AN IMPROVED SPECTRAL LARGE SIEVE INEQUALITY FOR $SL_3(\mathbb{Z})$

MATTHEW P. YOUNG

Abstract. We prove an improved spectral large sieve inequality for the family of $SL_3(\mathbb{Z})$ Hecke-Maass cusp forms. The method of proof uses duality and its structure reveals unexpected connections to Heath-Brown's large sieve for cubic characters.

1. Introduction

A large sieve inequality for a family of automorphic forms is a flexible and versatile tool that represents quantitative orthogonality properties of the family. Strong results are known for GL_1 and GL_2 (e.g. see [M] and [IK, Chapter 7] for some surveys), but progress has been more elusive in higher rank. The main focus of this article, the $SL_3(\mathbb{Z})$ spectral large sieve, has seen some recent attention in a series of papers [B, Y, BB]. For some other notable higher rank examples, see [DK, V, TZ].

We set some notation before continuing the discussion on the large sieve. Let $\mathcal{F}^{\text{cusp}}$ denote the family of Hecke-Maass cusp forms on $SL_3(\mathbb{Z})$. Similarly, let $\mathcal{F}^{\text{Eis}}$ denote the family of $SL_3(\mathbb{Z})$ Eisenstein series induced by $SL_2(\mathbb{Z})$ cusp forms and Eisenstein series (see [G, Section 10.5] for a definition). Let $\mathcal{F} = \mathcal{F}^{\text{cusp}} \cup \mathcal{F}^{\text{Eis}}$. For $F \in \mathcal{F}$, let $\mu_F = \mu = (\mu_1, \mu_2, \mu_3) \in \mathfrak{a}_{\mathbb{C}}^*$ be its Langlands parameters, so the Ramanujan conjecture predicts that $\mu \in i\mathbb{R}^3$. Let $\lambda_F(m,n)$ denote the Hecke eigenvalues of F. Let $\Omega \subset \mathfrak{a}^*$ be compact, Weyl group invariant, and disjoint from the Weyl chamber walls. Let $B = B_V$ be a box of sidelength $V/100$, with $100 \leq V \leq T$ and $B_V \subset T\Omega$, and let $\mathcal{F}_V \subset \mathcal{F}$ denote the set of Hecke-Maass cusp forms and Eisenstein series with $\mu_F \in W(B)$, where W is the Weyl group. Write

$$\mathcal{F}_V^{\text{cusp}} = \mathcal{F}^{\text{cusp}} \cap \mathcal{F}_V, \qquad \text{and} \qquad \mathcal{F}_V^{\text{Eis}} = \mathcal{F}^{\text{Eis}} \cap \mathcal{F}_V. \tag{1.1}$$

For $F \in \mathcal{F}^{\text{cusp}}$, let $\omega_F = \text{Res}_{s=1} L(F \otimes \overline{F}, s)$. The Weyl law proved by Lapid and Müller [LM] gives that the cardinality of $\mathcal{F}_V^{\text{cusp}}$ is $V^2 T^{3+o(1)}$; technically, their result holds only for congruence subgroups of $SL_3(\mathbb{Z})$ of sufficiently large level to ensure the non-existence of elliptic fixed points, which should not be a major obstacle in their method. Work of Blomer [B, Theorem 1] gives the weighted Weyl law for $SL_3(\mathbb{Z})$ that we need here. Any potential violation to the Ramanujan conjecture must occur near the Weyl chamber walls (e.g. see [B, p.678]), so automatically any cusp form with $\mu_F \in T\Omega$ satisfies Ramanujan. For $\mathbf{a} \in \ell^2$, we use the notation $|\mathbf{a}| = \|\mathbf{a}\|_2$.

The culmination of [B, Y, BB] is the following.

1991 *Mathematics Subject Classification.* Primary 11F03; Secondary 11F55, 11F72, 11M41.

Key words and phrases. Automorphic forms, large sieve inequality, Fourier coefficients, functional equation, Rankin-Selberg L-function.

This material is based upon work supported by the National Science Foundation under agreement No. DMS-2001306. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Theorem 1.1 ([BB]). *We have*

$$\sum_{*\in\{cusp,Eis\}}\sum_{F\in\mathcal{F}_T^*}\frac{1}{\omega_F}\Big|\sum_{N\le n\le 2N}a_n\lambda_F(1,n)\Big|^2\ll(T^5+T^2N)^{1+\varepsilon}|\mathbf{a}|^2, \tag{1.2}$$

$$\sum_{*\in\{cusp,Eis\}}\sum_{F\in\mathcal{F}_1^*}\frac{1}{\omega_F}\Big|\sum_{N\le n\le 2N}a_n\lambda_F(1,n)\Big|^2\ll(T^3+T^2N)^{1+\varepsilon}|\mathbf{a}|^2. \tag{1.3}$$

The formulations of (1.2) and (1.3) are imprecise because we have not fully described the meaning of $\sum_{F\in\mathcal{F}^{\mathrm{Eis}}}\omega_F^{-1}$ for F an Eisenstein series; see e.g. [BB, Section 4] for the correct normalizing factor. The proof of Theorem 1.1 notably relies on the GL_3 Kuznetsov formula. The spectral side of the Kuznetsov formula includes both the cusp forms as well as Eisenstein series, which explains why Theorem 1.1 covers both types of automorphic forms.

By general principles of bilinear forms (cf. [IK, Chapter 7]), the optimal bound one could expect on the right hand side of (1.2) would be $(T^5+N)^{1+\varepsilon}$, while that of (1.3) would be $(T^3+N)^{1+\varepsilon}$. However, Blomer and Buttcane showed that the term T^2N in (1.2) cannot be reduced in size, by constructing a choice of vector $\mathbf{a}$ so that the contribution from the Eisenstein series is at least $T^2N|\mathbf{a}|^2$, for $N\gg T^{3+\delta}$. An examination of the proof of [BB, Proposition 1.3] shows their method leads to a lower bound of size $TN|\mathbf{a}|^2$ for the left hand side of (1.3). In Section 8 we sketch an alternative method to produce this lower bound.

