21) will only pick up terms where $\rho_{\pi,j,k}$ and $\widetilde{\rho}_{\pi',j',k'}$ are in the same twist class, so it is natural to split it into a (finite) product over all twist classes $[\rho]$ of the unitary supercuspidals that arise in (5.20) for some $\pi \in \mathcal{S}$. This yields

$$L(s, \pi_p \times \widetilde{\pi}'_p) = \prod_{[\rho]} L_{[\rho]}(s, \pi_p \times \widetilde{\pi}'_p), \tag{5.23}$$

where by (5.21) and (5.22),

$$L_{[\rho]}(s, \pi_p \times \widetilde{\pi}'_p) := \prod_{\substack{1 \le j \le J_\pi \\ 1 \le k \le K_{\pi,j} \\ \rho_{\pi,j,k} = \rho[iu_{\pi,j,k}]}} \prod_{\substack{1 \le j' \le J_{\pi'} \\ 1 \le k \le K_{\pi',j'} \\ \rho_{\pi',j',k'} = \rho[iu_{\pi',j',k'}]}} \times \prod_{\ell=1}^{\min(L_{\pi,j,k}, L_{\pi',j',k'})} \left(1 - p^{-r_{[\rho]}\left(s + s_{\pi,j,k} - s_{\pi',j',k'} + \frac{L_{\pi,j,k} + L_{\pi',j',k'}}{2} - \ell\right)}\right)^{-1},$$

for $s_{\pi,j,k} := t_{\pi,j} + iu_{\pi,j,k}$; this corresponds to [RS96, (5.9)]. Note that for each $[\rho]$ and each π, by unitarity, we have $\{(s_{\pi,j,k}, L_{\pi,j,k}) : \rho_{\pi,j,k} \in [\rho]\} = \{(-\overline{s}_{\pi,j,k}, L_{\pi,j,k}) : \rho_{\pi,j,k} \in [\rho]\}$. Thus we can rewrite

$$L_{[\rho]}(s, \pi_p \times \widetilde{\pi}'_p) := \prod_{\substack{1 \le j \le J_\pi \\ 1 \le k \le K_{\pi,j} \\ \rho_{\pi,j,k} \in [\rho]}} \prod_{\substack{1 \le j' \le J_{\pi'} \\ 1 \le k \le K_{\pi',j'} \\ \rho_{\pi',j',k'} \in [\rho]}} \prod_{\ell=1}^{\min(L_{\pi,j,k}, L_{\pi',j',k'})} \left(1 - p^{-r_{[\rho]}\left(s + s_{\pi,j,k} + \overline{s}_{\pi',j',k'} + \frac{L_{\pi,j,k} + L_{\pi',j',k'}}{2} - \ell\right)}\right)^{-1},$$

By (5.23) and Lemma 5.4, it suffices to prove that for each class $[\rho]$, the family $(\log L_{[\rho]}(s, \pi_p \times \widetilde{\pi}'_p))_{\pi,\pi' \in \mathcal{S}}$ is positive semi-definite. In fact, we can further split

$$L_{[\rho]}(s, \pi_p \times \widetilde{\pi}'_p) = \prod_{\ell=1}^{\infty} L_{[\rho],\ell}(s, \pi_p \times \widetilde{\pi}'_p),$$

where

$$L_{[\rho],\ell}(s, \pi_p \times \widetilde{\pi}'_p) := \prod_{\substack{1 \le j \le J_\pi \\ 1 \le k \le K_{\pi,j} \\ \rho_{\pi,j,k} \in [\rho] \\ L_{\pi,j,k} \ge \ell}} \prod_{\substack{1 \le j' \le J_{\pi'} \\ 1 \le k \le K_{\pi',j'} \\ \rho_{\pi',j',k'} \in [\rho] \\ L_{\pi',j',k'} \ge \ell}} \left(1 - p^{-r_{[\rho]}\left(s + s_{\pi,j,k} + \overline{s}_{\pi',j',k'} + \frac{L_{\pi,j,k} + L_{\pi',j',k'}}{2} - \ell\right)}\right)^{-1},$$

and it suffices to prove that $(\log L_{[\rho],\ell}(s, \pi_p \times \widetilde{\pi}'_p))_{\pi,\pi' \in \mathcal{S}}$ is positive semi-definite for each ℓ. One can expand the formal logarithm as

$$\begin{aligned}
&\log L_{[\rho],\ell}(s, \pi_p \times \widetilde{\pi}'_p) \\
&= \sum_{\substack{1 \le j \le J_\pi \\ 1 \le k \le K_{\pi,j} \\ \rho_{\pi,j,k} = [\rho] \\ L_{\pi,j,k} \ge \ell}} \sum_{\substack{1 \le j' \le J_{\pi'} \\ 1 \le k \le K_{\pi',j'} \\ \rho_{\pi',j',k'} \in [\rho] \\ L_{\pi',j',k'} \ge \ell}} -\log\left(1 - p^{-r_{[\rho]}\left(s + s_{\pi,j,k} + \overline{s}_{\pi',j',k'} + \frac{L_{\pi,j,k} + L_{\pi',j',k'}}{2} - \ell\right)}\right) \\
&= \sum_{\substack{1 \le j \le J_\pi \\ 1 \le k \le K_{\pi,j} \\ \rho_{\pi,j,k} \in [\rho] \\ L_{\pi,j,k} \ge \ell}} \sum_{\substack{1 \le j' \le J_{\pi'} \\ 1 \le k \le K_{\pi',j'} \\ \rho_{\pi',j',k'} \in [\rho] \\ L_{\pi',j',k'} \ge \ell}} \sum_{q \ge 1} \frac{1}{q} p^{-qr_{[\rho]}\left(s + s_{\pi,j,k} + \overline{s}_{\pi',j',k'} + \frac{L_{\pi,j,k} + L_{\pi',j',k'}}{2} - \ell\right)},
\end{aligned}$$

which can be rearranged to

$$\sum_{q \ge 1} \frac{1}{q} p^{-qr_{[\rho]}(s-\ell)} \sum_{\substack{1 \le j \le J_\pi \\ 1 \le k \le K_{\pi,j} \\ \rho_{\pi,j,k} \in [\rho] \\ L_{\pi,j,k} \ge \ell}} p^{-qr_{[\rho]}\left(s_{\pi,j,k} + \frac{L_{\pi,j,k}}{2}\right)} \overline{\sum_{\substack{1 \le j' \le J_{\pi'} \\ 1 \le k \le K_{\pi',j'} \\ \rho_{\pi',j',k'} \in [\rho] \\ L_{\pi',j',k'} \ge \ell}} p^{-qr_{[\rho]}\left(s_{\pi',j',k'} + \frac{L_{\pi',j',k'}}{2}\right)}}.$$

But this is a positive linear combination of terms of the shape $w_{k,\pi}\overline{w}_{k,\pi'}\, p^{-ks}$, and thus positive semi-definite. This completes our proof.

Chapter 6

Unconditional large sieve and zero density estimates for GL_η

(based on joint work with Jesse Thorner)

6.1 Introduction

Let $n \geq 1$, F be a number field, and $\mathfrak{F}_n$ consist of all cuspidal automorphic representations of $\mathrm{GL}_n(\mathbb{A}_F)$, suitably normalized as in Section 2.3. For $\pi \in \mathfrak{F}_n$, we recall the standard automorphic L-function [Bum97; GH11a] given in $\operatorname{Re} s > 1$ by

$$L(s, \pi) = \sum_{\mathfrak{n}} \frac{\lambda_\pi(\mathfrak{n})}{\mathrm{N}\mathfrak{n}^s},$$

where $\mathfrak{n}$ varies over the nonzero integral ideals of $\mathcal{O}_F$. This has an associated arithmetic conductor $\mathfrak{q}_\pi$ as well as an analytic conductor C_π, following Iwaniec and Sarnak [IS00]; our normalization of the central characters ensures that the *universal family*

$$\mathfrak{F}_n(Q) = \{\pi \in \mathfrak{F}_n : C_\pi \leq Q\}$$

is finite. In fact, we expect that $|\mathfrak{F}_n(Q)| \asymp_{n,F} Q^{n+1}$, and a sharp lower bound follows from work of Brumley and Milićević [BM24]:

$$|\mathfrak{F}_n(Q)| \gg_{n,F} Q^{n+1}. \tag{6.1}$$

Now given $N, Q \geq 1$, a finite set $\mathcal{S} \subset \mathfrak{F}_n(Q)$, we denote[1]

$$\begin{aligned} \mathcal{L}(N, \mathcal{S}) &:= \max_{\|a\|_2=1} \sum_{\pi \in \mathcal{S}} \Big| \sum_{\mathrm{N}\mathfrak{n} \leq N} \lambda_\pi(\mathfrak{n}) a(\mathfrak{n}) \Big|^2, \\ \mathcal{L}^{\mathrm{unram}}(N, \mathcal{S}) &:= \max_{\|a\|_2=1} \sum_{\pi \in \mathcal{S}} \Big| \sum_{\substack{\mathrm{N}\mathfrak{n} \leq N \\ \gcd(\mathfrak{n}, \mathfrak{q}_\pi) = \mathcal{O}_F}} \lambda_\pi(\mathfrak{n}) a(\mathfrak{n}) \Big|^2, \end{aligned} \tag{6.2}$$

where the maximum ranges over all complex-valued sequences $(a(\mathfrak{n}))_{\mathrm{N}\mathfrak{n} \leq N}$ normalized so that $\|a\|_2^2 = \sum_{\mathrm{N}\mathfrak{n} \leq N} |a(\mathfrak{n})|^2 = 1$.

The Cauchy–Schwarz inequality yields $\mathcal{L}(N, \mathcal{S}) \ll_F N|\mathcal{S}|$, and a large sieve inequality is any improvement over this, possibly when N is large with respect to $|\mathcal{S}|$ or vice versa[2]. The best possible bound is

$$\mathcal{L}(N, \mathcal{S}) \ll_{n,F} (NQ)^{o(1)}(N + |\mathcal{S}|), \tag{6.3}$$

which can be viewed as a "quasi-orthogonality" result for the $\pi \in \mathcal{S}$. There are some choices of $\mathcal{S}$ for which (6.3) is provably not attainable [DR24; IL07].

When $F = \mathbb{Q}$, each $\pi \in \mathfrak{F}_1(Q)$ corresponds with a primitive Dirichlet character $\chi \pmod{q_\chi}$, and $C_\pi = 3q_\chi$. The classical large sieve inequality for Dirichlet characters, mentioned in (1.11), is the optimal bound

$$\mathcal{L}(N, \mathfrak{F}_1(Q)) \ll N + |\mathfrak{F}_1(Q)|. \tag{6.4}$$

The bound (6.4) frequently makes decisive appearances in analytic number theory. For example, it is crucial in the proof of the Bombieri–Vinogradov theorem, zero density estimates, and bounds for moments of Dirichlet L-functions.

