NON-ABELIAN AMPLIFICATION AND BILINEAR FORMS WITH KLOOSTERMAN SUMS

ALEXANDRU PASCADI

Abstract. We introduce a new method to bound bilinear (Type II) sums of Kloosterman sums with composite moduli c, using Fourier analysis on $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ and an amplification argument with non-abelian characters. For sums of length $\sqrt{c}$, our method produces a non-trivial bound for all moduli except near-primes, saving $c^{-1/12}$ for products of two primes of the same size. Combining this with previous results for prime moduli, we achieve savings beyond the Pólya–Vinogradov range for all moduli. We give applications to moments of twisted cuspidal L-functions, and to large sieve inequalities for exceptional cusp forms with composite levels.

Contents

1. Introduction

1.1. **Brief background.** There is by now a fairly comprehensive history of bounds for bilinear forms with Kloosterman sums and their applications [3, 1, 2, 14, 24, 25, 11, 28, 32, 40, 39, 22, 46, 44]. In the simplest form, the objects of interest are the sums

$$\sum_{m\le M}\sum_{n\le N}\alpha_m\beta_n S(m,n;c), \qquad \text{where} \qquad S(m,n;c) := \sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times} e\left(\frac{mx+n\overline{x}}{c}\right), \tag{1.1}$$

for positive integers c, M, N with $M, N \le c$ and complex sequences (α_m), (β_n); here $e(t) := \exp(2\pi i t)$ and $x\overline{x} \equiv 1 \pmod c$. In this work, we are mainly concerned with the 'Type II' setting where (α_m) and (β_n) are arbitrary sequences, and we search for an upper bound in terms of their ℓ^2 norms $\|\alpha\| := (\sum_m |\alpha_m|^2)^{1/2}$, $\|\beta\| := (\sum_n |\beta_n|^2)^{1/2}$. This is equivalent to bounding the operator norm, or the largest singular value, of the $M\times N$ matrix $(S(m,n;c))_{m\le M, n\le N}$.

For the Type II sums, it is in general necessary to incorporate a coprimality constraint $(m,n,c)=1$. In practice, since $S(gm,gn;gc) = \frac{\phi(gc)}{\phi(c)}S(m,n;c)$, one can separately consider each value of (m,n,c); therefore, in most bounds discussed henceforth, one can replace the restriction $(m,n,c)=1$ with the assumption that (α_m) and (β_n) are 1-bounded (and the norms $\|\alpha\|$, $\|\beta\|$ with $\sqrt{M}$, $\sqrt{N}$).

There are two main 'trivial' bounds to beat, packaged together into the following inequality with a slightly more general setup. For any (integer) intervals $\mathcal{I}, \mathcal{J} \subset \mathbb{Z}$ with $|\mathcal{I}| = M \leq c$, $|\mathcal{J}| = N \leq c$, any complex sequences $(\alpha_m)_{m\in\mathcal{I}}$, $(\beta_n)_{n\in\mathcal{J}}$, and any $a \in (\mathbb{Z}/c\mathbb{Z})^\times$, one has

$$\sum_{\substack{m\in\mathcal{I}, n\in\mathcal{J} \\ (m,n,c)=1}} \alpha_m \beta_n S(am, n; c) \ll \|\alpha\| \|\beta\| c^{o(1)} \min\left(c, \sqrt{MNc}\right). \tag{1.2}$$

The term c comes from Fourier analysis (a.k.a. the Pólya–Vinogradov method[1]) and is sharp when $M = N = c$, while the term $\sqrt{MNc}$ comes from the pointwise Weil bound $S(am, n; c) \ll c^{o(1)}\sqrt{(m,n,c)c}$, and performs better when $MN < c$. The best one could hope for is the perfect-orthogonality bound $\|\alpha\|\|\beta\|c^{o(1)}\sqrt{(M+N)c}$, but making any improvement over (1.2), particularly in the range $MN \approx c$ where the two trivial bounds match, is notoriously difficult. We note that while many applications [24, 25] require an improvement of the pointwise Weil bound when $MN \ll c$ (i.e., beyond the Pólya–Vinogradov range), some applications require an improvement of the Fourier-theoretic bound for larger values of $M, N \leq c^{1-o(1)}$; this is the case in our Section 9.

The first improvement of (1.2) when $MN \approx c$ was the celebrated breakthrough of Kowalski–Michel–Sawin [24], which requires a prime modulus $c = p$, and which saves a factor of $p^{-1/64}$ when $M, N \asymp \sqrt{p}$; see also [25] for their follow-up work, which outperforms the pointwise Weil bound for MN as small as $p^{3/4+o(1)}$. These results rely on a shift-by-ab trick of Vinogradov and Karatsuba, a Hölder step, and deep inputs of ℓ-adic cohomology; notably, the same bounds hold for more general algebraic trace functions, including hyper-Kloosterman sums.

Closer to our methods is an approach of Shkredov [38] for prime moduli p, which relies (in the Type II setting) on non-abelian Fourier analysis [37, Lemma 22], expansion in $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ [37, Theorem 50], and combinatorics; this beats (1.2) in the full range $M, N \in (p^{1/2-\delta}, p^{1-o(1)})$ with a small but effective power saving, and for sequences (α_m), (β_n) with more general additively-structured supports. We also mention some related additive-combinatorial approaches of Shparlinski–Zhang [40] for smooth sequences, of the author [32, §4] for additively-structured sequences, and of Kerr–Shparlinski–Wu–Xi [22] for Type I bilinear forms (where only the sequence (α_m) is smooth).

For moduli with a suitable factorization, the best Type II bounds so far have come from the q-van der Corput method [15], which relies on the twisted multiplicativity of Kloosterman sums, Cauchy–Schwarz, and a shifting trick; it was first applied in this setting by Blomer–Milićević [3]. The q-van der Corput method can also be iterated, leading to strong results for smooth square-free moduli [46, 44]. Unfortunately, these arguments fail to handle certain types of composite moduli when $MN \approx c$, including squares of primes and products of two distinct primes of the same size.

1.2. **Main results.** In this work, we develop a new method to bound bilinear forms with Kloosterman sums for essentially all composite moduli. Like Shkredov [38, 37], we rely on non-abelian Fourier analysis; unlike Shkredov, we use the normal subgroups of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ to our advantage, and we avoid relying on L^2-flattening results, to arrive at quantitatively-good power savings over (1.2) (up to $c^{-1/12}$; see Example 1.3). Our key innovation is a new type of amplification argument with non-abelian characters, detailed in Section 2.3, which may be of independent interest.

Combining our bounds with those of Kowalski–Michel–Sawin[2] [24] (as well as Blomer–Milićević [3] for an optimization), we obtain a non-trivial result for general moduli beyond the Pólya–Vinogradov range, given in Theorem 7.4. We state a particular case of this result below, when $M, N \ll c^{1/2+o(1)}$.

[1]See also [14, Theorem 1.17] for more standard versions of the Pólya–Vinogradov bound.

[2]One could also combine our bounds with the results of Shkredov [38, Theorem 4] for prime moduli, to obtain a result like Theorem 1.1 which does not rely on algebraic geometry.