A natural question is if the bounds (1.2)-(1.3) can be improved when the family is restricted to cusp forms. The main result of this article affirms this.

Theorem 1.2. *We have*

$$\sum_{F\in\mathcal{F}_1^{cusp}}\frac{1}{\omega_F}\Big|\sum_{N\le n\le 2N}a_n\lambda_F(1,n)\Big|^2\ll(T^5+N+T^2N^{2/3})^{1+\varepsilon}|\mathbf{a}|^2. \tag{1.4}$$

Note that the right hand side of (1.4) is smaller than the right hand side of (1.3) for $N\gg T^{3+\varepsilon}$, and is also just as good as the "$N+T^3$" theoretically optimal bound for $N\gg T^6$.

The starting point of our proof is to use the duality principle and the functional equation of Rankin-Selberg L-functions on $GL_3\times GL_3$. This method is most effective when N is large, since this makes the dual length of summation relatively shorter. The final step in our proof is an application of the Buttcane-Blomer bound (1.3), which is strongest for relatively small values of N.

The method of Thorner and Zaman [TZ] leads to the bound

$$\sum_{F\in\mathcal{F}_1^{\mathrm{cusp}}}\frac{1}{\omega_F}\Big|\sum_{N\le n\le 2N}a_n\lambda_F(1,n)\Big|^2\ll(N+T^6)^{1+\varepsilon}|\mathbf{a}|^2. \tag{1.5}$$

They also use duality and a contour-shifting argument, but apply the convexity bound for the $GL_3\times GL_3$ Rankin-Selberg L-functions after shifting near to the 0-line. For comparison, both (1.4) and (1.5) agree for $N\gg T^6$, but (1.4) is superior for $N\ll T^6$.

A curious aspect of the proof is that is reveals that certain aspects of the family $\mathcal{F}$ are in analogy with the family of cubic Hecke characters. A large sieve inequality for this latter family was proved by Heath-Brown [H-B] with an application to the problem of estimating sums of cubic Gauss sums of prime arguments. See Section 7 below for a more thorough discussion of Heath-Brown's work and its connections to our proof of Theorem 1.2. Very recently, Dunn and Radziwiłł [DR] have shown that Heath-Brown's result is (surprisingly!)

optimal, due to the existence of an exceptionally large main term related to Kummer's bias in cubic Gauss sums. This main term is comparable to the term $(T^3N)^{2/3}$ in (1.4).

For simplicity, Theorem 1.2 treats the localized family $\mathcal{F}_1$ but in principle one could use the same method of proof to study $\mathcal{F}_T$ as well. Typically, small families are more difficult than large families, so one might expect that a bound on $\mathcal{F}_T$ would be even easier to prove than that for $\mathcal{F}_1$. However, our proof of Theorem 1.2 using duality requires the conductor of $L(1/2, F \otimes \overline{G})$ for $F, G \in \mathcal{F}$ which is a bit simpler to express for $F, G \in \mathcal{F}_1$ than for general $F, G \in \mathcal{F}_T$. Generically, for $F, G \in \mathcal{F}_T$, the conductor of $L(1/2, F \otimes \overline{G})$ is of size T^9 but there are various conductor-dropping ranges to consider. Indeed, for $F, G \in \mathcal{F}_1$, the conductor of $L(1/2, F \otimes \overline{G})$ is of size T^6. Another issue to mention is that the root number in the wider family $\mathcal{F}_T$ is more difficult to handle (one wishes to separate variables) than in $\mathcal{F}_1$, since in this narrow family the root number is practically constant.

This discussion indicates that the approach via duality is beneficial when $N \gg T^{3+\delta}$ (since T^3 is the square-root of the conductor in the narrow family $\mathcal{F}_1$), consistent with the remark above that (1.4) is an improvement in this range.

2. Acknowledgements

I thank Valentin Blomer and Jack Buttcane for helpful feedback on this paper.

3. Preliminaries

3.1. Maass forms on $SL_3(\mathbb{Z})$.

Lemma 3.1 (Hecke relations). *Let $F \in \mathcal{F}$. Then*

$$\lambda_F(m,1)\lambda_F(1,n) = \sum_{d|(m,n)} \lambda_F\Big(\frac{m}{d}, \frac{n}{d}\Big), \tag{3.1}$$

and

$$\lambda_F(m,n) = \sum_{d|(m,n)} \mu(d)\lambda_F\Big(\frac{m}{d}, 1\Big)\lambda_F\Big(1, \frac{n}{d}\Big). \tag{3.2}$$

Moreover,

$$\lambda_F(m,n) = \overline{\lambda_F}(n,m). \tag{3.3}$$

The relation (3.1) appears in [G, Theorem 6.4.11], from which (3.2) follows from Möbius inversion. For (3.3), see [G, Theorem 9.3.11 Addendum].

Lemma 3.2 (Convexity bound). *For any $F \in \mathcal{F}_T$ and any $X \geq 1$ we have*

$$\sum_{m^2 n \leq X} |\lambda_F(m,n)|^2 \ll_\varepsilon X(XT)^\varepsilon. \tag{3.4}$$

This follows from work of Xiannan Li [Li].

Lemma 3.3. *Let $F, G \in \mathcal{F}^{cusp}$. The Rankin-Selberg L-function $L(s, F \otimes \overline{G})$ is defined by*

$$L(s, F \otimes \overline{G}) = \sum_{d,m,n \geq 1} \frac{\lambda_F(m,n)\overline{\lambda}_G(m,n)}{(d^3m^2n)^s}. \tag{3.5}$$

It has meromorphic continuation to $s \in \mathbb{C}$ with a possible pole at $s = 1$ only, and satisfies the functional equation

$$\gamma(s, F \otimes \overline{G})L(s, F \otimes \overline{G}) = \gamma(1-s, G \otimes \overline{F})L(1-s, G \otimes \overline{F}), \tag{3.6}$$

where

$$\gamma(s, F \otimes \overline{G}) = \gamma(s, \mu_F, \mu_G) = \prod_{i,j=1}^{3} \Gamma_{\mathbb{R}}(s + \mu_i(F) + \overline{\mu_j}(G)). \tag{3.7}$$

The pole at $s = 1$ exists if and only if $F = G$.