Assuming the generalized Ramanujan conjecture (GRC) from Conjecture 2.13 for all $\pi \in \mathcal{S}$, the first bound on $\mathcal{L}(N, \mathcal{S})$ that holds for all $\mathcal{S}$ is implicit in the work of Duke and Kowalski [DK00, Section 4]. It follows from their work that

$$\mathcal{L}(N, \mathcal{S}) \ll_{n,[F:\mathbb{Q}]} (NQ)^{o(1)}(N + Q^n|\mathcal{S}|^2).$$

Let $0 \leq \theta_n \leq \frac{1}{2} - \frac{1}{n^2+1}$ be the best exponent towards GRC that holds for all $\pi \in \mathcal{S}$. Thorner and Zaman [TZ21] unconditionally proved that[3]

$$\mathcal{L}^{\mathrm{unram}}(N, \mathcal{S}) \ll_{n,[F:\mathbb{Q}]} (NQ)^{o(1)}(N + Q^{4\theta_n n^2 + n}|\mathcal{S}|). \tag{6.5}$$

[1] Given an ordering $(\pi_1, \pi_2, \ldots, \pi_j, \ldots)$ of $\mathcal{S}$ with monotonically increasing analytic conductor and $(\mathfrak{n}_1, \mathfrak{n}_2, \ldots, \mathfrak{n}_i, \ldots)$ of the nonzero ideals of $\mathcal{O}_F$ with monotonically increasing norm in $[1, N]$, let A be the matrix $[\lambda_{\pi_j}(\mathfrak{n}_i)]$. Then $\mathcal{L}(N, \mathcal{S})$ equals the largest eigenvalue of the self-adjoint matrix $\overline{A}^t A$.

[2] The methods in this chapter perform well when N is large with respect to $|\mathcal{S}|$ and Q.

[3] The result stated here incorporates a small correction involving the contribution from the $\mathfrak{n}$ such that $\gcd(\mathfrak{n}, \mathfrak{q}_\pi) \neq \mathcal{O}_F$ and an extension to subsets $\mathcal{S}$, as outlined in [HT24, Section 4].

The exponent $4\theta_n n^2$ arises from addressing the $\mathfrak{n}$ such that $\gcd(\mathfrak{n}, \mathfrak{q}_\pi) \neq \mathcal{O}_F$. In particular, if each $\pi \in \mathcal{S}$ has trivial conductor, then the $4\theta_n n^2$ can be replaced with zero; similarly, if GRC were true, then the factor of $Q^{4\theta_n n^2}$ in (6.5) would be eliminated. In the absence of subconvexity bounds for the the family $\{L(s, \pi \times \widetilde{\pi}') \colon \pi, \pi' \in \mathcal{S}\}$ or a suitable trace formula for $\mathcal{S}$, this seems to be the limit of the current methods.

We build on the ideas from Chapter 5 to develop a new approach to large sieve inequalities that handles the contribution from ramified prime ideals more efficiently, fully eliminating the dependency on GRC. Here is the principal case of our main result, Theorem 6.11.

Theorem 6.1. *If $n, N, Q \geq 1$ and $\mathcal{S} \subset \mathfrak{F}_n(Q)$, then*

$$\mathcal{L}(N, \mathcal{S}) \ll_{n,[F:\mathbb{Q}]} (NQ)^{o(1)}(N + Q^n|\mathcal{S}|).$$

In particular, by (6.1), we have

$$\mathcal{L}(N, \mathfrak{F}_n(Q)) \ll_{n,F} (NQ)^{o(1)}(N + |\mathfrak{F}_n(Q)|^{2-\frac{1}{n+1}}).$$

Theorem 6.1 improves upon (6.5) in two ways. Firstly, it eliminates the need to avoid the $\mathfrak{n}$ that are not coprime to the ramification. Secondly, the removal of the factor $Q^{4\theta_n n^2}$ is significant when θ is bounded away from zero. Note that if $\mathcal{S} = \mathfrak{F}_n(Q)$, then $Q^{4\theta_n n^2}$ is quite large relative to the expected order of magnitude of $|\mathcal{S}|$ from (6.1).

As a corollary of our large sieve inequality, we extend the main result of Chapter 5, on density theorems towards GRC, to include representations that may be *ramified* at the non-Archimedean place $\mathfrak{p}$ considered. This also extends our previous results to number fields.

Corollary 6.2. *For $n, Q \geq 1$, $\theta \in (0, \frac{1}{2})$, and any prime ideal $\mathfrak{p} \subset \mathcal{O}_F$, one has*

$$\#\left\{\pi \in \mathfrak{F}_n(Q) : \max_j |\alpha_{\pi,j}(\mathfrak{p})| \geq \mathrm{N}\mathfrak{p}^\theta\right\} \ll_{\mathfrak{p},\theta,n,[F:\mathbb{Q}]} Q^{n\frac{1-2\theta}{2\theta}+o(1)}.$$

Using (6.1), the upper bound is at most $|\mathfrak{F}_n(Q)|^{\frac{n}{n+1}\cdot\frac{1-2\theta}{2\theta}+o(1)}$, which is nontrivial when

$$\theta > \frac{n}{4n+2} = \frac{1}{4} - \frac{1}{8n+4}.$$

Remark. As in Chapter 5, Corollary 6.2 also wins over the bound attainable with trace formulae [Blo23] when $\theta > \frac{1}{4} + o_{n\to\infty}(1)$.

Finally, our large sieve inequality also leads to zero density estimates for automorphic L-functions. Given $T \geq 0$ and $\sigma \geq 0$, we define[4]

$$N_\pi(\sigma, T) := \sum_{\substack{\rho=\beta+i\gamma \\ \beta\geq\sigma,\ |\gamma|\leq T \\ L(\rho,\pi)=0}} 1.$$

[4]In this work, all zeros are counted with multiplicity.

The generalized Riemann hypothesis (GRH) for $L(s,\pi)$ is equivalent to the statement that if $\sigma > \frac{1}{2}$, then $N_\pi(\sigma, T) = 0$. In the absence of strong zero-free regions for L-functions, strong bounds for $N_\pi(\sigma, T)$, for individual L-functions or in families, can sometimes serve as a substitute for GRH.

The methods in [HT24] combined with a small refinement of (6.5) yield the bound

$$\sum_{\pi\in\mathcal{S}} N_\pi(\sigma,T) \ll_{n,[F:\mathbb{Q}]} (Q^{2\theta_n n^2+\frac{2n+1}{4}} T^{\frac{[F:\mathbb{Q}]n(n+1)}{4}+1}|\mathcal{S}|)^{\frac{4(1-\sigma)}{3-2\sigma}+o(1)}. \tag{6.6}$$

Using Theorem 6.11, we obtain the following substantial improvement over (6.6).

Corollary 6.3. *If $n, Q \geq 1$ and $\mathcal{S} \subset \mathfrak{F}_n(Q)$, then*

$$\sum_{\pi\in\mathcal{S}} N_\pi(\sigma,T) \ll_{n,[F:\mathbb{Q}]} (Q^{\frac{2n+1}{4}} T^{\frac{[F:\mathbb{Q}]n(n+1)}{4}+1}|\mathcal{S}|)^{\frac{4(1-\sigma)}{3-2\sigma}+o(1)}.$$

In light of the sharpness of (6.1), we separately record the consequence

$$\sum_{\pi\in\mathfrak{F}_n(Q)} N_\pi(\sigma,T) \ll_{n,F} (|\mathfrak{F}_n(Q)|^{\frac{3}{2}-\frac{1}{4(n+1)}} T^{\frac{[F:\mathbb{Q}]n(n+1)}{2}+1})^{\frac{4(1-\sigma)}{3-2\sigma}+o(1)}.$$

Throughout the rest of this chapter, we view n and $[F:\mathbb{Q}]$ as fixed, so we will allow implied constants to depend on them.

6.2 Outline

Here we summarize our argument, ignoring various technical details, and emphasizing our key innovations. Let $n \in \mathbb{Z}_+$ be fixed, $N, Q \geq 1$, and $\mathcal{S} \subset \mathfrak{F}_n(Q)$. By the duality principle, the large sieve inequality for $\mathcal{L}(N,\mathcal{S})$ in Theorem 6.1 is equivalent to the bound

$$\sup_{\|w\|_2=1} \sum_{\mathrm{N}\mathfrak{n}\leq N} \Big|\sum_{\pi\in\mathcal{S}} w(\pi)\lambda_\pi(\mathfrak{n})\Big|^2 \ll (NQ)^{o(1)}(N+Q^n|\mathcal{S}|). \tag{6.7}$$

Duke and Kowalski [DK00, Section 4] expanded the square in (6.7), swapped the order of summation, and applied Mellin inversion, thus obtaining

$$\begin{aligned}\sum_{\mathrm{N}\mathfrak{n}\leq N} \Big|\sum_{\pi\in\mathcal{S}} w(\pi)\lambda_\pi(\mathfrak{n})\Big|^2 &= \sum_{\pi,\pi'\in\mathcal{S}} w(\pi)\overline{w_{\pi'}} \sum_{\mathrm{N}\mathfrak{n}\leq N} \lambda_\pi(\mathfrak{n})\overline{\lambda_{\pi'}(\mathfrak{n})} \\ &= \sum_{\pi,\pi'\in\mathcal{S}} w(\pi)\overline{w(\pi')} \frac{1}{2\pi i}\int_{3-i\infty}^{3+i\infty} \Big(\sum_{\mathfrak{n}} \frac{\lambda_\pi(\mathfrak{n})\overline{\lambda_{\pi'}(\mathfrak{n})}}{\mathrm{N}\mathfrak{n}^s}\Big)\frac{N^s}{s}ds.\end{aligned} \tag{6.8}$$

It is implicit in their work that if θ is the best bound towards GRC for each $\pi \in \mathcal{S}$, then there exists a constant $c_n > 0$ a and function $H(s,\pi,\pi') \ll_\epsilon C_\pi^\epsilon C_{\pi'}^\epsilon (\operatorname{Re} s - (\frac{1}{2}+2\theta))^{-c_n}$, holomorphic in the region $\operatorname{Re} s > \frac{1}{2} + 2\theta$, such that

$$\sum_{\mathfrak{n}} \frac{\lambda_\pi(\mathfrak{n})\overline{\lambda_{\pi'}(\mathfrak{n})}}{\mathrm{N}\mathfrak{n}^s} = H(s,\pi,\pi')L(s,\pi\times\widetilde{\pi}').$$

The basis for the results in [DK00] is that this factorization holds inside of the critical strip when $\theta < 1/4$ (and in particular, when GRC holds). Now, one pushes the contour as far to the left as progress towards GRC permits, bounding $L(s,\pi\times\widetilde{\pi}')$ in the critical strip using Lemma 2.14.