Theorem 1.1. *Let $c, M, N \in \mathbb{Z}_+$ with $M, N \ll c^{1/2+o(1)}$. Then for any complex sequences $(\alpha_m)_{m \le M}$, $(\beta_n)_{n \le N}$ and $a \in (\mathbb{Z}/c\mathbb{Z})^\times$, one has*

$$\sum_{\substack{m=1 \\ (m,n,c)=1}}^{M} \sum_{n=1}^{N} \alpha_m \beta_n S(am, n; c) \ll \|\alpha\| \|\beta\| c^{1-\frac{1}{700}+o(1)}.$$

Moreover, if $|\alpha_m| \le 1$ for all m (so $\|\alpha\| \le \sqrt{M}$), then

$$\sum_{\substack{m=1 \\ (n,c)=1}}^{M} \sum_{n=1}^{N} \alpha_m \beta_n S(am, n; c) \ll \sqrt{M} \|\beta\| c^{1-\frac{1}{276}+o(1)}.$$

Our main technical result leading to Theorem 1.1 is Theorem 7.1, which considers a factorization of the modulus into three parts, $c = dd'e$ (but one can usually take $d' = 1$ or $e = 1$). Below we state a particular case of Theorem 7.1, focusing on the same range $M, N \ll c^{1/2+o(1)}$ as in Theorem 1.1.

Theorem 1.2. *Let $c = dd'e$ for some $d, d', e \in \mathbb{Z}_+$ with $d' \mid d$ and $(d, e) = 1$, and let f be the largest integer such that $f^2 \mid cd$. Let $\mathcal{I}, \mathcal{J} \subset \mathbb{Z}$ be intervals with $|\mathcal{I}|, |\mathcal{J}| \ll c^{1/2+o(1)}$. Then for any complex sequences $(\alpha_m)_{m \in \mathcal{I}}$, $(\beta_n)_{n \in \mathcal{J}}$ and $a \in (\mathbb{Z}/c\mathbb{Z})^\times$, one has*

$$\sum_{\substack{m \in \mathcal{I}, n \in \mathcal{J} \\ (m,n,c)=1}} \sum \alpha_m \beta_n S(am, n; c) \ll \|\alpha\| \|\beta\| c^{1+o(1)} \left(\frac{f}{\min(c, d^2)} \right)^{\frac{1}{6}}.$$

Since $f \le \sqrt{cd}$, Theorem 1.2 automatically gives a non-trivial result when $d \in (c^{1/3+o(1)}, c^{1-o(1)})$. Unless c has a prime factor larger than $c^{1-o(1)}$, one can always find a factorization $c = dd'e$ with d in this range, $(d, e) = 1$, and $d' \mid d$, which makes the general result in Theorem 1.1 possible.

Example 1.3. *Let $d \mid c$ such that $\frac{c}{d}$ is square-free. Then one can take $d' := (\frac{c}{d}, d)$, $e := \frac{c}{dd'}$, $f = d$ in Theorem 1.2, so the saving over the trivial bound becomes $(d + \frac{c}{d})^{-1/6}$. If $d \asymp \sqrt{c}$, this is roughly*

$$c^{-\frac{1}{12}}.$$

In particular, this saving is achieved if $c \in \{p^2, pq\}$, where p and q are distinct primes with $p \asymp q$. For the same values of c, the more general Theorem 7.1 beats (1.2) in the range

$$M \asymp N \in [c^{\frac{5}{12}+o(1)}, c^{\frac{5}{8}-o(1)}].$$

Notably, while the values $c \in \{p^2, pq\}$, $M, N \asymp \sqrt{c}$ give blind spots of the q-van der Corput method, they happen to give one of the best cases for our methods; this case has until now constituted the remaining barrier towards the application in Theorem 1.5.

As a quick corollary of Theorem 1.2, we prove a trilinear-sum bound which includes a short averaging over c with a given large divisor q; the point is that only a factorization of q (rather than c) is assumed. Such sums arise in the spectral theory of automorphic forms, in particular in Section 9. Below is such a trilinear-sum bound, which is a particular case of Corollary 7.5.

Corollary 1.4. *Let $C \ge \frac{1}{2}$, $q = dd'e$ for some $d, d', e \in \mathbb{Z}_+$ with $d' \mid d$ and $(d, e) = 1$, and let f be the largest integer such that $f^2 \mid qd$. Let $\mathcal{I}, \mathcal{J} \subset \mathbb{Z}$ be intervals of lengths $|\mathcal{I}|, |\mathcal{J}| \ll C^{1/2+o(1)}$. Then or any complex sequences $(\alpha_m)_{m \in \mathcal{I}}$, $(\beta_n)_{n \in \mathcal{J}}$, one has*

$$\sum_{\substack{C < c \le 2C \\ q \mid c}} \left| \sum_{\substack{m \in \mathcal{I}, n \in \mathcal{J} \\ (m,n,q)=1}} \sum \alpha_m \beta_n S(m, n; c) \right| \ll \|\alpha\| \|\beta\| \frac{C^{2+o(1)}}{q} \left(\frac{f}{\min(C, d^2) + \min(q, d^2 \frac{C}{q})} \right)^{\frac{1}{6}}.$$

Remark. Milićević, Qin and Wu [29] have simultaneously and independently obtained results similar to our Theorems 1.1 and 1.5 using substantially different methods. The two papers are complementary, each performing slightly better in different ranges and for different types of moduli, and both achieving power savings for bilinear sums of square-root length and general moduli. The methods in [29] (which use algebraic geometry and build on Kowalski–Michel–Sawin [24] and Blomer–Milićević [3]) obtain better savings for general moduli and remove the dependency on the Ramanujan–Petersson conjecture in the application to moments of twisted cuspidal L-functions. Our methods (which use non-abelian Fourier analysis and are closer to the work of Shkredov [38, 37]) perform better and in longer ranges for specific classes of moduli c (see Examples 1.3 and 7.2), can handle more general supports of the sequences $(\alpha_m), (\beta_n)$ (intervals or other additively-structured subsets of $\mathbb{Z}/c\mathbb{Z}$), and find an application to exceptional-spectrum large sieve inequalities (Corollary 1.6).

Finally, we note that our methods might also lead to bounds for other exponential sums with $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ or $\mathrm{GL}_2(\mathbb{Z}/c\mathbb{Z})$ structure, such as

$$\sideset{}{^*}\sum_{x \in \mathbb{Z}/c\mathbb{Z}} e\left(\frac{mx + nF(x)}{c}\right),$$

where $F(x)$ is a suitable Möbius transformation (and the sum is restricted to those x such that $F(x)$ is well-defined). Related methods for $\mathrm{SL}_3(\mathbb{Z}/c\mathbb{Z})$ could be worth investigating as well.

1.3. **Applications.** As our first application, we prove an asymptotic for the averaged second moment of modular L-functions twisted by primitive Dirichlet characters modulo q, where the modulus q is arbitrary. Blomer–Milićević [1] established such an asymptotic for most moduli, more specifically whenever q is not close to a prime or to a product of two primes of the same size. The missing ingredient in these cases has been precisely a power-saving bound for bilinear forms with Kloosterman sums modulo q, where both sums have length $\approx \sqrt{q}$. The case of prime moduli q was established by Kowalski–Michel–Sawin [24, Theorem 1.5], and the remaining case can now be handled using our Theorem 1.1 (essentially in the setting from Example 1.3).

To state this application, we introduce some quick notation as in [3]. Given $q \in \mathbb{Z}_+$, we write

$$\phi^*(q) := \sum_{d \mid q} \phi(d)\, \mu\left(\frac{q}{d}\right)$$

for the number of primitive characters modulo q; this vanishes if and only if $q \equiv 2 \pmod 4$, and is otherwise of size $q^{1-o(1)}$. We write $\sum^*_{\chi \bmod q}$ for a sum over all primitive characters modulo q. Also following [3], given an L-function $L(s)$, we write $L_q(s)$ for the product $\prod_{p \mid q} L_p(s)$ over all local factors at primes dividing q; thus for example $\zeta_q(s) = \prod_{p \mid q}(1 - p^{-s})^{-1}$.