For a reference, see [G, Theorem 7.4.9, Proposition 11.6.17].

3.2. **Separation of variables.**

Lemma 3.4. *Suppose f is Schwartz-class. Then*

$$f(x) = \int_{-\infty}^{\infty} \widehat{f}(y) e(xy) dy, \tag{3.8}$$

where $\|\widehat{f}\|_1 \ll \|f\|_1 + \|f''\|_1$. The implied constant is absolute.

Proof. For any $y \in \mathbb{R}$, we have $|\widehat{f}(y)| \le \|f\|_1$. Integrating by parts twice, we similarly deduce $|\widehat{f}(y)| \le (2\pi y)^{-2}\|f''\|_1$. We use the former bound for $|y| \le 1$, and the latter for $|y| > 1$. □

To give an idea of how we wish to use Lemma 3.4 to separate variables, consider the following example.

Example 3.5. *Suppose that f is Schwartz-class. Moreover suppose that γ_m and δ_n are some sequences of real numbers and that I is some finite set of integers. Then*

$$\max_{|\mathbf{b}|=1} \Big| \sum_{m,n \in I} b_m \overline{b_n} f(\gamma_m + \delta_n) \Big| \le \|\widehat{f}\|_1 \max_{|\mathbf{b}|=1} \Big| \sum_{m \in I} b_m \Big|^2. \tag{3.9}$$

where $|\mathbf{b}| = (\sum_{n \in I} |b_n|^2)^{1/2}$.

Proof. By Lemma 3.4, we have

$$\Big| \sum_{m,n} b_m \overline{b_n} f(\gamma_m + \delta_n) \Big| = \Big| \int_{-\infty}^{\infty} \widehat{f}(y) \Big(\sum_m b_m e(\gamma_m y) \Big) \Big(\sum_n \overline{b_n} e(\delta_n y) \Big) dy \Big| \tag{3.10}$$

$$\le \|\widehat{f}\|_1 \max_{y \in \mathbb{R}} \Big| \sum_m b_m e(\gamma_m y) \Big| \Big| \sum_n \overline{b_n} e(\delta_n y) \Big|. \tag{3.11}$$

Taking the maximum over $|\mathbf{b}| = 1$ immediately gives the result. □

4. Definitions of norms and some relations between them

We begin by defining the basic norm that appears (implicitly) in Theorem 1.2:

$$\Delta_1(\mathcal{F}_V, N) = \max_{|\mathbf{a}|=1} \sum_{F \in \mathcal{F}_V^{\mathrm{cusp}}} \frac{1}{\omega_F} \Big| \sum_{N \le n \le 2N} a_n \lambda_F(1, n) \Big|^2. \tag{4.1}$$

By the duality principle (cf. [IK, p.170]), we have $\Delta_1(\mathcal{F}_V, N) = \Delta^{(1)}(\mathcal{F}_V, N)$, where

$$\Delta^{(1)}(\mathcal{F}_V, N) = \max_{|\mathbf{b}|=1} \sum_{N \le n \le 2N} \Big| \sum_{F \in \mathcal{F}_V^{\mathrm{cusp}}} b_F \omega_F^{-1/2} \lambda_F(1, n) \Big|^2, \tag{4.2}$$

and where $|\mathbf{b}|^2 = \sum_F |b_F|^2$. We also define a related norm $\Delta_2(\mathcal{F}_V, N) = \Delta^{(2)}(\mathcal{F}_V, N)$ by

$$\Delta_2(\mathcal{F}_V, N) = \max_{|\mathbf{a}|=1} \sum_{F\in\mathcal{F}_V^{\mathrm{cusp}}} \frac{1}{\omega_F}\Big| \sum_{N\le m^2n\le 2N} a_{m,n}\lambda_F(m,n)\Big|^2, \tag{4.3}$$

and where

$$\Delta^{(2)}(\mathcal{F}_V, N) = \max_{|\mathbf{b}|=1} \sum_{N\le m^2n\le 2N} \Big| \sum_{F\in\mathcal{F}_V^{\mathrm{cusp}}} b_F\omega_F^{-1/2}\lambda_F(m,n)\Big|^2. \tag{4.4}$$

Finally we define a third norm $\Delta_3(\mathcal{F}_V, N) = \Delta^{(3)}(\mathcal{F}_V, N)$ by

$$\Delta_3(\mathcal{F}_V, N) = \max_{|\mathbf{a}|=1} \sum_{F\in\mathcal{F}_V^{\mathrm{cusp}}} \frac{1}{\omega_F}\Big| \sum_{N\le d^3m^2n\le 2N} a_{d,m,n}\lambda_F(m,n)\Big|^2, \tag{4.5}$$

and where

$$\Delta^{(3)}(\mathcal{F}_V, N) = \max_{|\mathbf{b}|=1} \sum_{N\le d^3m^2n\le 2N} \Big| \sum_{F\in\mathcal{F}_V^{\mathrm{cusp}}} b_F\omega_F^{-1/2}\lambda_F(m,n)\Big|^2. \tag{4.6}$$

We obviously have $\Delta_1(\mathcal{F}_V, N) \le \Delta_2(\mathcal{F}_V, N) \le \Delta_3(\mathcal{F}_V, N)$. We also want relations in the other direction.