It follows from the work of Thorner and Zaman [TZ21, Section 4.2], which relies on the orthogonality of Schur polynomials evaluated at the local roots $\{\alpha_{\pi,j}(\mathfrak{p})\}$ and $\{\alpha_{\pi',j'}(\mathfrak{p})\}$, that the bound

$$\Big|\sum_{\pi\in\mathcal{S}} w(\pi)\lambda_\pi(\mathfrak{n})\Big|^2 \le \sum_{\pi,\pi'\in\mathcal{S}} w(\pi)\overline{w}(\pi)\lambda_{\pi\times\widetilde{\pi}'}(\mathfrak{n}), \qquad (\mathfrak{n},\mathfrak{q}_\pi\mathfrak{q}_{\pi'}) = \mathcal{O}_F \tag{6.9}$$

holds without recourse to unproven progress towards GRC. Summing over all $\mathfrak{n}$ with $\mathrm{N}\mathfrak{n} \le N$ and applying Mellin inversion, we arrive at

$$\begin{aligned}&\sum_{\mathrm{N}\mathfrak{n}\le N}\Big|\sum_{\pi\in\mathcal{S}} w(\pi)\lambda_\pi(\mathfrak{n})\mathbf{1}_{(\mathfrak{n},\mathfrak{q}_\pi)=\mathcal{O}_F}\Big|^2\\ &\le \sum_{\pi,\pi'\in\mathcal{S}} w(\pi)\overline{w}(\pi)\frac{1}{2\pi i}\int_{3-i\infty}^{3+i\infty}\frac{L(s,\pi\times\widetilde{\pi}')}{\prod_{\mathfrak{p}\mid\mathfrak{q}_\pi\mathfrak{q}_{\pi'}}L(s,\pi_\mathfrak{p}\times\widetilde{\pi}'_\mathfrak{p})}\frac{N^s}{s}ds.\end{aligned} \tag{6.10}$$

One then pushes the contour far to the left, bounding $L(s,\pi\times\widetilde{\pi}')$ using Lemma 2.14 and the product over $\mathfrak{p}\nmid\mathfrak{q}_\pi\mathfrak{q}_{\pi'}$ using the existing progress towards GRC in (2.44). It is this latter contribution that leads to the factor of $Q^{4\theta_n n^2}$ in (6.5).

One of our key observations is that (6.9) actually holds with no conditions on $\mathfrak{n}$:

$$\Big|\sum_{\pi\in\mathcal{S}} w(\pi)\lambda_\pi(\mathfrak{n})\Big|^2 \le \sum_{\pi,\pi'\in\mathcal{S}} w(\pi)\overline{w}(\pi')\lambda_{\pi\times\widetilde{\pi}'}(\mathfrak{n}). \tag{6.11}$$

To prove (6.11) in general, we need a more careful treatment of Rankin–Selberg coefficients at ramified places. While an explicit combinatorial argument as in [TZ21] becomes rather cumbersome and difficult to see, a linear-algebraic formulation inspired by our Chapter 5 will lead to a fairly short proof. In fact, a reformulation of (6.11) is the fact that for any $\mathfrak{n}$,

$$\lambda_{\pi\times\widetilde{\pi}'}(\mathfrak{n}) - \lambda_\pi(\mathfrak{n})\overline{\lambda}_{\pi'}(\mathfrak{n}) \qquad \text{form a } \textit{positive semi-definite matrix} \text{ in } \mathbb{C}^{\mathcal{S}\times\mathcal{S}}. \tag{6.12}$$

In Proposition 5.5, we proved that $\lambda_{\pi\times\widetilde{\pi}'}(\mathfrak{n})$ form a positive semi-definite matrix (in the case $F = \mathbb{Q}$), which is a priori a weaker statement. However, applying this to

the family $\mathcal{S}' := \mathcal{S} \cup \{\mathbb{1}\}$, where $\mathbb{1}$ denotes the trivial representation, we see that the $(|\mathcal{S}|+1) \times (|\mathcal{S}|+1)$ block matrix

$$A' := \begin{pmatrix} A & \vec{v} \\ \vec{v}^* & 1 \end{pmatrix}, \qquad \text{where} \qquad A_{\pi,\pi'} := \lambda_{\pi \times \widetilde{\pi}'}(\mathfrak{n}), \qquad v_\pi := \lambda_\pi(\mathfrak{n}),$$

is positive semi-definite. It is a nice exercise to deduce that the matrix $A - \vec{v}\vec{v}^*$ is positive semi-definite (either by matrix algebra or by spectral decomposition and Cauchy–Schwarz), which is precisely the content of (6.12).

A key further step, inspired by [HT24], is to extend our arguments to include Dirichlet coefficients $\lambda^\circ_{\pi \times \widetilde{\pi}'}(\mathfrak{n})$ of other related Dirichlet series, such as $L(s, \pi \times \widetilde{\pi}')^{-1}$ or $\log L(s, \pi \times \widetilde{\pi}')$. Although these coefficients do not form a positive semi-definite matrix, they turn out to have $\lambda_{\pi \times \widetilde{\pi}'}$ as a "positive semi-definite cover", in a sense that we introduce in Section 6.3. This leads to a general inequality of the form

$$\Big| \sum_{\pi \in \mathcal{S}} w(\pi) \lambda^\circ_\pi(\mathfrak{n}) \Big|^2 \leq \sum_{\pi, \pi' \in \mathcal{S}} w(\pi) \overline{w}(\pi') \lambda_{\pi \times \widetilde{\pi}'}(\mathfrak{n}). \tag{6.13}$$

The resulting large sieve inequality for $\lambda^\circ_\pi(\mathfrak{n})$ is given in Theorem 6.11. Our proof of Corollary 6.3 relies on inserting this result, in the specific case where $\lambda^\circ_\pi(\mathfrak{n})$ equals the $\mathfrak{n}$-th Dirichlet coefficient $\mu_\pi(\mathfrak{n})$ of $L(s, \pi)^{-1}$, into the zero detection arguments of Humphries and Thorner [HT24, Section 5].

6.2.1 Structure

For the rest of the chapter, the reader should be familiar with most of the notation and preliminaries from Section 2.3.

In Section 6.3, we introduce the notion of positive semi-definite covers for L-functions, which significantly develops the linear-algebraic formalism from Section 5.3.1. In Section 6.4, we prove our main result generalizing Theorem 6.1, and deduce Corollary 6.2. Finally, in Section 6.5, we establish the zero density estimate in Corollary 6.3.

6.3 Positive semi-definite covers

In this section, we significantly extend the ideas from Chapter 5 on positive semi-definite families of L-functions. Our goal is to prove the inequality (6.13) for Rankin–Selberg coefficients, reiterated in Proposition 6.10 – recall that such a result is implicit in the proof of [HT24, Theorem 4.1] when $\mathfrak{n}$ has no ramified prime factors. In our approach, we treat ramified and unramified places on the same footing; to this end, we use a fairly general linear-algebraic language, although our arguments are ultimately combinatorial.

6.3.1 Covers for matrices

We first introduce a generalization of the classical positive semi-definiteness property for matrices.

Definition 6.4. Let $\mathcal{S}$ be a finite set and $a, a^+ : \mathcal{S} \times \mathcal{S} \to \mathbb{C}$ (which can be viewed as matrices in $\mathbb{C}^{\mathcal{S}\times\mathcal{S}}$). We say that $a^+(x,y)$ is a *positive semi-definite cover* for $a(x,y)$ if and only if there exist countably-many functions $u_j : \mathcal{S} \to \mathbb{C}$ and complex numbers d_j with $|d_j| \le 1$ such that

$$a(x,y) = \sum_{j=1}^{\infty} d_j u_j(x)\overline{u}_j(y), \qquad a^+(x,y) = \sum_{j=1}^{\infty} u_j(x)\overline{u}_j(y), \tag{6.14}$$

for all $x, y \in \mathcal{S}$, where the convergence is absolute.

Remark. In particular, $a^+(x,y)$ is positive semi-definite iff it is a positive semi-definite cover for itself (or for the identically-zero matrix). Any Hermitian matrix has a positive semi-definite cover, obtained by replacing all eigenvalues in its spectral decomposition with their absolute values.

To motivate this definition, consider the following inequality.

Lemma 6.5. *Let $\mathcal{S}$ be a finite set and $a, a^+ : \mathcal{S} \times \mathcal{S} \to \mathbb{C}$, such that $a^+(x,y)$ is a positive semi-definite cover for $a(x,y)$. Then for any $\mathcal{S}_1, \mathcal{S}_2 \subset \mathcal{S}$ and $v : \mathcal{S}_1 \to \mathbb{C}$, $w : \mathcal{S}_2 \to \mathbb{C}$, one has*

$$\Big|\sum_{\substack{x\in\mathcal{S}_1\\ y\in\mathcal{S}_2}} v(x)w(y)a(x,y)\Big|^2 \le \Big(\sum_{x,x'\in\mathcal{S}_1} v(x)\overline{v}(x')a^+(x,x')\Big)\Big(\sum_{y,y'\in\mathcal{S}_2} w(y)\overline{w}(y')a^+(y',y)\Big).$$

Proof. By expanding $a(x,y)$ and $a^+(x,y)$ as in (6.14), swapping sums, and applying Cauchy–Schwarz in j, we obtain

$$\begin{aligned}
&\Big|\sum_{\substack{x\in\mathcal{S}_1\\ y\in\mathcal{S}_2}} v(x)w(y)a(x,y)\Big|^2 \\
&= \Big|\sum_{j=1}^{\infty} d_j \sum_{x\in\mathcal{S}_1} v(x)u_j(x) \sum_{y\in\mathcal{S}_2} w(y)\overline{u}_j(y)\Big|^2 \\
&\le \Big(\sum_{j=1}^{\infty}\sum_{x,x'\in\mathcal{S}_1} v(x)\overline{v}(x')u_j(x)\overline{u}_j(x')\Big)\Big(\sum_{j=1}^{\infty}\sum_{y,y'\in\mathcal{S}_2} w(y)\overline{w}(y')\overline{u}_j(y)u_j(y')\Big) \\
&= \Big(\sum_{x,x'\in\mathcal{S}_1} v(x)\overline{v}(x')\sum_{j=1}^{\infty} u_j(x)\overline{u}_j(x')\Big)\Big(\sum_{y,y'\in\mathcal{S}_2} w(y)\overline{w}(y')\sum_{j=1}^{\infty}\overline{u}_j(y)u_j(y')\Big),
\end{aligned}$$

which simplifies to the desired bound. □

Definition 6.4 is also well-behaved, in the sense that positive semi-definite covers are stable under several operations which also preserve positive semi-definiteness.