Theorem 1.5. *Let f_1, f_2 be fixed holomorphic cuspidal newforms for $\mathrm{SL}_2(\mathbb{Z})$ with even weights κ_1, κ_2 and Hecke eigenvalues $\lambda_1(n), \lambda_2(n)$ normalized as in (8.1). Provided that $\kappa_1 \equiv \kappa_2 \pmod 4$, one has the asymptotic*

$$\sideset{}{^*}\sum_{\chi \bmod q} L(\tfrac{1}{2}, f_1 \otimes \chi)\overline{L(\tfrac{1}{2}, f_2 \otimes \chi)} = \frac{2\phi^*(q)}{\zeta(2)} M(f_1, f_2; q) + O_{f_1, f_2}\left(q^{1 - \frac{1}{674} + o(1)}\right),$$

with a main term of

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

where $c(f_1)$ is a constant depending only on f_1 and

$$P(s) := \frac{\zeta_q(2s)}{L_q(s, \mathrm{sym}^2 f_1)}, \qquad Q(s) := \frac{\zeta_q(2s)}{L_q(s, f_1 \times f_2)}.$$

Remark. Similar results can be obtained for Maass cusp forms, with some care in removing the dependency on the Ramanujan conjecture; see also [29]. The case of (non-cuspidal) Eisenstein series reduces to the result of Young [47] on fourth moments of Dirichlet L-functions for prime moduli, extended to all moduli by Wu [45].

Our second application concerns the exceptional spectrum of the hyperbolic Laplacian on $\Gamma_0(q)\backslash\mathbb{H}$, consisting of Maass cusp forms of level $q \in \mathbb{Z}_+$ with eigenvalues $\lambda < \frac{1}{4}$. Selberg's eigenvalue conjecture [34], one of the central open problems in the theory of GL_2 automorphic forms, states that this exceptional spectrum is empty. However, unconditionally, exceptional forms often produce the worst contribution in applications of the Kuznetsov trace formula [11, 27] to analytic number theory problems [12, 43, 7, 32], losing exponential factors in the parameter $\theta := (\frac{1}{4} - \lambda)^{1/2}$. The best known pointwise bound is Kim–Sarnak $\theta \leq \frac{7}{64}$ [23, Appendix 2], but on-average results can also lead to savings in the θ-aspect, sometimes enough to match the conditional results [33, 17]. Following Deshouillers–Iwaniec [11], these on-average results often take the shape of large sieve inequalities for the Fourier coefficients of exceptional Maass forms, incorporating factors of $X^{2\theta}$. While improvements are now possible [32] for exceptional-spectrum large sieve inequalities with special sequences $(\alpha_n)_{n\leq N}$, the savings in the θ-aspect for arbitrary sequences have been limited to $(\frac{q}{N})^{2\theta}$, due to Deshouillers–Iwaniec [11, Theorem 5]. In fact, obtaining *any* power saving for arbitrary sequences when $N \asymp q$ is as hard as proving Selberg's eigenvalue conjecture [32, §2].

In Theorem 9.4, we overcome this barrier at $(\frac{q}{N})^{2\theta}$ if q is suitably-composite and N is not too large, using Theorem 7.1. We note that in many applications [4, 28, 12, 9, 10], the level q is a product of two factors of similar sizes, and $N \in (\sqrt{q}, q)$. We state below a particular case of Theorem 9.4, for $N \approx \sqrt{q}$. We point the reader to Section 9 for more background and notation.

Corollary 1.6. *Let $q \in \mathbb{Z}_+$ have a divisor $d \asymp \sqrt{q}$ such that $\frac{q}{d}$ is square-free. Consider an orthonormal basis of Maass cusp forms for $\Gamma_0(q)$, with Laplacian eigenvalues λ_j and Fourier coefficients $(\rho_j(n))_{n\in\mathbb{Z}}$ around ∞ (normalized as in [11, 32]). Let $N \ll q^{\frac{1}{2}+o(1)}$, and $(\alpha_n)_{N<n\leq 2N}$ be a complex sequence supported on $(n, q) = 1$. Then with $\theta_j := (\frac{1}{4} - \lambda_j)^{1/2} \leq \frac{7}{64}$, one has*

$$\sum_{\lambda_j < \frac{1}{4}} q^{\frac{6}{5}\theta_j} \left| \sum_{N<n\leq 2N} \alpha_n \, \rho_j(n) \right|^2 \ll (qN)^{o(1)} \|\alpha\|^2. \tag{1.3}$$

For reference, [11, Theorem 5] of Deshouillers–Iwaniec would include a factor of $(\frac{q}{N})^{2\theta_j}$ (which is q^{θ_j} when $N = \sqrt{q}$) in the left-hand side, so Corollary 1.6 wins a factor of $q^{\theta/5}$ in this case.

Remark. It follows from the more general Theorem 9.4 that one can relax the condition that $\frac{q}{d}$ is square-free when some averaging over levels $q \leq Q$ with $d \mid q$ (and $d \asymp \sqrt{Q}$) is available. The sequence (α_n) inside the large sieve may depend on q in this case, unlike in [11, Theorem 6].

1.4. **Acknowledgements.** The author is deeply grateful to Valentin Blomer, James Maynard, Sary Drappeau, Philippe Michel, Emmanuel Kowalski, and Ilya Shkredov for helpful comments and discussions. This work was supported by the ERC Advanced Grant 101054336 and Germany's Excellence Strategy grant EXC-2047/1-390685813. For a part of the duration of this project, the author was also supported by an EPSRC Scholarship, as well as a Campus France Scholarship.

2. Outline

2.1. Structure of the paper. Our proof of Theorem 1.2 has three main steps:

I (*Fourier analysis*). In Section 4 (particularly, Proposition 4.10), we relate matrices of Kloosterman sums to the Fourier transform of certain functions at a special representation ρ_c° of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$. This involves Fourier analysis on both abelian ($\mathbb{Z}/c\mathbb{Z}$, $\mathbb{R}$) and non-abelian ($\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$) groups, and a Möbius inversion process for representations of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$.

II (*Amplification*). In Section 5 (particularly, Proposition 5.5), we upper bound the spectral norm of the above Fourier coefficients by a weighted count of solutions to an equation in $\mathrm{PSL}_2(\mathbb{Z}/d\mathbb{Z})$, with $d \mid c$. This is where our non-abelian amplification argument comes in.

III (*Combinatorics*). In Section 6 (particularly, Proposition 6.4), we analyze this counting problem using elementary arguments. In particular, we do not rely on expansion techniques.

In Section 7, we combine these ingredients to deduce Theorem 1.2 and its variations. The applications to moments of twisted cuspidal L-functions and large sieve inequalities for exceptional cusp forms are handled in Sections 8 and 9, respectively.

For the rest of this section, we give a brief informal overview of our argument, ignoring various technical details. We will use the symbols '$\approx$', '$\lesssim$' for identities and inequalities that are 'morally' true (and can be made rigorous with minor modifications, such as including $c^{o(1)}$ factors).