Lemma 4.1. *We have*

$$\Delta_3(\mathcal{F}_V, N) \ll (\log N) \max_{R\ll N} \Big(\frac{N}{R}\Big)^{1/3} \Delta_2(\mathcal{F}_V, R). \tag{4.7}$$

Proof. We prove this on the dual side, using (4.6) and (4.4). By breaking the sum up so $R \le m^2n \le 2R$ and summing R over dyadic segments, we obtain

$$\Delta^{(3)}(\mathcal{F}_V, N) \ll (\log N) \max_{1\ll R\ll N} \Big(\frac{N}{R}\Big)^{1/3} \max_{|\mathbf{b}|=1} \sum_{R\le m^2n\le 2R} \Big| \sum_{F\in\mathcal{F}_V^{\mathrm{cusp}}} b_F\omega_F^{-1/2}\lambda_F(m,n)\Big|^2. \tag{4.8}$$

The result follows immediately. □

Lemma 4.2. *We have*

$$\Delta_2(\mathcal{F}_V, N) \ll (NT)^{\varepsilon} \max_{Y^2X\ll N} \min\Big(Y\Delta_1(\mathcal{F}_V, X), X\Delta_1(\mathcal{F}_V, Y)\Big). \tag{4.9}$$

Proof. Again, we prove this on the dual side, using (4.4) and (4.2). By the Hecke relation (3.2), we deduce

$$\Delta^{(2)}(\mathcal{F}_V, N) = \max_{|\mathbf{b}|=1} \sum_{N\le m^2n\le 2N} \Big| \sum_{d|(m,n)} \mu(d) \sum_{F\in\mathcal{F}_V^{\mathrm{cusp}}} b_F\omega_F^{-1/2}\lambda_F\Big(\frac{m}{d}, 1\Big)\lambda_F\Big(1, \frac{n}{d}\Big)\Big|^2. \tag{4.10}$$

Applying Cauchy's inequality and a divisor function bound to take the sum over d to the outside, we deduce

$$\Delta^{(2)}(\mathcal{F}_V, N) \ll N^{\varepsilon} \max_{|\mathbf{b}|=1} \sum_{N\le m^2n\le 2N} \sum_{d|(m,n)} \Big| \sum_{F\in\mathcal{F}_V^{\mathrm{cusp}}} b_F\omega_F^{-1/2}\lambda_F\Big(\frac{m}{d}, 1\Big)\lambda_F\Big(1, \frac{n}{d}\Big)\Big|^2. \tag{4.11}$$

Interchanging the order of summation and changing variables $m \to dm$ and $n \to dn$, we obtain

$$\Delta^{(2)}(\mathcal{F}_V, N) \ll N^{\varepsilon} \max_{|\mathbf{b}|=1} \sum_{N\le d^3m^2n\le 2N} \Big| \sum_{F\in\mathcal{F}_V^{\mathrm{cusp}}} b_F\omega_F^{-1/2}\lambda_F(m,1)\lambda_F(1,n)\Big|^2. \tag{4.12}$$

Now we further restrict d, m and n so $dm \asymp Y$ and $n \asymp X$, and let $b'_F = b_F \lambda_F(m,1)$. Then

$$\begin{aligned}\Delta^{(2)}(\mathcal{F}_V, N) &\ll N^\varepsilon \max_{XY^2 \ll N} \max_{|\mathbf{b}|=1} \sum_{\substack{N \le d^3m^2n \le 2N \\ dm \asymp Y \\ n \asymp X}} \Big| \sum_{F \in \mathcal{F}_V^{\mathrm{cusp}}} b'_F \omega_F^{-1/2} \lambda_F(1,n) \Big|^2 \\ &\ll N^\varepsilon \max_{XY^2 \ll N} \max_{|\mathbf{b}|=1} \sum_{dm \asymp Y} \Delta^{(1)}(\mathcal{F}_V, X) \sum_{F \in \mathcal{F}_V^{\mathrm{cusp}}} |b_F|^2 |\lambda_F(m,1)|^2.\end{aligned}$$

From Lemma 3.2 we deduce $\sum_{dm \asymp Y} |\lambda_F(m,1)|^2 \ll Y(NT)^\varepsilon$, uniformly in F, leading to

$$\Delta^{(2)}(\mathcal{F}_V, N) \ll (NT)^\varepsilon \max_{Y^2X \ll N} Y \Delta^{(1)}(\mathcal{F}_V, X).$$

It remains to show that a similar bound holds but with $Y\Delta^{(1)}(\mathcal{F}_V, X)$ replaced by $X\Delta^{(1)}(\mathcal{F}_V, Y)$. This follows by going through the same proof but reversing the roles of m and n, and using (3.3) along the way. □

Chaining together Lemmas 4.1 and 4.2, we immediately deduce the following.

Lemma 4.3. *We have*

$$\Delta_3(\mathcal{F}_V, N) \ll (NT)^\varepsilon \max_{Y^2X \ll N} \Big(\frac{N}{XY^2}\Big)^{1/3} \min\Big(Y\Delta_1(\mathcal{F}_V, X), X\Delta_1(\mathcal{F}_V, Y)\Big). \tag{4.13}$$

See Section 7 for a comparison of Lemma 4.3 with [H-B, Lemma 6].

We also observe that the analogs of Lemmas 4.1–4.3 hold equally well for Eisenstein series.

5. Functional equation

In this section we use the functional equation of the Rankin-Selberg L-function to deduce the following estimate.

Lemma 5.1. *We have*

$$\Delta^{(3)}(\mathcal{F}_1, N) \ll N + \frac{N}{T^3}(NT)^\varepsilon \max_{1 \le Z \ll \frac{T^6}{N}(TN)^\varepsilon} \Delta^{(3)}\Big(\mathcal{F}_1, Z\Big). \tag{5.1}$$

The proof of Lemma 5.1 crucially uses that the family is restricted to cusp forms. The reader may examine the proof of Proposition 8.1 below to see how the family of Eisenstein series exhibits different behavior than the cusp forms.