Lemma 6.6. *Let $\lambda \in \mathbb{C}$. If $a^+(x,y), b^+(x,y)$ are positive semi-definite covers of $a(x,y), b(x,y)$ respectively, then*

$$|\lambda| a^+(x,y), \qquad (a^+ + b^+)(x,y), \qquad (a^+ \cdot b^+)(x,y)$$

are positive semi-definite covers, respectively, of

$$\lambda a(x,y), \qquad (a+b)(x,y), \qquad (a \cdot b)(x,y).$$

Proof. Let us write $a(x,y) = \sum_{j=1}^{\infty} c_j u_j(x)\overline{u}_j(y)$, $a^+(x,y) = \sum_{j=1}^{\infty} u_j(x)\overline{u}_j(y)$ according to Definition 6.4, and similarly $b(x,y) = \sum_{j=1}^{\infty} d_j v_j(x)\overline{v}_j(y)$, $b^+(x,y) = \sum_{j=1}^{\infty} v_j(x)\overline{v}_j(y)$, where $|c_j|, |d_j| \leq 1$. The claims for scaling and addition are immediate; for multiplication, we expand

$$a(x,y)\, b(x,y) = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} c_j d_k \; (u_j \cdot v_k)(x) \; \overline{(u_j \cdot v_k)(y)},$$

and $a^+(x,y)\, b^+(x,y)$ similarly, then re-index to a single countable sum. □

6.3.2 Covers for families of Dirichlet series

We now pass from matrices $a(x,y)$ to families of Dirichlet series $A(s;x,y)$. The reader should keep in mind the case when $\mathcal{S}$ is a family of automorphic representations, and the families $A(s;x,y)$ come from Rankin–Selberg convolutions.

Definition 6.7. Let F be a number field, and let $\mathfrak{n}$ range over the nonzero ideals of $\mathcal{O}_F$. Let $\mathcal{S}$ be a finite set, and $(A(s;x,y))_{x,y\in\mathcal{S}}$, $(A^+(s;x,y))_{x,y\in\mathcal{S}}$ be families of formal Dirichlet series with expansions

$$A(s;x,y) = \sum_{\mathfrak{n}} \frac{a(\mathfrak{n};x,y)}{\mathrm{N}\mathfrak{n}^s}, \qquad A^+(s;x,y) = \sum_{\mathfrak{n}} \frac{a^+(\mathfrak{n};x,y)}{\mathrm{N}\mathfrak{n}^s}.$$

We say that $A^+(s;x,y)$ is a *positive semi-definite cover* for $A(s;x,y)$ if and only if for each $\mathfrak{n}$, $a^+(\mathfrak{n};x,y)$ is a positive semi-definite cover for $a(\mathfrak{n};x,y)$ (indexing over $x,y \in \mathcal{S}$). In other words, there exist functions $u_j : \mathcal{S} \to \mathbb{C}$, complex numbers d_j with $|d_j| \leq 1$, and integral ideals $\mathfrak{n}_j$ such that

$$A(s;x,y) = \sum_{j=1}^{\infty} d_j \frac{u_j(x)\overline{u}_j(y)}{\mathrm{N}\mathfrak{n}_j^s}, \qquad A^+(s;x,y) = \sum_{j=1}^{\infty} \frac{u_j(x)\overline{u}_j(y)}{\mathrm{N}\mathfrak{n}_j^s},$$

where the convergence of the Dirichlet coefficient of each $\mathrm{N}\mathfrak{n}^{-s}$ is absolute, pointwise in x,y. We say that $A^+(s;x,y)$ is *positive semi-definite* if and only if it is a positive semi-definite cover for 0.

Fortunately, some natural operations on L-functions preserve positive semi-definite covers; the following lemma is a generalization of Lemma 5.4.

Lemma 6.8. *Let $z \in \mathbb{C}$. If $A^+(s;x,y)$, $B^+(s;x,y)$ are positive semi-definite covers of $A(s;x,y)$, $B^+(s;x,y)$ respectively, then the families*

$$|z|A^+(s;x,y), \quad (A^+ + B^+)(s;x,y), \quad (A^+ \cdot B^+)(s;x,y), \quad \exp A^+(s;x,y)$$

are positive semi-definite covers, respectively, of

$$\lambda A(s;x,y), \quad (A+B)(s;x,y), \quad (A \cdot B)(s;x,y), \quad \exp A(s;x,y)$$

Moreover, if $C(s;x,y)$ is positive semi-definite, it has $\exp C(s;x,y)$ as a positive semi-definite cover.

Proof. The claims about the covers of $\lambda A(s;x,y)$, $(A+B)(s;x,y)$, and $(A\cdot B)(s;x,y)$ follow from Lemma 6.6. For the exponentiation property we can expand

$$\exp A(s;x,y) = \sum_{k=0}^{\infty} \frac{1}{k!} A(s;x,y)^k, \qquad \exp A^+(s;x,y) = \sum_{k=0}^{\infty} \frac{1}{k!} A^+(s;x,y)^k$$

and use the previous properties (note that $A^+(s;x,y)^k$ is a positive semi-definite cover for $A(s;x,y)^k$ for each k). For the final claim, expand $\exp C(s;x,y)$ as above and consider the terms with $k=1$. □

Finally, we apply these notions to Rankin–Selberg L-functions.

Lemma 6.9. *If $\mathcal{S}$ is a finite subset of $\bigcup_{n=1}^{\infty} \mathfrak{F}_n$, then the family $(\log L(s, \pi \times \widetilde{\pi}'))_{\pi,\pi' \in \mathcal{S}}$ is positive semi-definite.*

Proof. This is purely formal generalization of Proposition 5.5 to variable ranks n and general number fields. When $F = \mathbb{Q}$, the computations of Brumley [ST19, Appendix A.2, (A.8)] express each factor in the Euler product $L(s, \pi \times \widetilde{\pi}') = \prod_{p \text{ prime}} L(s, \pi_p \times \widetilde{\pi}'_p)$, with minor changes of notation, as

$$\log L(s, \pi_p \times \widetilde{\pi}'_p) = \sum_{[\varrho_\ell] \text{ for } \mathbb{Q}_p} \sum_{f,\nu \geq 1} \frac{1}{f p^{e_\ell f (s-\nu)}} \sum_{\substack{\rho_j(\pi_p) \in [\varrho_\ell] \\ n_j(\pi_p) \geq \nu}} z_j(\pi_p, \ell)^{e_\ell f} \overline{\sum_{\substack{\rho_k(\pi'_p) \in [\varrho_\ell] \\ n_k(\pi'_p) \geq \nu}} z_k(\pi'_p, \ell)^{e_\ell f}},$$

where $[\varrho_\ell]$ ranges over all the twist-equivalence classes of unitary supercuspidal representations of a general linear group over $\mathbb{Q}_p$, e_ℓ is the torsion number of ϱ_ℓ, $\rho_j(\pi_p)$

(resp., $\rho_k(\pi'_p)$) ranges over the supercuspidals in a deconstruction of the p-adic components π_p (resp., π'_p) that are twist-equivalent to ϱ_ℓ, $n_j(\pi_p)$ are positive integers depending only on π_p, j, and $z_j(\pi_p, \ell)$ are complex numbers depending only on π_p, ℓ, j. Summing over primes p and collecting terms, one can ultimately write

$$\log L(s, \pi \times \widetilde{\pi}') = \sum_j \frac{S_j(\pi)\overline{S}_j(\pi')}{n_j^s},$$

where $j = (p, \ell, f, \nu)$ is a tuple of parameters with countably many values (such that n_j only achieves a given value finitely many times), and $S_j(\pi)$ are complex numbers depending only on π, j. For a general number field F, an analogous argument expresses

$$\log L(s, \pi \times \widetilde{\pi}') = \sum_j \frac{S_j(\pi)\overline{S}_j(\pi')}{\mathrm{N}\mathfrak{n}_j^s}, \tag{6.15}$$

where all parameters involved depend implicitly on F. This is precisely the form required by Definition 6.7. □

Putting together our ingredients in this section, we obtain the following result, which may be used as a black-box.

Proposition 6.10. *Let $\mathcal{S}_1, \mathcal{S}_2$ be finite subsets of $\bigcup_{n=1}^{\infty} \mathfrak{F}_n$. With the understanding that families are indexed over $\mathcal{S}_1 \cup \{\widetilde{\pi} : \pi \in \mathcal{S}_2\}$:*

(*i*). *$L(s, \pi \times \widetilde{\pi}')$ is a positive semi-definite cover for itself, $L(s, \pi \times \widetilde{\pi}')^{-1}$, as well as $\log L(s, \pi \times \widetilde{\pi}')$.*

(*ii*). *If $L^+(s, \pi \times \widetilde{\pi}') = \sum_{\mathfrak{n}} \lambda^+_{\pi \times \widetilde{\pi}'}(\mathfrak{n}) \mathrm{N}\mathfrak{n}^{-s}$ is a positive semi-definite cover for $L^\circ(s, \pi \times \widetilde{\pi}') = \sum_{\mathfrak{n}} \lambda^\circ_{\pi \times \widetilde{\pi}'}(\mathfrak{n}) \mathrm{N}\mathfrak{n}^{-s}$, then for all $\mathfrak{n}$ and $v : \mathcal{S}_1 \to \mathbb{C}$, $w : \mathcal{S}_2 \to \mathbb{C}$, one has*

$$\Big| \sum_{\substack{\pi_1 \in \mathcal{S}_1 \\ \pi_2 \in \mathcal{S}_2}} v(\pi_1) w(\pi_2) \lambda^\circ_{\pi_1 \times \pi_2}(\mathfrak{n}) \Big|^2 \leq \sum_{\pi_1, \pi'_1 \in \mathcal{S}_1} v(\pi_1) \overline{v}(\pi'_1) \lambda^+_{\pi_1 \times \widetilde{\pi}'_1}(\mathfrak{n}) \times \sum_{\pi_2, \pi'_2 \in \mathcal{S}_2} w(\pi_2) \overline{w}(\pi'_2) \lambda^+_{\widetilde{\pi}'_2 \times \pi_2}(\mathfrak{n}).$$

In particular, if $\mathcal{S}_1 = \{\pi_1\}$ consists of only one element, this reads

$$\Big| \sum_{\pi_2 \in \mathcal{S}_2} w(\pi_2) \lambda^\circ_{\pi_1 \times \pi_2}(\mathfrak{n}) \Big|^2 \leq \lambda^+_{\pi_1 \times \widetilde{\pi}_1}(\mathfrak{n}) \sum_{\pi_2, \pi'_2 \in \mathcal{S}_2} w(\pi_2) \overline{w}(\pi'_2) \lambda^+_{\widetilde{\pi}'_2 \times \pi_2}(\mathfrak{n}). \tag{6.16}$$

If $\mathcal{S}_2 = \{\pi_2\}$, this further simplifies to

$$|\lambda^\circ_{\pi_1 \times \pi_2}(\mathfrak{n})|^2 \leq \lambda^+_{\pi_1 \times \widetilde{\pi}_1}(\mathfrak{n}) \lambda^+_{\widetilde{\pi}_2 \times \pi_2}(\mathfrak{n}). \tag{6.17}$$

Proof. For the first claim in (i), note that $\log L(s, \pi \times \widetilde{\pi}')$ is a positive semi-definite cover for $-\log L(s, \pi \times \widetilde{\pi}')$, and exponentiate both sides. The second claim in (i) follows directly from the last part of Lemma 6.8. Part (ii) is simply an application of Lemma 6.5 for $\mathcal{S}_1$ and $\{\widetilde{\pi} : \pi \in \mathcal{S}_2\}$. □

Remark. Proposition 6.10 generalizes several results found in literature, with no dependencies of the conductors of the representations involved, and no assumptions about ramification:

(1). Take $L^\circ = L^+ = L$, i.e., $\lambda^\circ_{\pi\times\pi'}(\mathfrak{n}) = \lambda^+_{\pi\times\widetilde{\pi}'}(\mathfrak{n}) = \lambda_{\pi\times\widetilde{\pi}'}(\mathfrak{n})$. Then (6.16) implies our claim from (6.11) by taking π_1 to be the trivial representation; recall that (6.11) extends the bound (6.9) used by Thorner–Zaman [TZ21] in the unramified case. Meanwhile, (6.17) generalizes [TZ21, Corollary 3.2] and a remark from Section 5.3.1.