2.2. First steps: Fourier analysis. Let us focus on the balanced case $M = N$. We begin by considering the $N \times N$ complex matrix

$$K := \left(S(m, n; c)\right)_{m,n \le N}.$$

where $c, N \in \mathbb{Z}_+$ with $N \le c$. Our task is to bound the operator norm $\|K\|$ by less than $\min(c, N\sqrt{c})$, to beat (1.2). We extend K to a $c \times c$ matrix, and multiply it on both sides by the unitary matrix $(\frac{1}{\sqrt{c}} e(\frac{xy}{c}))_{x,y \in \mathbb{Z}/c\mathbb{Z}}$, which preserves the operator norm and essentially amounts to taking a Fourier transform in the m, n variables. Letting $H \approx \frac{c}{N}$, a truncated version of Poisson summation yields

$$c\, U^* K U^* \approx \frac{1}{H^2} \left(\sum_{|h| \le H} \mathcal{T}^h \right) \mathcal{S} \left(\sum_{|h| \le H} \mathcal{T}^h \right), \qquad \text{where} \qquad \begin{cases} \mathcal{T} := (\mathbb{1}_{u = x+1})_{u,x \in \mathbb{Z}/c\mathbb{Z}}, \\ \mathcal{S} := (\mathbb{1}_{xy = -1})_{x,y \in \mathbb{Z}/c\mathbb{Z}}. \end{cases}$$

By inserting a few more rows and columns, we can in fact work over the projective line $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ rather than $\mathbb{Z}/c\mathbb{Z}$. The matrices $\mathcal{T}$ and $\mathcal{S}$ then extend to $\rho_c(T)$ and $\rho_c(S)$, where T and S are the usual generators of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ (see (3.13)), and

$$\rho_c : \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \to \left\{ \text{Unitary maps of } \mathbb{C}^{\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})} \right\}$$

is the $c^{1+o(1)}$-dimensional permutation representation corresponding to the action of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ on $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ by Möbius transformations. It then remains to bound the spectral norm of the matrix

$$\frac{1}{H^2} \sum_{|h_1|, |h_2| \le H} \rho_c(T^{h_1} S T^{h_2})$$

by less than $\min(1, N/\sqrt{c})$. In this form, our task is actually impossible: the matrix above decomposes as a direct sum corresponding to the irreducible representations inside ρ_c, one of which is the trivial representation—and this contributes exactly one singular value of size 1. Other small-dimensional subrepresentations of ρ_c are also problematic for similar reasons.

This is where the coprimality constraint $(m,n,c)=1$ comes in. Incorporating this weight into the matrix K and expanding it by Möbius inversion ultimately results in a 'sifted' representation,

$$K^\circ := (S(m,n;c)\mathbb{1}_{(m,n,c)=1})_{m\le M, n\le N} \qquad \rightsquigarrow \qquad \rho_c^\circ : \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}),$$

where ρ_c° is essentially obtained by removing from ρ_c the contribution of all subrepresentations isomorphic to

$$\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \xrightarrow{\text{Reduction mod } d} \mathrm{SL}_2(\mathbb{Z}/d\mathbb{Z}) \xrightarrow{\rho_d} \left\{\text{Unitary maps of } \mathbb{C}^{\mathbb{P}^1(\mathbb{Z}/d\mathbb{Z})}\right\},$$

for $d \mid c$. Although ρ_c° is not irreducible in general, it has the key property that all of its $c^{o(1)}$ irreducible subrepresentations are large, of dimension $c^{1-o(1)}$ (see Proposition 4.6).

2.3. **The key step: Amplification.** We are left to bound the spectral norm of the non-abelian Fourier coefficient

$$\widehat{F}(\rho) = \sum_{g\in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})} F(g)\rho(g), \qquad F := \frac{1}{H^2}\sum_{|h_1|,|h_2|\le H} \mathbb{1}_{T^{h_1}ST^{h_2}} : \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \to \mathbb{C},$$

where ρ is any irreducible subrepresentation of ρ_c°. A natural approach is to use the trace method, i.e., to bound the top singular value $\|\widehat{F}(\rho_c^\circ)\|$ by an even moment of all singular values, and then to expand the latter as a trace; this brings in the character $\chi := \mathrm{Tr}\rho$.

One can then attempt to use non-abelian Fourier analysis by summing over all irreducible characters χ' of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$. However, this sum must somehow amplify the contribution of $\chi' = \chi$ compared to other irreducible characters of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$, especially the small-dimensional ones—otherwise, our construction of ρ_c° by eliminating various subrepresentations from ρ_c will have been useless.

If χ was an abelian character of $(\mathbb{Z}/c\mathbb{Z})^\times$, i.e., a Dirichlet character, then following the ideas of Duke–Friendlander–Iwaniec [13], one could weigh the sum by an amplifier of the shape

$$A(\chi') := \left|\sum_{\ell\in\mathcal{L}} \overline{\chi}'(\ell)\chi(\ell)\right|^2 = \sum_{\ell_1,\ell_2\in\mathcal{L}} \overline{\chi}'(\ell_1\ell_2^{-1})\chi(\ell_1\ell_2^{-1}), \qquad \chi \in \widehat{(\mathbb{Z}/c\mathbb{Z})^\times},$$

where $\mathcal{L}$ is some set of positive integers (e.g., the primes in a dyadic interval). This $A(\chi')$ has size $\approx |\mathcal{L}|^2$ when $\chi' = \chi$, and should typically obey square-root cancellation when $\chi' \neq \chi$.

Inspired by this, we construct a general amplifier for irreducible representations of a finite non-abelian group G—which is to the best of our knowledge the first instance of such a construction, and which might find applications to other problems. We set

$$A(\chi') := \left\|\sum_{\ell\in\mathcal{L}} \overline{\rho'(\ell)} \otimes \rho(\ell)\right\|_{S^2}^2 = \sum_{\ell_1,\ell_2\in\mathcal{L}} \overline{\chi}'(\ell_1\ell_2^{-1})\chi(\ell_1\ell_2^{-1}), \qquad \rho' \in \widehat{G},\ \chi' := \mathrm{Tr}\rho',$$

where $\|\cdot\|_{S^2}$ denotes the Frobenius norm of a map (which is the ℓ^2 norm of its singular values), and $\mathcal{L}$ is a well-chosen subset of G. It is most convenient to pick $\mathcal{L}$ to be a normal subgroup of G; note that when $G = \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$, this is only possible if c is composite. We will in fact pick

$$\mathcal{L} = \Gamma_c(d) := \ker\left(\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/d\mathbb{Z})\right),$$

for a suitable divisor d of c. The result of this amplification argument for the sixth moment of singular values is a bound of the shape (see Proposition 5.1)

$$\|\widehat{F}(\rho)\|^6 \lesssim \frac{c^3H^{-6}}{\sum_{\ell\in\Gamma_c(d)}|\chi(\ell)|^2} \sum_{\substack{|h_1|,\dots,|h_6|\le H \\ T^{h_1}S\cdots T^{h_6}S\in\Gamma_c(d)}} \chi(T^{h_1}S\cdots T^{h_6}S). \tag{2.1}$$

To go any further, we need to know the typical size of the character χ on $\Gamma_c(d)$, based on the information that $\dim \chi \gg c^{1-o(1)}$. This is a somewhat challenging computation involving Clifford theory, and depends on the factorizations of c and d; see Lemmas 5.2 and 5.4.