Proof. Select a smooth nonnegative bump function w with compact support on the positive reals, satisfying $w(x) \ge 1$ for $1 \le x \le 2$. Then

$$\Delta^{(3)}(\mathcal{F}_1, N) \le \max_{|\mathbf{b}|=1} \sum_{d,m,n} w\Big(\frac{d^3m^2n}{N}\Big) \Big| \sum_{F \in \mathcal{F}_1^{\mathrm{cusp}}} b_F \omega_F^{-1/2} \lambda_F(m,n) \Big|^2. \tag{5.2}$$

Next open the square, apply Mellin inversion, and evaluate the resulting Dirichlet series using (3.5), giving

$$\Delta^{(3)}(\mathcal{F}_1, N) \le \max_{|\mathbf{b}|=1} \sum_{F,G \in \mathcal{F}_1^{\mathrm{cusp}}} \frac{b_F \overline{b_G}}{\omega_F^{1/2} \omega_G^{1/2}} \frac{1}{2\pi i} \int_{(3/2)} N^s \widetilde{w}(s) L(s, F \otimes \overline{G}) ds. \tag{5.3}$$

Next we shift the contour of integration to the line $\mathrm{Re}(s) = -\varepsilon$, change variables $s \to 1-s$, and apply the functional equation (3.6). In this process we cross a potential pole at $s = 1$

only, which exists if and only if $F = G$. Recalling that $\omega_F = \mathrm{Res}_{s=1} L(F \otimes \overline{F}, s)$, this pole contributes the term of size $O(N)$ to the right hand side of (5.1). In all we obtain $\Delta^{(3)}(\mathcal{F}_1, N)$ is at most $O(N)$ plus

$$\max_{|\mathbf{b}|=1} \Big| \sum_{F,G \in \mathcal{F}_1^{\mathrm{cusp}}} \frac{b_F \overline{b_G}}{\omega_F^{1/2} \omega_G^{1/2}} \frac{1}{2\pi i} \int_{(3/2)} N^{1-s} \widetilde{w}(1-s) \frac{\gamma(s, \mu_G, \mu_F)}{\gamma(1-s, \mu_F, \mu_G)} L(s, G \otimes \overline{F}) ds \Big|. \tag{5.4}$$

Now we examine the ratio of gamma factors appearing in (5.4). Six out of the nine gamma factors in (3.7) have $|\mu_i(F) + \overline{\mu_j}(G)|$ large, of size T (the precise size determined up to $O(1)$ by the location of the box B). The remaining three gamma factors have $|\mu_i(F) + \overline{\mu_j}(G)|$ of size $O(1)$. Moreover, since F and G automatically satisfy Ramanujan by the location of the box B, then $\mu_i(F) + \overline{\mu_j}(G) \in i\mathbb{R}$. This means that for $\mathrm{Re}(s) > 0$, we have that the ratio of gamma factors appearing in (5.4) is analytic, and satisfies the bound

$$\Big| Q^{\frac{1}{2}-s} \widetilde{w}(1-s) \frac{\gamma(s, \mu_G, \mu_F)}{\gamma(1-s, \mu_F, \mu_G)} \Big| \ll_{\mathrm{Re}(s), A} (1 + |s|)^{-A}, \tag{5.5}$$

for any $A > 0$, where $Q = T^6$. Now in (5.4) we open up the Dirichlet series, obtaining

$$\max_{|\mathbf{b}|=1} \Big| \sum_{d,m,n} \sum_{F,G \in \mathcal{F}_1^{\mathrm{cusp}}} \frac{b_F \overline{b_G} N}{\omega_F^{1/2} \omega_G^{1/2}} \frac{1}{2\pi i} \int_{(3/2)} \widetilde{w}(1-s) \frac{\gamma(s, \mu_G, \mu_F)}{\gamma(1-s, \mu_F, \mu_G)} \frac{\lambda_G(m,n) \overline{\lambda_F}(m,n)}{(d^3 m^2 n N)^s} ds \Big|. \tag{5.6}$$

We may truncate the Dirichlet series at $d^3 m^2 n \ll \frac{Q}{N} (TN)^{\varepsilon}$ with a very small error term (certainly smaller than the $O(N)$ term already accounted for), by shifting contours far to the right. Having imposed this truncation, we may then shift the contour to the line $\mathrm{Re}(s) = \varepsilon$. We may also truncate the integral at $|\mathrm{Im}(s)| \ll (NT)^{\varepsilon}$, without producing a new error term.

Then (5.6) is reduced to

$$\begin{aligned} \max_{|\mathbf{b}|=1} \Big| \sum_{d^3 m^2 n \ll \frac{T^6}{N} (NT)^{\varepsilon}} \sum_{F,G \in \mathcal{F}_1^{\mathrm{cusp}}} \frac{b_F \overline{b_G} N}{\omega_F^{1/2} \omega_G^{1/2}} \\ \frac{1}{2\pi i} \int_{\substack{\mathrm{Re}(s)=\varepsilon \\ |\mathrm{Im}(s)| \ll (NT)^{\varepsilon}}} \widetilde{w}(1-s) \frac{\gamma(s, \mu_G, \mu_F)}{\gamma(1-s, \mu_F, \mu_G)} \frac{\lambda_G(m,n) \overline{\lambda_F}(m,n)}{(d^3 m^2 n N)^s} ds \Big|. \end{aligned} \tag{5.7}$$

At a first pass, the reader is encouraged to "pretend" that $\frac{\gamma(s, \mu_G, \mu_F)}{\gamma(1-s, \mu_F, \mu_G)}$ equals $Q^{s-\frac{1}{2}}$ (which is a good first-order approximation) and continue with (5.10) to finish the proof. Unfortunately, a rigorous argument is a bit more technical. The plan is to separate the variables μ_F and μ_G in the ratio of gamma factors. The basic idea is encoded in Example 3.5. To this end, let $\mu_i(B)$, $i = 1, 2, 3$ denote a point inside the box B (the choice of point is irrelevant). Then

$$\frac{\Gamma_{\mathbb{R}}(s + \mu_i(F) + \overline{\mu_j}(G))}{\Gamma_{\mathbb{R}}(1 - s + \mu_i(F) + \overline{\mu_j}(G))} = \frac{\Gamma_{\mathbb{R}}(s + \mu_i(B) + \overline{\mu_j}(B) + i(\delta_i + \nu_j))}{\Gamma_{\mathbb{R}}(1 - s + \mu_i(B) + \overline{\mu_j}(B) + i(\delta_i + \nu_j))}, \tag{5.8}$$

where $i\delta_i = \mu_i(F) - \mu_i(B)$ and $i\nu_j = \overline{\mu_j(G)} - \overline{\mu_j(B)}$. Here $\delta_i, \nu_j = O(1)$ and are real. The goal is to separate δ_i from ν_j. Let

$$f(x) = \frac{\Gamma_{\mathbb{R}}(s + \mu_i(B) + \overline{\mu_j}(B) + ix)}{\Gamma_{\mathbb{R}}(s + \mu_i(B) + \overline{\mu_j}(B))} \frac{\Gamma_{\mathbb{R}}(1 - s + \mu_i(B) + \overline{\mu_j}(B))}{\Gamma_{\mathbb{R}}(1 - s + \mu_i(B) + \overline{\mu_j}(B) + ix)}. \tag{5.9}$$