(2). Take $\lambda^\circ_{\pi\times\widetilde{\pi}'}(\mathfrak{n}) = \lambda^+_{\pi\times\widetilde{\pi}'}(\mathfrak{n}) = \lambda_{\pi\times\widetilde{\pi}'}(\mathfrak{n}) - \lambda_\pi(\mathfrak{n})\overline{\lambda}_{\pi'}(\mathfrak{n})$, which is positive semi-definite in π, π' by (6.12). Then (6.17) becomes the bound

$$|\lambda_{\pi_1\times\pi_2}(\mathfrak{n}) - \lambda_{\pi_1}(\mathfrak{n})\lambda_{\pi_2}(\mathfrak{n})|^2 \leq (\lambda_{\pi_1\times\widetilde{\pi}_1}(\mathfrak{n}) - |\lambda_{\pi_1}(\mathfrak{n})|^2)(\lambda_{\widetilde{\pi}_2\times\pi_2}(\mathfrak{n}) - |\lambda_{\pi_2}(\mathfrak{n})|^2),$$

which removes the real part and the coprimality constraint from another result of Thorner–Zaman [TZ21, Proposition 3.1].

(3). Take $L^\circ = L^{-1}$ and $L^+ = L$, i.e., $\lambda^\circ_{\pi\times\widetilde{\pi}'}(\mathfrak{n})$ are the coefficients of $L(s, \pi \times \widetilde{\pi}')^{-1}$, and $\lambda^+_{\pi\times\widetilde{\pi}'}(\mathfrak{n}) = \lambda_{\pi\times\widetilde{\pi}'}(\mathfrak{n})$. Then (6.16) extends a bound implicitly used by Humphries–Thorner [HT24, Section 4] in the unramified case.

(4). When $L^+ = L^\circ = \log L$, (6.17) recovers [ST19, Proposition A.1]. Moreover, when $L^\circ = \log L$, $L^+ = L$, and $\pi_1 = \pi_2$, (6.17) recovers [ST19, (6.10)]. These were the key inequalities for Rankin–Selberg coefficients used in the work of Soundararajan–Thorner [ST19].

6.4 The large sieve

We now state a general large sieve inequality for the Dirichlet coefficients of various L-functions. This improves on the first bound in [TZ21, Theorem 4.2] by a factor of $Q^{4\theta_n n^2}$; the second bound therein can be improved similarly. Once again, we fix $n \in \mathbb{Z}_+$ and the degree $[F : \mathbb{Q}]$, so all implicit constants may depend on $n, [F : \mathbb{Q}]$.

Theorem 6.11. *Let $\varepsilon > 0$, $x, Q, T \geq 1$, $a(\mathfrak{n})$ be a complex-valued function, and $\mathcal{S} \subset \mathfrak{F}_n$ be such that $\max_{\pi\in\mathcal{S}} C_\pi \leq Q$. Denote by $\lambda^\circ_\pi(\mathfrak{n})$ the Dirichlet coefficients of either $L(s, \pi)$, $L(s, \pi)^{-1}$, or $\log L(s, \pi)$. Then one has*

$$\sum_{\pi\in\mathcal{S}} \Big| \sum_{\mathrm{N}\mathfrak{n}\in(x, e^{1/T}x]} a(\mathfrak{n})\lambda^\circ_\pi(\mathfrak{n}) \Big|^2 \ll_\epsilon Q^\varepsilon \left(x + Q^n T^{\frac{[F:\mathbb{Q}]n^2}{2}+\varepsilon} |\mathcal{S}| \right) \sum_{\mathrm{N}\mathfrak{n}\in(x, e^{1/T}x]} |a(\mathfrak{n})|^2.$$

Proof of Theorem 6.1 assuming Theorem 6.11. Theorem 6.1 follows from (2.38) and a standard dyadic decomposition argument, by taking $\lambda^\circ = \lambda$ and $T = 1$. □

To prove Theorem 6.11, we require a simpler version of [TZ21, Lemma 4.1] which does not include a coprimality constraint, and which is closely related to Lemma 5.6 from Chapter 5. We recall our notation for the Fourier transform from Section 2.1.2.

Lemma 6.12. *Fix a smooth test function ϕ with compact support in $[-2, 2]$, and let $x, T \geq 1$, $\pi, \pi' \in \mathfrak{F}_n$, $C_\pi, C_{\pi'} \leq Q$. Then for all $\varepsilon > 0$, one has*

$$\left| \sum_{\mathfrak{n}} \phi\left(T \log \frac{\mathrm{N}\mathfrak{n}}{x}\right) \lambda_{\pi \times \widetilde{\pi}'}(\mathfrak{n}) - x \frac{\widehat{\phi}(-(2\pi i T)^{-1})}{T} \operatorname*{Res}_{s=1} L(s, \pi \times \widetilde{\pi}') \right| \ll_\epsilon Q^{n+\epsilon} T^{\frac{[F:\mathbb{Q}]n^2}{2}+\varepsilon}.$$

Proof. This calculation is almost identical to the proof of [TZ21, Lemma 4.1], except that we do not need to bound the contribution of the ramified prime ideals $\mathfrak{p} \mid \mathfrak{q}_\pi \mathfrak{q}_{\pi'}$ separately. This removes a factor of $Q^{4\theta_n n^2}$ from our final bound compared to [TZ21]. Note that the subtracted main term vanishes unless $\pi = \pi'$. □

Proof of Theorem 6.11. Our key new input over [TZ21, Theorem 4.2] is Proposition 6.10, which lets us handle the ramified and unramified primes on the same footing. We may assume without loss of generality that $\|a\|_2^2 := \sum_{\mathrm{N}\mathfrak{n} \in (x, e^{1/T}x]} |a(\mathfrak{n})|^2 = 1$. Then by duality, we have

$$\mathcal{L} := \max_{\|a\|_2 = 1} \sum_{\pi \in \mathcal{S}} \left| \sum_{\mathrm{N}\mathfrak{n} \in (x, e^{1/T}x]} a(\mathfrak{n}) \lambda_\pi^\circ(\mathfrak{n}) \right|^2 = \sum_{\mathrm{N}\mathfrak{n} \in (x, e^{1/T}x]} \left| \sum_{\pi \in \mathcal{S}} w(\pi) \lambda_\pi^\circ(\mathfrak{n}) \right|^2,$$

for some function $w : \mathcal{S} \to \mathbb{C}$ with $\|w\|_2^2 := \sum_{\pi \in \mathcal{S}} |w(\pi)|^2 = 1$. Now apply inequality (6.16) for each $\mathfrak{n}$, with $\lambda^+ = \lambda$ (as allowed by Proposition 6.10.(i)) and π_1 being the trivial representation, to obtain

$$\mathcal{L} \leq \sum_{\mathrm{N}\mathfrak{n} \in (x, e^{1/T}x]} \sum_{\pi, \pi' \in \mathcal{S}} w(\pi) \overline{w}(\pi') \lambda_{\pi \times \widetilde{\pi}'}(\mathfrak{n}).$$

Moreover, each summand is nonnegative, so we can insert a smooth majorant for the indicator function of $(x, e^{1/T}x]$. Let $\phi : \mathbb{R} \to [0, 1]$ be an infinitely differentiable, bounded function supported in $[-2, 2]$, which is a pointwise upper bound for the indicator function of $[0, 1]$. We find that

$$\begin{aligned} \mathcal{L} &\leq \sum_{\mathfrak{n}} \phi\Big(T \log \frac{\mathrm{N}\mathfrak{n}}{x}\Big) \sum_{\pi, \pi' \in \mathcal{S}} w(\pi) \overline{w}(\pi') \lambda_{\pi \times \widetilde{\pi}'}(\mathfrak{n}) \\ &= \sum_{\pi, \pi' \in \mathcal{S}} w(\pi) \overline{w}(\pi') \sum_{\mathfrak{n}} \phi\Big(T \log \frac{\mathrm{N}\mathfrak{n}}{x}\Big) \lambda_{\pi \times \widetilde{\pi}'}(\mathfrak{n}). \end{aligned}$$

Using Lemma 6.12 to evaluate the inner sums, we obtain

$$\mathcal{L} \leq x\frac{\widehat{\phi}(-(2\pi iT)^{-1})}{T}\sum_{\pi\in\mathcal{S}}|w(\pi)|^2\operatorname*{Res}_{s=1}L(s,\pi\times\widetilde{\pi}) \\ +O_\varepsilon\Big(Q^{n+\varepsilon}T^{\frac{[F:\mathbb{Q}]n^2}{2}+\varepsilon}\sum_{\pi,\pi'\in\mathcal{S}}|w(\pi)w(\pi')|\Big),$$

where the main term comes from the diagonal terms $\pi=\pi'$. Using $\widehat{\phi}(-(2\pi iT)^{-1})\ll 1$ and $\|w\|_2=1$, we conclude that

$$\mathcal{L} \ll_\epsilon x\max_{\pi\in\mathcal{S}}|\operatorname*{Res}_{s=1}L(s,\pi\times\widetilde{\pi})|+Q^{n+\varepsilon}T^{\frac{[F:\mathbb{Q}]n^2}{2}+\varepsilon}|\mathcal{S}|. \tag{6.18}$$

Finally, using Lemma 2.14, we recover the desired bound of $Q^\varepsilon(x+Q^nT^{\frac{[F:\mathbb{Q}]n^2}{2}+\varepsilon}|\mathcal{S}|)$. □

From Theorem 6.11, we can now quickly deduce Corollary 6.2. Again, the key point is that using our positive semi-definite cover formalism, we can include automorphic representations whose local components at the non-Archimedean place $\mathfrak{p}$ are ramified.