Let us now focus on the case when $c = p^2$ is the square of a prime and $N \approx H \approx p$; we naturally pick $d = p$. It turns out that χ typically has size $\approx p$ on $\Gamma_{p^2}(p)$, and roughly $\approx p^2$ at $\pm I \in \mathrm{SL}_2(\mathbb{Z}/p^2\mathbb{Z})$. To beat (1.2), it essentially remains to bound

$$\sum_{|h_1|,\ldots,|h_6|\le p} \mathbb{1}_{T^{h_1}S\cdots T^{h_6}S\equiv \pm I \pmod{p^2}} \overset{?}{<} p^4, \qquad \sum_{|h_1|,\ldots,|h_6|\le p} \mathbb{1}_{T^{h_1}S\cdots T^{h_6}S\equiv \pm I \pmod{p}} \overset{?}{<} p^5. \tag{2.2}$$

2.4. **Final steps: Combinatorics.** The estimates in (2.2) amount to counting the number of solutions to the system of congruences

$$\begin{cases} 1 - h_2h_3 \equiv \mp(1 - h_5h_6) \\ h_1(1-h_2h_3) + h_3 \equiv \pm h_5 \\ h_4(1-h_2h_3) + h_2 \equiv \pm h_6 \end{cases} \quad (\mathrm{mod}\ p^2,\text{ respectively, } p),$$

with $|h_1|, \ldots, |h_6| \le p$. For the generic solutions, one can expect each congruence to cut down the total number of solutions p^6 by the size of the modulus—but one must also account for certain diagonal solutions where some $h_i = 0$. A careful but elementary analysis (which becomes more involved when the modulus c is arbitrary) shows that these congruences have $\approx p^2$ solutions modulo p^2 and $\approx p^3$ solutions modulo p; see Proposition 6.4. Both of these counts are sharp, and save a factor of p over the bounds required in (2.2). This saving is ultimately raised to the power $\frac{1}{6}$ in (2.1) (since we considered a sixth moment of singular values), and putting everything together yields

$$\left\| \left(S(m,n;p^2)\right)_{m,n\le p} \mathbb{1}_{(m,n,p)=1} \right\| \lesssim p^{2-\frac{1}{6}},$$

as in Example 1.3. We note that for the other case $c = pq$ from Example 1.3, one can simplify the amplification argument by noting that all irreducible characters of $\mathrm{SL}_2(\mathbb{Z}/pq\mathbb{Z})$ are tensor products of irreducible characters of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ and $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$, but the end result is the same.

2.5. **Comments on prime moduli.** When $c = p$ is a prime and $d = p$, the amplifier from Section 2.3 reduces to the 'trivial' choice

$$A(\chi') = \overline{\chi}'(I)\chi(I) = \dim \chi' \dim \chi,$$

since $\mathcal{L} = \Gamma_c(c) = \{I\}$. In this setting, (2.1) (with 6 replaced by another even integer q) reads

$$\|\widehat{F}(\rho)\|^q \lesssim p^2 H^{-q} \sum_{|h_1|,\ldots,|h_q|\le H} \mathbb{1}_{T^{h_1}S\cdots T^{h_q}S = I}.$$

This was observed, using a somewhat different language, by Shkredov [37, proofs of Lemmas 22 and 53]. Shkredov then relied on an L^2-flattening lemma [37, Theorem 50], which stems from a result of Helfgott [19], to bound the right-hand side above by $O(p^2 H^{-q} H^q p^{-3}) = O(p^{-1})$ for a quite large value of q depending on $\frac{\log p}{\log H}$. This leads to a bound for bilinear (Type II) sums of Kloosterman sums with prime moduli [38, (6) of Theorem 4], with a power saving of $p^{-\delta}$, where $\delta \approx \frac{1}{q}$.

To obtain a more quantitatively-relevant power saving, competitive with [24, 25], one must use a smaller value of q; one would then need to solve a counting problem with few variables $h_1, \ldots, h_q$, as in Section 2.4. We do not know how to do this, but such an approach could produce good results when, e.g., $q \in \{8, 10, 12\}$. In particular, assuming Conjecture 6.2 for $q = 8$, one could prove a non-trivial bound for bilinear sums of Kloosterman sums with prime moduli p and sequences of lengths $M \asymp N > p^{3/8+o(1)}$; interestingly, the same limit at $p^{3/8+o(1)}$ appears in the results of Kowalski–Michel–Sawin [25], so our work reaffirms the difficulty of this barrier.

Alternatively, to obtain non-trivial results at prime moduli, it might be possible to use a different choice of subset $\mathcal{L} \subset \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ in the construction of the amplifier from Section 2.4. Indeed, although a normal subgroup is the most natural choice for $\mathcal{L}$, it is possible that another conjugation-invariant subset might produce a useful amplifier when normal subgroups are not available.

3. Preliminaries

3.1. **Analytic and arithmetic notation.** We use the standard asymptotic notation from analytic number theory, indicating dependencies of implicit constants on a parameter ε through subscripts. In particular, $f \ll_\varepsilon g$ and $f = O_\varepsilon(g)$ both mean $|f| \le C_\varepsilon g$ for some constant $C_\varepsilon > 0$ depending only on ε; $f \asymp_\varepsilon g$ means $f \ll_\varepsilon g \ll_\varepsilon f$, $f = \Omega_\varepsilon(g)$ means $f \gg_\varepsilon g$; $f(x) = o(g(x))$ means $\frac{f(x)}{g(x)} \to 0$ as $x \to \infty$; $f(x) \ll x^{o(1)} g(x)$ is equivalent to the statement that $f(x) \ll_\varepsilon x^\varepsilon g(x)$ for all $\varepsilon > 0$. With this notation, the divisor bound reads $\sum_{d|c} 1 \ll c^{o(1)}$.

We use the notation $\mathbb{1}_S$ for both indicator functions of sets S and truth values (0 or 1) of statements S; we also abbreviate $\mathbb{1}_x := \mathbb{1}_{\{x\}}$ for singletons. We write $n \sim N$ for the range $N < n \le 2N$, $\|\alpha\| := (\sum_n |\alpha_n|^2)^{1/2}$ for the ℓ^2 norm of a sequence $(\alpha_n)_{n \in \mathcal{N}}$ for some $\mathcal{N} \subset \mathbb{Z}$ (or $\mathcal{N} \subset \mathbb{Z}/c\mathbb{Z}$), and $e(t) := \exp(2\pi i t)$ for $t \in \mathbb{R}/\mathbb{Z}$. Given a positive integer c, we let $c\mathbb{Z}$ (resp., $c\mathbb{Z}_+$) be the sets of integers (resp., positive integers) divisible by c, and $\overline{x}$ be the inverse of x modulo c (here c may be implied from context, e.g., in an exponential phase $e(\overline{x}/c)$). We use μ and ϕ be the Möbius and Euler totient functions. Given $a, b \in \mathbb{Z}_+$, we write (a, b) for their greatest common divisor (and similarly for more positive integers), and (a, b^∞) for the greatest divisor of a whose prime factors all divide b. We write $p^k \| c$ when a prime power exactly divides a positive integer, meaning that $p^k \mid c$ but $p^{k+1} \nmid c$. We will reserve the letter ψ for functions on $\mathbb{Z}/c\mathbb{Z}$, and Φ, Ψ for functions on $\mathbb{R}$. We denote the Fourier transform of an L^1 function $\Phi : \mathbb{R} \to \mathbb{C}$ by

$$\widehat{\Phi}(\xi) := \int_{-\infty}^{\infty} \Phi(t)\, e(-t\xi)\, dt. \tag{3.1}$$

In particular, if $\Psi(t) := \Phi(At)e(Bt)$ for some $A > 0$ and $B \in \mathbb{R}$, then a change of variables yields

$$\widehat{\Psi}(\xi) = \int_{-\infty}^{\infty} \Phi(At)\, e(-t(\xi - B))\, dt = \frac{1}{A} \widehat{\Phi}\left(\frac{\xi - B}{A}\right), \tag{3.2}$$

and the Poisson summation identity reads

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

Given a map M between finite-dimensional complex Hilbert spaces, we write its operator norm as

$$\|M\| := \sup_{\|\vec{v}\|=1} \|M\vec{v}\| = \sup_{\|\vec{v}\|=\|\vec{w}\|=1} |\vec{w}^T M \vec{v}|. \tag{3.4}$$

On the Hilbert space $\mathbb{C}^n$ equipped with the Euclidean norm, we define $\|M\|_{S^q}$ as the ℓ^q norm of singular values of a map (or a matrix) M, for $q \in [1, \infty]$. In particular, we have

$$\|M\|_{S^\infty} = \|M\| \qquad \text{and} \qquad \|M\|_{S^q}^q = \mathrm{Tr}\left((MM^*)^{q/2}\right)^{\frac{1}{q}} \ \text{ for } q \in 2\mathbb{Z}_+, \tag{3.5}$$

where M^* denotes the adjoint (conjugate transpose) of M. We quickly record the following simple fact about projections and operator norms.