By Stirling, for $x \in \mathbb{R}$ and $|x| \ll 1$, we have $|f(x)| + |f''(x)| \ll T^\varepsilon$. By Lemma 3.4 and (3.9), in effect this means we can separate the variables δ_i, ν_j at "cost" at most T^ε. Applying this with each of the gamma factors, we obtain that (5.7) is bounded by

$$\frac{N}{T^3}(NT)^\varepsilon \max_{|\mathbf{b}|=1} \sum_{d^3m^2n \ll \frac{T^6}{N}(NT)^\varepsilon} \Big| \sum_{F,G \in \mathcal{F}_1^{\mathrm{cusp}}} \frac{b_F\overline{b_G}}{\omega_F^{1/2}\omega_G^{1/2}} \lambda_G(m,n)\overline{\lambda_F}(m,n)\Big| \ll \frac{N}{T^3}(NT)^\varepsilon \max_{1 \le Z \ll \frac{T^6}{N}(TN)^\varepsilon} \Delta^{(3)}\Big(\mathcal{F}_1, Z\Big). \quad \square \tag{5.10}$$

6. Completion of the proof

Now we prove Theorem 1.2. We chain together the results from Section 4 as well as Lemma 5.1, giving

$$\begin{aligned} \Delta_1(\mathcal{F}_1, N) \le \Delta^{(3)}(\mathcal{F}_1, N) &\ll N + \frac{N}{T^3}(NT)^\varepsilon \max_{1 \le Z \ll \frac{T^6}{N}(TN)^\varepsilon} \Delta^{(3)}\Big(\mathcal{F}_1, Z\Big) \\ &\ll N + \frac{N}{T^3}(NT)^\varepsilon \max_{Y^2X \ll \frac{T^6}{N}(NT)^\varepsilon} \Big(\frac{T^6/N}{XY^2}\Big)^{1/3} \min\Big(Y\Delta_1(\mathcal{F}_1, X), X\Delta_1(\mathcal{F}_1, Y)\Big). \end{aligned} \tag{6.1}$$

This sequence of inequalities is reminiscent of (and somewhat inspired by) [H-B, Section 8]. Finally we insert the Blomer-Buttcane bound $\Delta_1(\mathcal{F}_1, M) \ll (T^3 + T^2M)(TM)^\varepsilon$ from Theorem 1.1. In all, we obtain

$$\begin{aligned} \Delta_1(\mathcal{F}_1, N) &\ll N + \frac{N}{T^3}(NT)^\varepsilon \max_{Y^2X \le \frac{T^6}{N}} \Big(\frac{T^6/N}{XY^2}\Big)^{1/3} \min\Big(Y(T^3 + T^2X), X(T^3 + T^2Y)\Big) \\ &\ll N + \frac{N}{T^3}(NT)^\varepsilon \max_{Y^2X \le \frac{T^6}{N}} \Big(\frac{T^6/N}{XY^2}\Big)^{1/3} T^2\Big(XY + T\min(X,Y)\Big) \\ &\ll N + \frac{N}{T^3}T^2\Big(\frac{T^6}{N} + T\Big(\frac{T^6}{N}\Big)^{1/3}\Big)(NT)^\varepsilon \ll N + (NT)^\varepsilon((T^3N)^{2/3} + T^5), \end{aligned}$$

completing the proof of Theorem 1.2.

7. Cubic characters

In this section we briefly recall the large sieve inequality of Heath-Brown for cubic characters [H-B] for the purpose of developing an analogy with the $SL_3(\mathbb{Z})$ cusp form family considered in this paper.

Let $\theta = \exp(2\pi i/3)$, and for nonzero $m, n \in \mathbb{Z}[\theta]$ let $(m/n)_3$ denote the cubic residue symbol. The cubic reciprocity law gives that $(m/n)_3 = (n/m)_3$. Let $N(\cdot)$ denote the norm map of $\mathbb{Q}[\omega]/\mathbb{Q}$. Let

$$\Delta_1(M, Q) = \max_{|\mathbf{a}|=1} \sideset{}{^*}\sum_{\substack{n \in \mathbb{Z}[\omega] \\ N(n) \le M}} \Big| \sideset{}{^*}\sum_{\substack{q \in \mathbb{Z}[\theta] \\ N(q) \le Q}} a_q \Big(\frac{n}{q}\Big)_3\Big|^2, \tag{7.1}$$

where the symbol $\sum^*$ means the sums are restricted to (nonzero) square-free integers. Heath-Brown's cubic large sieve is the bound

$$\Delta_1(M, Q) \ll (M + Q + (MQ)^{2/3})(MQ)^\varepsilon. \tag{7.2}$$

To make the notation appear more similar to the $SL_3(\mathbb{Z})$ family, define (for $m, n, q \in \mathbb{Z}[\theta]$)

$$\lambda_q(m,n) = \Big(\frac{n}{q}\Big)_3 \overline{\Big(\frac{m}{q}\Big)_3}. \tag{7.3}$$

Note the simple identities which the reader is invited to compare with Lemma 3.1:

$$\lambda_q(m,n) = \overline{\lambda_q(n,m)} = \lambda_q(mn^2,1) = \lambda_q(1,nm^2) = \lambda_q(m,1)\lambda_q(1,n).$$

Also observe $\lambda_q(d^3,1) = 1$ for $(d,q) = 1$.