Proof of Corollary 6.2. We apply Theorem 6.11 with

$$\mathcal{S}:=\{\pi\in\mathfrak{F}_n(Q):\max_j|\alpha_{\pi,j}(\mathfrak{p})|\geq \mathrm{N}\mathfrak{p}^\theta\},$$

$T=1$, and $\lambda_\pi^\circ(\mathfrak{n})$ being the Dirichlet coefficients of $\log L(s,\pi)$. This yields

$$\sum_{\pi\in\mathcal{S}}\Big|\sum_{\mathrm{N}\mathfrak{n}\in(x,ex]}a(\mathfrak{n})\lambda_\pi^\circ(\mathfrak{n})\Big|^2\ll Q^{o(1)}\left(x+Q^n|\mathcal{S}|\right)\sum_{\mathrm{N}\mathfrak{n}\in(x,ex]}|a(\mathfrak{n})|^2, \tag{6.19}$$

where in $\operatorname{Re}s>1$,

$$\log L(s,\pi)=\sum_{n=1}^\infty\frac{\lambda_\pi^\circ(\mathfrak{n})}{\mathfrak{n}^s}.$$

Now let $k\geq 1$, and take $a(\mathfrak{n})$ to be supported on the single ideal $\mathfrak{n}=\mathfrak{p}^k$ (so also $x\asymp\mathrm{N}\mathfrak{p}^k$). In particular, by (2.42) we have

$$\lambda_\pi^\circ(\mathfrak{p}^k)=\sum_{j=1}^n\frac{\alpha_{\pi,j}(\mathfrak{p})^k}{k},$$

so (6.19) becomes

$$\frac{1}{k}\sum_{\pi\in\mathcal{S}}\Big|\sum_{j=1}^n\alpha_{\pi,j}(\mathfrak{p})^k\Big|^2\ll Q^{o(1)}\left(\mathrm{N}\mathfrak{p}^k+Q^n|\mathcal{S}|\right),$$

which implies, by Cauchy–Schwarz,

$$\frac{1}{\sqrt{k|\mathcal{S}|}}\sum_{\pi\in\mathcal{S}}\left|\sum_{j=1}^{n}\alpha_{\pi,j}(\mathfrak{p})^k\right| \ll Q^{o(1)}\sqrt{\mathrm{N}\mathfrak{p}^k + Q^n|\mathcal{S}|}.$$

Summing over k in a range $[k_0+1, k_0+n]$ and using Turán's lower bound for power sums, exactly as in Section 5.4.1, we conclude that

$$\frac{\sqrt{|\mathcal{S}|}}{k_0^{n+1/2}}\mathrm{N}\mathfrak{p}^{k_0\theta} \le \frac{1}{k_0^{n+1/2}\sqrt{|\mathcal{S}|}}\sum_{\pi\in\mathcal{S}}\max_j|\alpha_{\pi,j}(\mathfrak{p})|^{k_0} \ll_{\mathfrak{p}} Q^{o(1)}\sqrt{\mathrm{N}\mathfrak{p}^{k_0}+Q^n|\mathcal{S}|},$$

Finally, picking k_0 to balance the right-hand side, we conclude that

$$\sqrt{|\mathcal{S}|}(Q^n|\mathcal{S}|)^{\theta} \ll_{\mathfrak{p}} (Q|\mathcal{S}|)^{o(1)}\sqrt{Q^n|\mathcal{S}|},$$

which rearranges to the desired bound, $|\mathcal{S}| \ll Q^{n\frac{1-2\theta}{2\theta}+o(1)}$. □

6.5 Zero density estimates

When the large sieve inequality in Theorem 6.11 is combined with Plancherel's theorem, we can deduce "hybrid" large sieve inequalities. To state these in our setting of interest, we define the numbers $\mu_\pi(\mathfrak{n})$ by the Dirichlet series identity

$$\sum_{\mathfrak{n}}\frac{\mu_\pi(\mathfrak{n})}{\mathrm{N}\mathfrak{n}^s} = \frac{1}{L(s,\pi)} = \prod_{\mathfrak{p}}\prod_{j=1}^{n}(1-\alpha_{\pi,j}(\mathfrak{p})\mathrm{N}\mathfrak{p}^{-s}), \qquad \mathrm{Re}\, s > 1.$$

Corollary 6.13. *Let $Q, T \ge 1$, and $\epsilon > 0$. If $Y \ge e$ and*

$$X \ge Q^n T^{\frac{[F:\mathbb{Q}]n^2}{2}+1}|\mathcal{S}|(QT)^{\varepsilon}, \quad \log Y \asymp_\epsilon \log X,$$

then

$$\sum_{\pi\in\mathcal{S}}\int_{-T}^{T}\Big|\sum_{\mathrm{N}\mathfrak{n}>X}\frac{\mu_\pi(\mathfrak{n})}{\mathrm{N}\mathfrak{n}^{1+\frac{1}{\log Y}+iv}}\Big|^2 dv \ll_\epsilon Q^\epsilon T^\epsilon \log X,$$
$$\sum_{\pi\in\mathcal{S}}\int_{-T}^{T}\Big|\sum_{\mathrm{N}\mathfrak{n}\le X}\frac{\mu_\pi(\mathfrak{n})}{\mathrm{N}\mathfrak{n}^{\frac{1}{2}+iv}}\Big|^2 dv \ll_\epsilon Q^\epsilon T^\epsilon X\log X.$$

Proof. The proof is the same as that of [HT24, Corollary 4.4], with Theorem 6.11 in the role of [HT24, Theorem 4.1], using $\lambda_\pi^\circ(\mathfrak{n}) = \mu_\pi(\mathfrak{n})$. □

Proof of Corollary 6.3. We proceed along the same lines as in the proof of [HT24, Theorem 1.1], with $n' = 1$, π' given by the trivial representation, and with the

coprimality restrictions (as well as the dependencies on θ_n) completely removed. We summarize as follows. The parameters in the beginning of [HT24, §5] become

$$X = Q^n T^{\frac{[F:\mathbb{Q}]n^2}{2}+1}|\mathcal{S}|(QT)^\epsilon, \qquad Y = \big(Q^{\frac{1}{2}} T^{\frac{[F:\mathbb{Q}]n}{2}+1}|\mathcal{S}|X(QT)^\epsilon\big)^{\frac{1}{3-2\sigma}},$$

$$M_X(s,\pi) = \sum_{\mathrm{N}\mathfrak{n}\le X} \frac{\mu_\pi(\mathfrak{n})}{\mathrm{N}\mathfrak{n}^s}, \qquad LM_X(s,\pi) = L(s,\pi)M_X(s,\pi),$$

in $\operatorname{Re} s > 1$, so that

$$1 - LM_X(s,\pi) = L(s,\pi)\sum_{\mathrm{N}\mathfrak{n}>X} \frac{\mu_\pi(\mathfrak{n})}{\mathrm{N}\mathfrak{n}^s}.$$

Following the ideas Montgomery [Mon69], we note that $LM_X(\rho,\pi) = 0$ when ρ is a zero of $L(s,\pi)$, and use a contour-shifting argument to bound the number of zeros $N_\pi(\sigma,T)$ in terms of an integral of $1 - LM_X(s,\pi)$ near the 1-line, and an integral of $LM_X(s,\pi)$ on the $\frac{1}{2}$-line. As in [HT24, near (5.6)], summing over π and using Cauchy–Schwarz, we conclude that

$$\begin{aligned}\sum_{\pi\in\mathcal{S}} N_\pi(\sigma,T) \ll_\epsilon X^\epsilon \Big[Y^{2(1-\sigma)} \int_{-T}^{T} \sum_{\pi\in\mathcal{S}} \Big| \sum_{\mathrm{N}\mathfrak{n}>X} \frac{\mu_\pi(\mathfrak{n})}{\mathrm{N}\mathfrak{n}^{1+\frac{1}{\log Y}+iv}}\Big|^2 dv \\ + Y^{\frac{1}{2}-\sigma}\Big(\int_{-T}^{T}\sum_{\pi\in\mathcal{S}} |L(\tfrac{1}{2}+iv,\pi)|^2 dv\Big)^{\frac{1}{2}}\Big(\int_{-T}^{T}\sum_{\pi\in\mathcal{S}}\Big|\sum_{\mathrm{N}\mathfrak{n}\le X}\frac{\mu_\pi(\mathfrak{n})}{\mathrm{N}\mathfrak{n}^{\frac{1}{2}+iv}}\Big|^2 dv\Big)^{\frac{1}{2}} + Y^{1-\sigma}\Big].\end{aligned}$$

It should be noted that one of the intermediate steps in [HT24, §5] uses the bound in [HT24, Lemma 3.3]:

$$|\mu_{\pi\times\pi'}(\mathfrak{n})| \le \frac{\lambda_{\pi\times\tilde{\pi}}(\mathfrak{n}) + \lambda_{\pi'\times\tilde{\pi}'}(\mathfrak{n})}{2}, \qquad \gcd(\mathfrak{n}, \mathfrak{q}_\pi\mathfrak{q}_{\pi'}) = \mathcal{O}_F. \tag{6.20}$$

Fortunately, (6.20) holds independently of the coprimality condition per (6.17).

We estimate the two v-integrals involving $\mu_\pi(\mathfrak{n})$ using Corollary 6.13 in place of [HT24, Corollary 4.4]. The remaining integral is estimated trivially using Lemma 2.14. These estimates yield the bound

$$\sum_{\pi\in\mathcal{S}} N_\pi(\sigma,T) \ll_\epsilon Y^{2(1-\sigma)}X^\epsilon,$$

which is equivalent to desired result in Corollary 6.3. □

References

[AB24] Edgar Assing and Valentin Blomer. "The density conjecture for principal congruence subgroups". *Duke Math. J.* 173.7 (2024), pp. 1359–1426.

[ABL21] Edgar Assing, Valentin Blomer, and Junxian Li. "Uniform Titchmarsh divisor problems". *Adv. Math.* 393 (2021), Paper No. 108076, 51.

[BB11] Valentin Blomer and Farrell Brumley. "On the Ramanujan conjecture over number fields". *Ann. of Math. (2)* 174.1 (2011), pp. 581–605.

[BB13] Valentin Blomer and Farrell Brumley. "The role of the Ramanujan conjecture in analytic number theory". *Bull. Amer. Math. Soc. (N.S.)* 50.2 (2013), pp. 267–320.

[BCR17] Sandro Bettin, Vorrapan Chandee, and Maksym Radziwiłł. "The mean square of the product of the Riemann zeta-function with Dirichlet polynomials". *J. Reine Angew. Math.* 729 (2017), pp. 51–79.

[BD20] Régis de la Bretèche and Sary Drappeau. "Niveau de répartition des polynômes quadratiques et crible majorant pour les entiers friables". *J. Eur. Math. Soc.* 22.5 (2020), pp. 1577–1624.

[BFI19] E Bombieri, JB Friedlander, and H Iwaniec. "Some corrections to an old paper". *Preprint, arXiv:1903.01371* (2019).

[BFI86] Enrico Bombieri, John B. Friedlander, and Henryk Iwaniec. "Primes in arithmetic progressions to large moduli". *Acta Math.* 156.3-4 (1986), pp. 203–251.

[BFI87] Enrico Bombieri, John B. Friedlander, and Henryk Iwaniec. "Primes in arithmetic progressions to large moduli. II". *Math. Ann.* 277.3 (1987), pp. 361–393.

[BFI89] Enrico Bombieri, John B. Friedlander, and Henryk Iwaniec. "Primes in arithmetic progressions to large moduli. III". *J. Amer. Math. Soc.* 2.2 (1989), pp. 215–224.