Lemma 3.1. *Let V be a finite-dimensional complex Hilbert space, $W \subset V$ be a subspace, and $P_W : V \to V$ be the orthogonal projection onto W. Suppose that W is an invariant subspace of a linear map $M : V \to V$ (i.e., the restriction $M|_W : W \to W$ is well-defined). Then $\|M|_W\| = \|MP_W\|$.*

Proof. By definition, we have $\|M|_W\| = \sup_{\vec{w}\in W, \|\vec{w}\|=1} \|M\vec{w}\|$ and $\|MP_W\| = \sup_{\vec{v}\in V, \|\vec{v}\|=1} \|MP_W\vec{v}\|$. Since $P_W\vec{v} \in W$ with $\|P_W\vec{v}\| \le \|\vec{v}\| = 1$ for all $\vec{v} \in V$ with $\|\vec{v}\| = 1$, we have $\|MP_W\| \le \|M|_W\|$. On the other hand, for each $\vec{w} \in W$ with $\|\vec{w}\| = 1$, we have $P_W\vec{w} = \vec{w}$, so $\|M|_W\| \le \|MP_W\|$. □

3.2. **Bounds for Kloosterman sums.** We now recall the Ramanujan and Weil bounds for Kloosterman sums, as well as some results of Kowalski–Michel–Sawin [24] and Blomer–Milićević [3].

Lemma 3.2 (Ramanujan bound). *For $c \in \mathbb{Z}_+$ and $n \in \mathbb{Z}$, one has*

$$|S(0, n; c)| \le (n, c).$$

Proof. This is a classical result which follows from Möbius inversion. □

Lemma 3.3 (Weil bound). *For $c \in \mathbb{Z}_+$ and $m, n \in \mathbb{Z}$, one has*

$$S(m, n; c) \ll c^{o(1)} \sqrt{(m, n, c)c}.$$

Proof. This is [21, Corollary 11.12] followed by the divisor bound. □

For the sake of completeness, we give a quick proof of the trivial bound from (1.2).

Proof of (1.2). The second bound implicit in (1.2), with a term of $\sqrt{MNc}$, follows immediately from Lemma 3.3 and Cauchy–Schwarz. For the first bound implicit in (1.2), we eliminate the constraint $(m, n, c) = 1$ by Möbius inversion and use the identity $S(dm, dn; c) = \frac{\phi(c)}{\phi(c/d)} S(m, n; \frac{c}{d})$ to write

$$\sum_{\substack{m\in\mathcal{I}, n\in\mathcal{J} \\ (m,n,c)=1}} \sum \alpha_m \beta_n S(am, n; c) \ll c^{o(1)} \max_{d|c} d \left| \sum_{dm\in\mathcal{I}} \alpha_{dm} \sum_{dn\in\mathcal{J}} \beta_{dn} S(am, n; \tfrac{c}{d}) \right|.$$

Now apply Cauchy–Schwarz in the sum over m, and complete the sum over $m \pmod{\frac{c}{d}}$ to get

$$d \left| \sum_{dm\in\mathcal{I}} \alpha_{dm} \sum_{dn\in\mathcal{J}} \beta_{dn} S(am, n; \tfrac{c}{d}) \right| \le \|\alpha\| \left(d^2 \sum_{m \pmod{\frac{c}{d}}} \left| \sum_{dn\in\mathcal{J}} \beta_{dn} S(am, n; \tfrac{c}{d}) \right|^2 \right)^{\frac{1}{2}}.$$

Expanding the square and the Kloosterman sums, then performing the sum over m, one reaches

$$d^2 \sum_{m \pmod{\frac{c}{d}}} \left| \sum_{dn\in\mathcal{J}} \beta_{dn} S(am, n; \tfrac{c}{d}) \right|^2 = dc \sum_{x\in(\mathbb{Z}/\frac{c}{d}\mathbb{Z})^\times} \left| \sum_{dn\in\mathcal{J}} \beta_{dn} e\left(\frac{nx}{c/d}\right) \right|^2.$$

Finally, complete the sum over $x \pmod{\frac{c}{d}}$, expand the square, and perform the sum over x to obtain

$$dc \sum_{x\in(\mathbb{Z}/\frac{c}{d}\mathbb{Z})^\times} \left| \sum_{dn\in\mathcal{J}} \beta_{dn} e\left(\frac{nx}{c/d}\right) \right|^2 \le c^2 \|\beta\|^2.$$

Putting these bounds together completes our proof. □

Theorem 3.4 (Kowalski–Michel–Sawin [24]). *Let p be a prime and $M, N \in \mathbb{Z}$ be such that $1 \le N \le M \le p - 1$ and $p^{1/4} < MN < p^{5/4}$. Then for any complex sequences $(\alpha_m)_{m\le M}$, $(\beta_n)_{n\le N}$ and any $a \in (\mathbb{Z}/p\mathbb{Z})^\times$, one has*

$$\sum_{m=1}^{M} \sum_{n=1}^{N} \alpha_m \beta_n S(am, n; p) \ll \|\alpha\|\|\beta\| p^{o(1)} \sqrt{MNp} \left(N^{-\frac{1}{2}} + (MN)^{-\frac{3}{16}} p^{\frac{11}{64}} \right).$$

Proof. This is [24, Theorem 1.1] with $k = 2$ and M, N swapped. □

Remark. The constraint $p^{1/4} < MN < p^{5/4}$ from Theorem 3.4 can be removed in light of the trivial bound (1.2). Indeed, if $MN \leq p^{1/4}$, then

$$\sqrt{MNp} \cdot (MN)^{-\frac{3}{16}} \cdot p^{\frac{11}{64}} \geq \sqrt{MNp} \cdot p^{\frac{11}{64} - \frac{3}{64}} > \sqrt{MNp},$$

so the bound (1.2) is better. Similarly, if $MN \geq p^{5/4}$, then

$$\sqrt{MNp} \cdot (MN)^{-\frac{3}{16}} \cdot p^{\frac{11}{64}} \geq p^{\frac{5}{4}(\frac{1}{2} - \frac{3}{16})} \cdot p^{\frac{1}{2} + \frac{11}{64}} = p^{\frac{25}{64} + \frac{43}{64}} > p.$$

Theorem 3.5 (Blomer–Milićević [3])**.** *Let $c, d, M, N \in \mathbb{Z}_+$ such that $d \mid c$ and d is odd. Then for any complex sequences $(\alpha_m)_{m \leq M}$ and $(\beta_n)_{n \leq N}$ such that $|\alpha_m| \leq 1$ for all m, and any $a \in (\mathbb{Z}/c\mathbb{Z})^\times$, one has*

$$\sum_{m=1}^{M} \sum_{\substack{n=1 \\ (n,c)=1}}^{N} \alpha_m \beta_n S(am, n; c) \ll \sqrt{M} \|\beta\| (MNc)^{\frac{1}{2} + o(1)} \left(\frac{c^{1/2}}{d^{1/2} M^{1/2}} + \frac{1}{d^{1/4}} + \frac{d^{1/4}}{N^{1/2}} \right).$$