Heath-Brown's first step is to drop the condition that n is square-free in (7.2), leading to the definition

$$\Delta_3(M,Q) = \max_{|\mathbf{a}|=1} \sum_{\substack{d,m,n\in\mathbb{Z}[\theta]\\ N(d^3m^2n)\le M}} |\mu(mn)| \Big| \sum_{\substack{q\in\mathbb{Z}[\theta]\\ N(q)\le Q}}{}^{*} a_q \lambda_q(m,n)\Big|^2 \tag{7.4}$$

Obviously $\Delta_1(M,Q) \le \Delta_3(M,Q)$, which parallels our relation $\Delta_1(\mathcal{F}_V,N) \le \Delta_3(\mathcal{F}_V,N)$. The same steps used to prove Lemma 4.3 can be applied here to show

$$\Delta_3(M,Q) \ll (MQ)^{\varepsilon} \max_{XY^2\ll M} \Big(\frac{M}{XY^2}\Big)^{1/3} \min(X\Delta_1(Y,Q), Y\Delta_1(X,Q)), \tag{7.5}$$

which is essentially [H-B, Lemma 6].

Heath-Brown also gives a relationship between $\Delta_3(M,Q)$ and $\Delta_3(Q^2/M,Q)$ (see [H-B, Lemmas 7 and 8] for the precise statement) which arises from the functional equation and is analogous to Lemma 5.1.

8. Lower bound via duality

Proposition 8.1. *There exists a choice of vector* $\mathbf{a}$ *so that*

$$\sum_{F\in\mathcal{F}_1^{Eis}} \frac{1}{\omega_F} \Big| \sum_{N\le n\le 2N} a_n \lambda_F(1,n)\Big|^2 \gg (TN)^{1-\varepsilon}|\mathbf{a}|^2, \tag{8.1}$$

for $N \gg T^{7/3+\delta}$.

Since a lower bound of this type (however, for the wide family $\mathcal{F}_T^{\mathrm{Eis}}$) was already proved in [BB, Proposition 1.3], for brevity we only give a sketch which could be made rigorous with more work. Blomer and Buttcane [BB, Section 4] showed that the lower bound of size "T^2N" in (1.2) comes from the Eisenstein series $E(z, 1/2+it, u_j)$ induced from $SL_2(\mathbb{Z})$ cusp forms u_j. This Eisenstein series E has Hecke eigenvalues

$$\lambda_E(1,n) = \lambda(n) = \sum_{d_1d_2=n} \lambda_j(d_1) d_1^{-it} d_2^{2it},$$

and Langlands parameters $\mu = (2it, -it+it_j, -it-it_j)$, where t_j is the spectral parameter of u_j. Moreover, $\omega_F = T^{o(1)}$, so we will drop this aspect in the proof.

Proof. To simplify notation, say that B is the spectral ball of size $O(1)$ centered at $i(2T,T,-3T)$. This means $t = T+O(1)$ and $t_j = T+O(1)$. The contribution of this family of Eisenstein series, on the dual side, takes the form (after smoothing)

$$\mathcal{S} := \sum_n w(n/N) \Big| \int_{t,t_j=T+O(1)} \beta_{t,t_j} \lambda(n)\Big|^2. \tag{8.2}$$

Expanding the square, we obtain

$$\mathcal{S} = \int_{t,t',t_j,t_j'=T+O(1)} \beta\overline{\beta'} \frac{1}{2\pi i} \int_{(1+\varepsilon)} N^s \widetilde{w}(s) \underbrace{\sum_{n=1}^{\infty} \frac{\lambda(n)\overline{\lambda'}(n)}{n^s}}_{Z_{u_j,u_j',t,t'}(s)} ds. \tag{8.3}$$

Note

$$\begin{aligned} Z_{u_j,u_j',t,t'}(s) &= \sum_{d_1d_2=e_1e_2} \frac{\lambda_j(d_1)d_1^{-it}d_2^{2it}\lambda_j'(e_1)e_1^{it'}e_2^{-2it'}}{(d_1d_2)^s} \\ &= \prod_p (1 + p^{-s}[\lambda_j(p)\lambda_j'(p)p^{-it+it'} + \lambda_j(p)p^{-it-2it'} + \lambda_j'(p)p^{2it+it'} + p^{2it-2it'}] + O_{u_j,u_j'}(p^{-2s})). \end{aligned} \tag{8.4}$$

With some care, including use of the convexity bound for $GL_2 \times GL_2$ L-functions due to Iwaniec [I], one may then derive

$$Z_{u_j,u_j',t,t'}(s) = \zeta(s-2it+2it')L(s+it+2it', u_j)L(s-2it-it', u_j')L(s+it-it', u_j \otimes u_j')A(s), \tag{8.5}$$

where $A(s) = A_{u_j,u_j',t,t'}(s)$ is given by an absolutely convergent Euler product for $\mathrm{Re}(s) > 1/2$, satisfying $|A(\sigma + iy)| \ll_\sigma T^\varepsilon$, for $\sigma > 1/2$.

Returning to (8.3), we shift contours to the line $\mathrm{Re}(s) = 1/2 + \varepsilon$. Note the pole of zeta at $s = 1 + 2it - 2it'$ which occurs for all pairs u_j, u_j'. This polar term, say denoted $\mathcal{S}_0$, contributes (roughly)

$$\int_{t,t',t_j,t_j'=T+O(1)} \beta\beta' N^{1+2it-2it'} \widetilde{w}(1+2it-2it')L(1+3it, u_j)L(1-3it', u_j')L(1+3it-3it', u_j \otimes u_j'). \tag{8.6}$$

If we choose $\beta_{t,t_j} = L(1 + 3it, u_j)^{-1}$ (alternatively, one could take $\beta_{t,t_j} = \overline{L(1 + 3it, u_j)}$) then this polar term becomes approximately

$$N \sum_{t_j,t_j'=T+O(1)} L(1, u_j \otimes u_j') \approx NT^2. \tag{8.7}$$

With a bit more care, one can derive $|\mathcal{S}_0| \gg (NT^2)^{1-\varepsilon}$ for this choice of β.