[BH97] C. J. Bushnell and G. Henniart. "An upper bound on conductors for pairs". *J. Number Theory* 65.2 (1997), pp. 183–196.

[Blo23] Valentin Blomer. "Density theorems for GL(n)". *Invent. Math.* 232.2 (2023), pp. 683–711.

[BM24] Farrell Brumley and Djordje Milićević. "Counting cusp forms by analytic conductor". *Ann. Sci. Éc. Norm. Supér. (4)* 57.5 (2024), pp. 1473–1597.

[Bom65] Enrico Bombieri. "On the large sieve". *Mathematika* 12 (1965), pp. 201–225.

[Bru06] Farrell Brumley. "Effective multiplicity one on GL_N and narrow zero-free regions for Rankin-Selberg L-functions". *Amer. J. Math.* 128.6 (2006), pp. 1455–1474.

[BT24] Valentin Blomer and Jesse Thorner. "Zeros of L-functions in families near the critical line". *Preprint, arXiv:2410.17158* (2024).

[BTZ22] Farrell Brumley, Jesse Thorner, and Asif Zaman. "Zeros of Rankin-Selberg L-functions at the edge of the critical strip". *J. Eur. Math. Soc. (JEMS)* 24.5 (2022). With an appendix by Colin J. Bushnell and Guy Henniart, pp. 1471–1541.

[Bum97] Daniel Bump. *Automorphic forms and representations*. Vol. 55. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 1997, pp. xiv+574.

[CG11] Javier Cilleruelo and Moubariz Z. Garaev. "Concentration of points on two and three dimensional modular hyperbolas and applications". *Geom. Funct. Anal.* 21.4 (2011), pp. 892–904.

[Cor19] Andrew Corbett. "An explicit conductor formula for $\mathrm{GL}_n \times \mathrm{GL}_1$". *Rocky Mountain J. Math.* 49.4 (2019), pp. 1093–1110.

[Del71] Pierre Deligne. "Formes modulaires et représentations l-adiques". *Séminaire Bourbaki. Vol. 1968/69: Exposés 347–363*. Vol. 175. Lecture Notes in Math. Springer, Berlin, 1971, Exp. No. 355, 139–172.

[DFI95] 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), pp. 423–441.

[DGS17] Sary Drappeau, Andrew Granville, and Xuancheng Shao. "Smooth-supported multiplicative functions in arithmetic progressions beyond the $x^{1/2}$-barrier". *Mathematika* 63.3 (2017), pp. 895–918.

[DI82a] J.-M. Deshouillers and H. Iwaniec. "On the greatest prime factor of n^2+1". *Ann. Inst. Fourier (Grenoble)* 32.4 (1982), pp. 1–11.

[DI82b] J.-M. Deshouillers and H. Iwaniec. "Power mean values of the Riemann zeta function". *Mathematika* 29.2 (1982), pp. 202–212.

[DI82c] Jean-Marc Deshouillers and Henryk Iwaniec. "Kloosterman sums and Fourier coefficients of cusp forms". *Invent. Math.* 70.2 (1982), pp. 219–288.

[DI84] J.-M. Deshouillers and H. Iwaniec. "Power mean-values for Dirichlet's polynomials and the Riemann zeta-function. II". *Acta Arith.* 43.3 (1984), pp. 305–312.

[DI90] W. Duke and H. Iwaniec. “Estimates for coefficients of L-functions. I”. *Automorphic forms and analytic number theory (Montreal, PQ, 1989)*. Univ. Montréal, Montreal, QC, 1990, pp. 43–47.

[DK00] W. Duke and E. Kowalski. “A problem of Linnik for elliptic curves and mean-value estimates for automorphic representations”. *Invent. Math.* 139.1 (2000). With an appendix by Dinakar Ramakrishnan, pp. 1–39.

[DPR23] Sary Drappeau, Kyle Pratt, and Maksym Radziwiłł. “One-level density estimates for Dirichlet L-functions with extended support”. *Algebra Number Theory* 17.4 (2023), pp. 805–830.

[DR24] Alexander Dunn and Maksym Radziwiłł. “Bias in cubic Gauss sums: Patterson’s conjecture”. *Ann. of Math. (2)* 200.3 (2024), pp. 967–1057.

[Dra15] Sary Drappeau. “Théorèmes de type Fouvry-Iwaniec pour les entiers friables”. *Compos. Math.* 151.5 (2015), pp. 828–862.

[Dra17] Sary Drappeau. “Sums of Kloosterman sums in arithmetic progressions, and the error term in the dispersion method”. *Proc. Lond. Math. Soc. (3)* 114.4 (2017), pp. 684–732.

[EH68] Peter D.T.A. Elliott and Heini Halberstam. “A conjecture in prime number theory”. *Symposia Mathematica, Vol. IV (INDAM, Rome, 1968/69)*. Academic Press, London-New York, 1968, pp. 59–72.

[FI10] John Friedlander and Henryk Iwaniec. *Opera de cribro*. Vol. 57. American Mathematical Society Colloquium Publications. American Mathematical Society, Providence, RI, 2010, pp. xx+527.

[FI80] Étienne Fouvry and Henryk Iwaniec. “On a theorem of Bombieri-Vinogradov type”. *Mathematika* 27.2 (1980), pp. 135–152.

[FI83] Étienne Fouvry and Henryk Iwaniec. “Primes in arithmetic progressions”. *Acta Arith.* 42.2 (1983), pp. 197–218.

[FI85] John B. Friedlander and Henryk Iwaniec. “Incomplete Kloosterman sums and a divisor problem”. *Ann. of Math. (2)* 121.2 (1985). With an appendix by Bryan J. Birch and Enrico Bombieri, pp. 319–350.

[FKM14] Étienne Fouvry, Emmanuel Kowalski, and Philippe Michel. “Algebraic trace functions over the primes”. *Duke Math. J.* 163.9 (2014), pp. 1683–1736.

[FKM15] Étienne Fouvry, Emmanuel Kowalski, and Philippe Michel. “On the exponent of distribution of the ternary divisor function”. *Mathematika* 61.1 (2015), pp. 121–144.

[Fou82] Étienne Fouvry. “Répartition des suites dans les progressions arithmétiques”. *Acta Arith.* 41.4 (1982), pp. 359–382.

[Fou84] Étienne Fouvry. “Autour du théorème de Bombieri-Vinogradov”. *Acta Math.* 152.3-4 (1984), pp. 219–244.

[Fou85] Étienne Fouvry. "Sur le problème des diviseurs de Titchmarsh". *J. Reine Angew. Math.* 357 (1985), pp. 51–76.

[Fou87] Étienne Fouvry. "Autour du théorème de Bombieri-Vinogradov. II". *Ann. Sci. École Norm. Sup. (4)* 20.4 (1987), pp. 617–640.

[Fri95] John B. Friedlander. "Bounds for L-functions". *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*. Birkhäuser, Basel, 1995, pp. 363–373.

[FT91] Étienne Fouvry and Gérald Tenenbaum. "Entiers sans grand facteur premier en progressions arithmetiques". *Proc. London Math. Soc. (3)* 63.3 (1991), pp. 449–494.

[FT96] Étienne Fouvry and Gérald Tenenbaum. "Répartition statistique des entiers sans grand facteur premier dans les progressions arithmétiques". *Proc. London Math. Soc. (3)* 72.3 (1996), pp. 481–514.

[GH11a] Dorian Goldfeld and Joseph Hundley. *Automorphic representations and L-functions for the general linear group. Volume I.* Vol. 129. Cambridge Studies in Advanced Mathematics. With exercises and a preface by Xander Faber. Cambridge University Press, Cambridge, 2011, pp. xx+550.

[GH11b] Dorian Goldfeld and Joseph Hundley. *Automorphic representations and L-functions for the general linear group. Volume II.* Vol. 130. Cambridge Studies in Advanced Mathematics. With exercises and a preface by Xander Faber. Cambridge University Press, Cambridge, 2011, pp. xx+188.

[Gra93a] Andrew Granville. "Integers, without large prime factors, in arithmetic progressions. I". *Acta Math.* 170.2 (1993), pp. 255–273.

[Gra93b] Andrew Granville. "Integers, without large prime factors, in arithmetic progressions. II". *Philos. Trans. Roy. Soc. London Ser. A* 345.1676 (1993), pp. 349–362.

[Har08] Glyn Harman. "On values of $n^2 + 1$ free of large prime factors". *Arch. Math. (Basel)* 90.3 (2008), pp. 239–245.

[Har12] Adam J. Harper. "Bombieri–Vinogradov and Barban–Davenport–Halberstam type theorems for smooth numbers". *Preprint, arXiv:1208.5992* (2012).

[Har24] Glyn Harman. "Two problems on the greatest prime factor of n^2+1". *Acta Arith.* 213.3 (2024), pp. 273–287.

[HB19] Peter Humphries and Farrell Brumley. "Standard zero-free regions for Rankin-Selberg L-functions via sieve theory". *Math. Z.* 292.3-4 (2019), pp. 1105–1122.

[Hil86] Adolf Hildebrand. "On the number of positive integers $\leq x$ and free of prime factors $> y$". *J. Number Theory* 22.3 (1986), pp. 289–307.

[HL23] G. H. Hardy and J. E. Littlewood. "Some problems of 'Partitio numerorum'; III: On the expression of a number as a sum of primes". *Acta Math.* 44.1 (1923), pp. 1–70.

[Hoo67] Christopher Hooley. "On the greatest prime factor of a quadratic polynomial". *Acta Math.* 117 (1967), pp. 281–299.

[HT22] Peter Humphries and Jesse Thorner. "Towards a GL_n variant of the Hoheisel phenomenon". *Trans. Amer. Math. Soc.* 375.3 (2022), pp. 1801–1824.

[HT24] Peter Humphries and Jesse Thorner. "Zeros of Rankin-Selberg L-functions in families". *Compos. Math.* 160.5 (2024), pp. 1041–1072.

[Hum18] Peter Humphries. "Density theorems for exceptional eigenvalues for congruence subgroups". *Algebra Number Theory* 12.7 (2018), pp. 1581–1610.

[IK21] Henryk Iwaniec and Emmanuel Kowalski. *Analytic number theory*. Vol. 53. Providence, RI: American Mathematical Society, 2021.

[IL07] Henryk Iwaniec and Xiaoqing Li. "The orthogonality of Hecke eigenvalues". *Compos. Math.* 143.3 (2007), pp. 541–565.

[IS00] H. Iwaniec and P. Sarnak. "Perspectives on the analytic theory of L-functions". *Geom. Funct. Anal.* Special Volume, Part II (2000). GAFA 2000 (Tel Aviv, 1999), pp. 705–741.

[Iwa02] Henryk Iwaniec. *Spectral methods of automorphic forms*. Second. Vol. 53. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI; Revista Matemática Iberoamericana, Madrid, 2002, pp. xii+220.

[Iwa80] Henryk Iwaniec. "A new form of the error term in the linear sieve". *Acta Arith.* 37 (1980), pp. 307–320.