Proof. Dyadically summing instances of [3, Theorem 5] with $(q, r, s, M, K, \lambda(k))$ in loc. cit. replaced by $(c, c, \frac{c}{d}, N, M, \alpha_m)$, one obtains the bound[3]

$$\sum_{\substack{n \leq N \\ (n,c)=1}} \left| \sum_{m \leq M} \alpha_m S(am, n; c) \right|^2 \ll (cMN)^{o(1)} M^2 N c \left(\frac{c}{dM} + \frac{1}{\sqrt{d}} + \frac{\sqrt{d}}{N} \right).$$

The desired bound now follows from Cauchy–Schwarz in the shape

$$\left| \sum_{m=1}^{M} \sum_{\substack{n=1 \\ (n,c)=1}}^{N} \alpha_m \beta_n S(am, n; c) \right|^2 \leq \|\beta\|^2 \sum_{\substack{n \leq N \\ (n,c)=1}} \left| \sum_{m \leq M} \alpha_m S(am, n; c) \right|^2.$$

(Since (β_n) can be chosen to attain equality in this Cauchy–Schwarz step, Theorem 3.5 is in fact a restatement of [3, Theorem 5].) □

3.3. **Fourier analysis on finite groups.** Here we recall some general facts and notation from representation theory on finite groups; we point the reader to [35, 16, 42, 20] for more background. Let G be a finite group with identity element e. A (unitary) representation of G is a homomorphism

$$\rho : G \to U(V),$$

where V is a finite-dimensional complex Hilbert space and $U(V)$ is the set of unitary transformations of V. In particular, $\rho(e) = \mathrm{Id}_V$ is the identity transformation on V. We write[4]

$$\dim \rho := \dim V$$

for the dimension of ρ. We say that two representations $\rho_1 : G \to U(V_1)$, $\rho_2 : G \to U(V_2)$ are *isomorphic* iff there is an invertible linear map $M : V_1 \to V_2$ such that $M \circ \rho_1(g) = \rho_2(g) \circ M$ for all $g \in G$ (since we normalize all representations to be unitary, the map M can also be taken unitary).

[3] [3, Theorem 5] does not include an a-scalar inside the Kloosterman sum, but it holds in this slightly more general form with the same proof, and it is in fact applied this way in [3, p. 471, after (4.2)].

[4] Given a choice of orthonormal basis of $V \cong \mathbb{C}^{\dim \rho}$, one may of course represent the transformations $\rho(g)$ for $g \in G$ as matrices in $\mathbb{C}^{\dim \rho \times \dim \rho}$.

Example 3.6. *We write* $\mathbf{0} : G \to U(\{0\})$ *for the* zero representation *given by* $\mathbf{0}(g) = 0\ \forall g \in G$*, and* $\mathbf{1} : G \to U(\mathbb{C})$ *for the* trivial representation *given by* $\mathbf{1}(g) = \mathrm{Id}_{\mathbb{C}}\ \forall g \in G$*. Any action of* G *on a finite set* X *induces a* permutation representation $\rho : G \to U(\mathbb{C}^X)$*, defined by* $(\rho(g)f)(x) := f(g^{-1}x)$ *for* $g \in G$*,* $x \in X$*. The* regular representation R_G *is the permutation representation induced by the action by left-multiplication on* $X = G$*, so* $\dim R_G = |G|$*.*

Given two representations $\rho_1 : G \to U(V_1)$ and $\rho_2 : G \to U(V_2)$, we write $\rho_1 \oplus \rho_2 : G \to U(V_1 \oplus V_2)$ and $\rho_1 \otimes \rho_2 : G \to U(V_1 \otimes V_2)$ for their direct sum and product. The operations $\oplus$ and $\otimes$ have identity elements $\mathbf{0}$ and $\mathbf{1}$ respectively (up to isomorphism). Given $\rho : G \to U(V)$, we write

$$\rho^{\oplus m} := \underbrace{\rho \oplus \cdots \oplus \rho}_{m \text{ times}}$$

for all nonnegative integers m; when $m = 0$, we interpret this as the zero representation $\mathbf{0}$. We use a similar notation for repeated direct sums of linear maps.

An *invariant subspace* W of a representation $\rho : G \to U(V)$ is a subspace of V such that $\rho(g)W \subset W$ for all $g \in G$. For such W, we define $\rho|_W : G \to U(W)$ by $\rho|_W(g) := \rho(g)$ for all $g \in G$, which is automatically unitary, and we say that $\rho|_W$ is a *subrepresentation* of ρ. One can decompose $\rho \cong \rho_W \oplus \rho_{W^\perp}$; conversely, if $\rho \cong \rho_1 \oplus \rho_2$ then ρ_1 and ρ_2 are isomorphic to subrepresentations of ρ.

We say that a representation of G is *irreducible* iff it is nonzero and has no nonzero subrepresentations other than itself. We write $\widehat{G}$ for a complete set of irreducible representations of G up to isomorphism, which always includes the trivial representation $\mathbf{1}$. Any representation ρ of G has a unique decomposition (up to permutation and isomorphism) into irreducible representations,

$$\rho \cong \bigoplus_{\rho' \in \widehat{G}} \rho'^{\oplus \mathrm{Mult}(\rho',\rho)}, \tag{3.6}$$

where $\mathrm{Mult}(\rho', \rho)$ is called the *multiplicity* of ρ' inside ρ. In particular, $\mathrm{Mult}(\rho', R_G) = \dim \rho'$.

Given two finite groups G_1, G_2 and representations $\rho_1 : G_1 \to U(V_1)$ and $\rho_2 : G_2 \to U(V_2)$, we write $\rho_1 \boxtimes \rho_2 : G_1 \times G_2 \to U(V_1 \otimes V_2)$ for the representation of $G_1 \times G_2$ given by

$$(\rho_1 \boxtimes \rho_2)(g_1, g_2) := \rho_1(g_1) \otimes \rho_2(g_2), \qquad g_1 \in G_1,\ g_2 \in G_2.$$

The irreducible representations of $G_1 \times G_2$ are (up to isomorphism) precisely those of the form $\rho_1 \boxtimes \rho_2$ where $\rho_1 \in \widehat{G}_1$ and $\rho_2 \in \widehat{G}_2$ [35, §3.2].

Notation 3.7. If G_1, G_2, ρ_1, ρ_2 are as above, and $G_{1,2}$ is a group isomorphic to $G_1 \times G_2$ by a fixed implicit map (such as (3.16)), we also use the notation $\rho_1 \boxtimes \rho_2$ to describe representations of $G_{1,2}$.

A *character* $\chi : G \to \mathbb{C}$ is any function of the form $\chi(g) = \mathrm{Tr}\rho(g)$, where ρ is a representation of G; note that characters are constant on conjugacy classes, that $\chi(e) = \dim \rho$ and $\chi(g^{-1}) = \overline{\chi}(g)$, and that isomorphic representations induce the same character. If ρ_1, ρ_2 are two representations of G with characters χ_1, χ_2, then $\mathrm{Tr}(\rho_1 \oplus \rho_2) = \chi_1 + \chi_2$ and $\mathrm{Tr}(\rho_1 \otimes \rho_2) = \chi_1\chi_2$. If ρ_1, ρ_2 are representations of G_1, G_2 with characters χ_1, χ_2 (respectively), then $\mathrm{Tr}(\rho_1 \boxtimes \rho_2)(g_1, g_2) = \chi_1(g_1)\chi_2(g_2)$. We say that χ is irreducible iff ρ is, and write $\mathrm{Irr}(G)$ for the set of all irreducible characters of G. The character table of G satisfies the following orthogonality relations.