Next we estimate the contribution to $\mathcal{S}$ from the new line of integration at $\mathrm{Re}(s) = 1/2+\varepsilon$; call this $\mathcal{S}'$. Jutila and Motohashi [JM] showed a Weyl bound for the $SL_2(\mathbb{Z})$ cusp forms, namely

$$L(1/2 + it, u_j) \ll (1 + |t| + |t_j|)^{1/3+\varepsilon}. \tag{8.8}$$

Combining this with the convexity bound for the $GL_2 \times GL_2$ factor in (8.5) (which has conductor of size T^2) gives $|Z(\sigma + iy)| \ll T^{7/6+\varepsilon}$. Note that the ζ-factor is evaluated at $\sigma + iy$ with $|y| \ll T^\varepsilon$, so it practically gives no contribution here. Hence, for this choice of β, we have

$$\mathcal{S}' \ll N^{1/2+\varepsilon}T^{7/6+\varepsilon} \int_{t,t',t_j,t_j'=T+O(1)} |\beta\beta'| \ll N^{1/2+\varepsilon}T^{7/6+\varepsilon}T^2. \tag{8.9}$$

Thus

$$|\mathcal{S}| \gg (NT^2)^{1-\varepsilon} + O(N^{1/2+\varepsilon}T^{7/6+\varepsilon}T^2). \tag{8.10}$$

Note the polar term dominates the error term provided $N \gg T^{7/3+\delta}$. Finally, we observe that

$$\int_{t,t_j=T+O(1)} |\beta_{t,t_j}|^2 dt = T^{1+o(1)}, \tag{8.11}$$

whence $|\mathcal{S}| \gg (NT)^{1-\varepsilon} \int |\beta|^2$ with this choice of β. □

9. Loose ends

We list a few possible directions for future work.

(1) Extend Theorem 1.2 to cover the family $\mathcal{F}_V$ for more general V, with $1 \ll V \ll T$.
(2) Extend Theorem 1.1, which gives a bound on the norm Δ_1 using the Kuznetsov formula, to directly give a bound on the norm Δ_2 (modified to include the Eisenstein series as well as the cusp forms). The point would be to bypass the use of Lemma 4.2 in (6.1), though it is unclear if any improvement is possible this way.
(3) Is it possible to use the $SL_3(\mathbb{Z})$ Kuznetsov formula to directly bound the cuspidal part of the spectrum in the large sieve inequality? (By subtracting off the Eisenstein parts, which one would presumably then control with lower-rank tools such as the GL_2 Kuznetsov formula.) As a point of reference, Luo has succesfully isolated the cuspidal spectrum in a GL_2 large sieve problem [Lu].
(4) Can the term T^2N in (1.3) be reduced to TN, to match the lower bound from Proposition 8.1? If so, this would immediately improve Theorem 1.2 by replacing the term T^5 by T^4.
(5) By analogy with [H-B, DR], determine if the term $(T^3N)^{2/3}$ in (1.4) is not removable.

References

[B] V. Blomer, *Applications of the Kuznetsov formula on GL*(3). Invent. Math. 194 (2013), no. 3, 673–729.

[BB] V. Blomer and J. Buttcane, *Global decomposition of GL*(3) *Kloosterman sums and the spectral large sieve.* J. Reine Angew. Math. 757 (2019), 51–88

[DK] W. Duke, and E. Kowalski, *A problem of Linnik for elliptic curves and mean-value estimates for automorphic representations.* With an appendix by Dinakar Ramakrishnan. Invent. Math. 139 (2000), no. 1, 1–39.

[DR] A. Dunn and M. Radziwiłł, *Bias in cubic Gauss sums: Patterson's conjecture.* arXiv:2109.07463

[G] D. Goldfeld, *Automorphic forms and L-functions for the group* $\mathrm{GL}(n,\mathbb{R})$. With an appendix by Kevin A. Broughan. Cambridge Studies in Advanced Mathematics, 99. Cambridge University Press, Cambridge, 2006.

[H-B] D.R. Heath-Brown, *Kummer's conjecture for cubic Gauss sums.* Israel J. Math. 120 (2000), part A, 97–124.

[I] H. Iwaniec, *The spectral growth of automorphic L-functions.* J. Reine Angew. Math. 428 (1992), 139–159.

[IK] H. Iwaniec and E. Kowalski, *Analytic number theory.* American Mathematical Society Colloquium Publications, 53. American Mathematical Society, Providence, RI, 2004.

[JM] M. Jutila and Y. Motohashi, *Uniform bound for Hecke L-functions.* Acta Math. 195 (2005), 61–115.

[LM] E. Lapid and W. Müller, *Spectral asymptotics for arithmetic quotients of* $\mathrm{SL}(n,\mathbb{R})/\mathrm{SO}(n)$. Duke Math. J. 149 (2009), no. 1, 117–155.

[Li] Xiannan Li, *Upper bounds on L-functions at the edge of the critical strip.* Int. Math. Res. Not. IMRN 2010, no. 4, 727–755.

[Lu] W. Luo, *The spectral mean value for linear forms in twisted coefficients of cusp forms.* Acta Arith. 70 (1995), no. 4, 377–391.

[M] H. Montgomery, *Topics in multiplicative number theory.* Lecture Notes in Mathematics, Vol. 227. Springer-Verlag, Berlin-New York, 1971. ix+178 pp.

[TZ] J. Thorner and A. Zaman, *An unconditional GL_n large sieve*, Adv. Math. 378 (2021), 107529, 24 pp.

[V] A. Venkatesh, *Large sieve inequalities for $GL(n)$-forms in the conductor aspect.* Adv. Math. 200 (2006), no. 2, 336–356.
[Y] M. Young, *Bilinear forms with GL_3 Kloosterman sums and the spectral large sieve.* Int. Math. Res. Not. IMRN 2016, no. 21, 6453–6492.

Department of Mathematics, Texas A&M University, College Station, TX 77843-3368, U.S.A.
Email address: myoung@math.tamu.edu