[Iwa90] Henryk Iwaniec. "Small eigenvalues of Laplacian for $\Gamma_0(N)$". *Acta Arith.* 56.1 (1990), pp. 65–82.

[Iwa97] Henryk Iwaniec. *Topics in classical automorphic forms*. Vol. 17. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 1997, pp. xii+259.

[Jan21] Subhajit Jana. "Applications of analytic newvectors for $\mathrm{GL}(n)$". *Math. Ann.* 380.3-4 (2021), pp. 915–952.

[JPS81] H. Jacquet, I. I. Piatetski-Shapiro, and J. Shalika. "Conducteur des représentations du groupe linéaire". *Math. Ann.* 256.2 (1981), pp. 199–214.

[JPS83] H. Jacquet, I. I. Piatetskii-Shapiro, and J. A. Shalika. "Rankin-Selberg convolutions". *Amer. J. Math.* 105.2 (1983), pp. 367–464.

[JS81] H. Jacquet and J. A. Shalika. "On Euler products and the classification of automorphic representations. I". *Amer. J. Math.* 103.3 (1981), pp. 499–558.

[Ker23] Kerr, Bryce and Shparlinski, Igor E. and Wu, Xiaosheng and Xi, Ping. "Bounds on bilinear forms with Kloosterman sums". *J. Lond. Math. Soc. (2)* 108.2 (2023), pp. 578–621.

[Kim03] Henry H. Kim. "Functoriality for the exterior square of GL_4 and the symmetric fourth of GL_2". *J. Amer. Math. Soc.* 16.1 (2003). With appendix 1 by Dinakar Ramakrishnan and appendix 2 by Kim and Peter Sarnak, pp. 139–183.

[KMS17] Emmanuel Kowalski, Philippe Michel, and Will Sawin. "Bilinear forms with Kloosterman sums and applications". *Ann. of Math. (2)* 186.2 (2017), pp. 413–500.

[KMS20] Emmanuel Kowalski, Philippe Michel, and Will Sawin. "Stratification and averaging for exponential sums: bilinear forms with generalized Kloosterman sums". *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* 21 (2020), pp. 1453–1530.

[Kuz80] Nikolai 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), pp. 334–383, 479.

[Li10] Xiannan Li. "Upper bounds on L-functions at the edge of the critical strip". *Int. Math. Res. Not. IMRN* 4 (2010), pp. 727–755.

[Lic23] Jared Duker Lichtman. "Primes in arithmetic progressions to large moduli, and Goldbach beyond the square-root barrier". *Preprint, arXiv:2309.08522v1* (2023).

[Lic25] Jared Duker Lichtman. "A modification of the linear sieve, and the count of twin primes". *Algebra Number Theory* 19.1 (2025), pp. 1–38.

[Lin63] Ju. V. Linnik. *The dispersion method in binary additive problems*. Translated by S. Schuur. American Mathematical Society, Providence, RI, 1963, pp. x+186.

[LP24] Jared Duker Lichtman and Alexandru Pascadi. "Density theorems for GL_n via Rankin-Selberg L-functions". *Preprint, arXiv:2408.13682* (2024).

[LRS95] Wenzhi Luo, Zeév Rudnick, and Peter Sarnak. "On Selberg's eigenvalue conjecture". *Geom. Funct. Anal.* 5.2 (1995), pp. 387–401.

[LRS99] Wenzhi Luo, Zeév Rudnick, and Peter Sarnak. "On the generalized Ramanujan conjecture for $GL(n)$". *Automorphic forms, automorphic representations, and arithmetic (Fort Worth, TX, 1996)*. Vol. 66. Proc. Sympos. Pure Math. Amer. Math. Soc., Providence, RI, 1999, pp. 301–310.

[May15] James Maynard. "Small gaps between primes". *Ann. of Math. (2)* 181.1 (2015), pp. 383–413.

[May25a] James Maynard. "Primes in Arithmetic Progressions to Large Moduli I: Fixed Residue Classes". *Mem. Amer. Math. Soc.* 306.1542 (2025).

[May25b] James Maynard. "Primes in Arithmetic Progressions to Large Moduli II: Well-Factorable Estimates". *Mem. Amer. Math. Soc.* 306.1543 (2025).

[May25c] James Maynard. "Primes in Arithmetic Progressions to Large Moduli III: Uniform Residue Classes". *Mem. Amer. Math. Soc.* 306.1544 (2025).

[Mer23] Jori Merikoski. "On the largest prime factor of n^2+1". *J. Eur. Math. Soc. (JEMS)* 25.4 (2023), pp. 1253–1284.

[Mon69] H. L. Montgomery. "Zeros of L-functions". *Invent. Math.* 8 (1969), pp. 346–354.

[Mon94] Hugh L. Montgomery. *Ten lectures on the interface between analytic number theory and harmonic analysis*. Vol. 84. CBMS Regional Conference Series in Mathematics. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 1994, pp. xiv+220.

[MS04] W. Müller and B. Speh. "Absolute convergence of the spectral side of the Arthur trace formula for GL_n". *Geom. Funct. Anal.* 14.1 (2004). With an appendix by E. M. Lapid, pp. 58–93.

[MW89] C. Mœglin and J.-L. Waldspurger. "Le spectre résiduel de $\mathrm{GL}(n)$". *Ann. Sci. École Norm. Sup. (4)* 22.4 (1989), pp. 605–674.

[Pas25a] Alexandru Pascadi. "Large sieve inequalities for exceptional Maass forms and the greatest prime factor of n^2+1". *Forum Math. Pi, to appear. Preprint, arXiv:2404.04239* (2025).

[Pas25b] Alexandru Pascadi. "On the exponents of distribution of primes and smooth numbers". *Preprint, arXiv:2505.00653* (2025).

[Pas25c] Alexandru Pascadi. "Smooth numbers in arithmetic progressions to large moduli". *Compos. Math.* 161.8 (2025), pp. 1923–1974.

[Pol14] D. H. J. Polymath. "Variants of the Selberg sieve, and bounded intervals containing many primes". *Res. Math. Sci.* 1 (2014), Art. 12, 83.

[PY23] Ian Petrow and Matthew P. Young. "The fourth moment of Dirichlet L-functions along a coset and the Weyl bound". *Duke Math. J.* 172.10 (2023), pp. 1879–1960.

[RS96] Zeév Rudnick and Peter Sarnak. "Zeros of principal L-functions and random matrix theory". *Duke Math. J.* 81 (1996). A celebration of John F. Nash, Jr., pp. 269–322.

[Sar05] Peter Sarnak. "Notes on the generalized Ramanujan conjectures". *Harmonic analysis, the trace formula, and Shimura varieties*. Vol. 4. Clay Math. Proc. Amer. Math. Soc., Providence, RI, 2005, pp. 659–685.

[Sar91] Peter C. Sarnak. "Diophantine problems and linear groups". *Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990)*. Math. Soc. Japan, Tokyo, 1991, pp. 459–471.

[Sar95] Peter Sarnak. "Selberg's eigenvalue conjecture". *Notices Amer. Math. Soc.* 42.11 (1995), pp. 1272–1277.

[Sel65] Atle Selberg. "On the estimation of Fourier coefficients of modular forms". *Proc. Sympos. Pure Math., Vol. VIII*. Amer. Math. Soc., Providence, RI, 1965, pp. 1–15.

[Ser81] Jean-Pierre Serre. “Letter to J.M. Deshouillers”. A copy can be found in Blomer–Brumley’s survey: The role of the Ramanujan conjecture in analytic number theory (Bull. Amer. Math. Soc., 2013). 1981.

[Ser98] Jean-Pierre Serre. *Abelian l-adic representations and elliptic curves*. Vol. 7. Research Notes in Mathematics. With the collaboration of Willem Kuyk and John Labute, Revised reprint of the 1968 original. A K Peters, Ltd., Wellesley, MA, 1998, p. 199.

[Sha81] Freydoon Shahidi. “On certain L-functions”. *Amer. J. Math.* 103.2 (1981), pp. 297–355.

[Shp18] Igor E. Shparlinski. “Character sums with smooth numbers”. *Arch. Math. (Basel)* 110.5 (2018), pp. 467–476.

[ST19] Kannan Soundararajan and Jesse Thorner. “Weak subconvexity without a Ramanujan hypothesis”. *Duke Math. J.* 168.7 (2019). With an appendix by Farrell Brumley, pp. 1231–1268.

[Sta25] Julia Stadlmann. “On primes in arithmetic progressions and bounded gaps between many primes”. *Adv. Math.* 468 (2025), Paper No. 110190.

[SX91] Peter Sarnak and Xiao Xi Xue. “Bounds for multiplicities of automorphic representations”. *Duke Math. J.* 64.1 (1991), pp. 207–227.

[SZ16] Igor E. Shparlinski and Tianping Zhang. “Cancellations amongst Kloosterman sums”. *Acta Arith.* 176.3 (2016), pp. 201–210.

[Ter88] Audrey Terras. *Harmonic analysis on symmetric spaces and applications. II*. Springer-Verlag, Berlin, 1988, pp. xii+385.

[Top17] Berke Topacogullari. “On a certain additive divisor problem”. *Acta Arith.* 181.2 (2017), pp. 143–172.

[Top18] Berke Topacogullari. “The shifted convolution of generalized divisor functions”. *Int. Math. Res. Not. IMRN* 24 (2018), pp. 7681–7724.

[TZ21] Jesse Thorner and Asif Zaman. “An unconditional GL_n large sieve”. *Adv. Math.* 378 (2021), Paper No. 107529, 24.

[Vin65] A. I. Vinogradov. “The density hypothesis for Dirichet L-series”. *Izv. Akad. Nauk SSSR Ser. Mat.* 29 (1965), pp. 903–934.

[Wat95] N. Watt. “Kloosterman sums and a mean value for Dirichlet polynomials”. *J. Number Theory* 53.1 (1995), pp. 179–210.

[Wol73a] Dieter Wolke. “Über die mittlere Verteilung der Werte zahlentheoretischer Funktionen auf Restklassen. I”. *Math. Ann.* 202 (1973), pp. 1–25.

[Wol73b] Dieter Wolke. “Über die mittlere Verteilung der Werte zahlentheoretischer Funktionen auf Restklassen. II”. *Math. Ann.* 204 (1973), pp. 145–153.

[Wu23] Xiaosheng Wu. “The fourth moment of Dirichlet L-functions at the central value”. *Math. Ann.* 387.3-4 (2023), pp. 1199–1248.

[Xi18] Ping Xi. “Ternary divisor functions in arithmetic progressions to smooth moduli”. *Mathematika* 64.3 (2018), pp. 701–729.

[You11] Matthew P. Young. “The fourth moment of Dirichlet L-functions”. *Ann. of Math. (2)* 173.1 (2011), pp. 1–50.

[Zha14] Yitang Zhang. “Bounded gaps between primes”. *Ann. of Math. (2)* 179.3 (2014), pp. 1121–1174.