Lemma 3.8 (Character orthogonality)**.** *One has*

$$\sum_{g \in G} \chi_1(g)\overline{\chi}_2(g) = |G|\mathbb{1}_{\chi_1 = \chi_2}, \qquad \chi_1, \chi_2 \in \mathrm{Irr}(G), \tag{3.7}$$

$$\sum_{\chi \in \mathrm{Irr}(G)} \chi(g_1)\overline{\chi}(g_2) = \begin{cases} \frac{|G|}{|C|}, & g_1, g_2 \textit{ belong to the same conjugacy class } C \textit{ of } G, \\ 0, & g_1, g_2 \in G \textit{ are not conjugate.} \end{cases} \tag{3.8}$$

Proof. See, e.g., [16, Theorem 2.12 and Exercise 2.21]. □

It follows from (3.6) and (3.7) that for an arbitrary character $\chi = \mathrm{Tr}\rho$ of G, one has

$$\frac{1}{|G|}\sum_{g\in G}|\chi(g)|^2 = \sum_{\rho'\in\widehat{G}}\mathrm{Mult}(\rho',\rho)^2. \tag{3.9}$$

We may also restrict a representation $\rho : G \to U(V)$ and its character $\chi = \mathrm{Tr}\rho$ to a subgroup $H \leq G$, to obtain a representation of $\rho|_H : H \to U(V)$ with character $\chi|_H = \mathrm{Tr}\rho|_H$. If ρ is irreducible, $\rho|_H$ is not necessarily irreducible. When H is a normal subgroup, the structure of $\rho|_H$ can be better understood using Clifford theory [6].

Lemma 3.9 (Clifford)**.** *Let G be a group, $N \lhd G$ be a normal subgroup, and $\rho \in \widehat{G}$ be an irreducible representation. Then there exist positive integers L, m, d with $\dim\rho = Lmd$, and non-isomorphic irreducible representations $\sigma_1, \ldots, \sigma_L \in \widehat{N}$ of dimension d, all lying in the same orbit of the action of G by conjugation (i.e., $(g\cdot\sigma)(n) := \sigma(gng^{-1})$ for $g \in G$ and $n \in N$), such that*

$$\rho|_N \cong \bigoplus_{\ell=1}^{L}\sigma_\ell\oplus^m.$$

Proof. See, e.g., [20, Theorem 6.5]. □

Given a function $F : G \to \mathbb{C}$ and a (not necessarily irreducible) representation $\rho : G \to U(V)$, we define the *Fourier coefficient* $\widehat{F}(\rho) : V \to V$ by

$$\widehat{F}(\rho) := \sum_{g\in G}F(g)\rho(g). \tag{3.10}$$

This obeys $\widehat{F_1 * F_2}(\rho) = \widehat{F_1}(\rho)\widehat{F_2}(\rho)$, where $(F_1 * F_2)(g) := \sum_{g_1g_2=g}F_1(g)F_2(g)$ denotes the convolution of two functions $F_1, F_2 : G \to \mathbb{C}$. In particular, if $G = \mathbb{Z}/c\mathbb{Z}$, the irreducible representations (which are all 1-dimensional since G is abelian) are of the shape $\rho_a(g) := e(\frac{ag}{c})$ for $a, g \in \mathbb{Z}/c\mathbb{Z}$. In this case, we write

$$\widehat{F}(a) := \widehat{F}(\rho_{-a}) = \sum_{g\in\mathbb{Z}/c\mathbb{Z}}F(g)\,e\left(-\frac{ag}{c}\right). \tag{3.11}$$

Lemma 3.10. *Let $F : G \to \mathbb{C}$, $\rho : G \to U(V)$ be a representation, and $q \in [1,\infty)$. Then one has*

$$\|\widehat{F}(\rho)\|_{S^q}^q = \sum_{\rho'\in\widehat{G}}\mathrm{Mult}(\rho',\rho)\|\widehat{F}(\rho')\|_{S^q}^q, \qquad \|\widehat{F}(\rho)\| = \max_{\substack{\rho'\in\widehat{G}\\ \mathrm{Mult}(\rho',\rho)>0}}\|\widehat{F}(\rho')\|.$$

Proof. By (3.6), there exists a unitary map U (from V to the direct sum of ρ's irreducible invariant subspaces) such that for any $g \in G$,

$$U\rho(g)U^* = \bigoplus_{\rho'\in\widehat{G}}\rho'(g)\oplus^{\mathrm{Mult}(\rho',\rho)},$$

using some implicit ordering of $\widehat{G}$. But then, by (3.10), we have

$$\begin{aligned}U\widehat{F}(\rho)U^* = \sum_{g\in G}F(g)U\rho(g)U^* &= \sum_{g\in G}F(g)\bigoplus_{\rho'\in\widehat{G}}\rho'(g)\oplus^{\mathrm{Mult}(\rho',\rho)}\\ &= \bigoplus_{\rho'\in\widehat{G}}\left(\sum_{g\in G}F(g)\rho'(g)\right)\oplus^{\mathrm{Mult}(\rho',\rho)} = \bigoplus_{\rho'\in\widehat{G}}\widehat{F}(\rho')\oplus^{\mathrm{Mult}(\rho',\rho)},\end{aligned}$$

and the conclusion follows from the fact that the multiset of singular values of a direct sum of matrices is the union of the multisets of singular values of those matrices. □

3.4. **Facts about** $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$**.** Let $c \in \mathbb{Z}_+$. Recall the special linear groups $\mathrm{SL}_2(\mathbb{Z})$ and $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ of matrices in $\mathbb{Z}^{2\times 2}$ (resp., $(\mathbb{Z}/c\mathbb{Z})^{2\times 2}$) with determinant 1, and the projective special linear groups,

$$\mathrm{PSL}_2(\mathbb{Z}) := \mathrm{SL}_2(\mathbb{Z})/\{\pm I\}, \qquad \mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z}) := \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})/_{\{\gamma I : \gamma \in \mathbb{Z}/c\mathbb{Z}, \gamma^2 = 1\}}. \tag{3.12}$$

When the group $\mathrm{SL}_2(\mathbb{Z})$, $\mathrm{PSL}_2(\mathbb{Z})$, $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ or $\mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})$ is understood from context, we write

$$I := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad T := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad S := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{3.13}$$

which satisfy the relations $-S^2 = -(ST)^3 = I$, and in the case of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ or $\mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})$, $T^c = I$. Note that T and S generate $\mathrm{SL}_2(\mathbb{Z})$.

Notation 3.11 (Projective line)**.** For $c \in \mathbb{Z}_+$, we recall the projective line

$$\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z}) := \{(x, y) : x, y \in \mathbb{Z}/c\mathbb{Z},\ x(\mathbb{Z}/c\mathbb{Z}) + y(\mathbb{Z}/c\mathbb{Z}) = \mathbb{Z}/c\mathbb{Z}\}/_\sim,$$

where $\sim$ is the equivalence relation generated by $(x, y) \sim (\alpha x, \alpha y)$ for $\alpha \in (\mathbb{Z}/c\mathbb{Z})^\times$. We write the equivalence class of (x, y) as $[x : y]$, and we will typically use the letters u, v to denote projective points in $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$, reserving x, y for elements of $\mathbb{Z}/c\mathbb{Z}$. For $d \mid c$, we write the natural map $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z}) \to \mathbb{P}^1(\mathbb{Z}/d\mathbb{Z})$ which reduces both entries modulo d as $u \mapsto u \bmod d$.

The group $\mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})$ (and, through it, $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$) acts on $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ by

$$\begin{pmatrix} m & n \\ p & q \end{pmatrix} [x : y] := [mx + ny : px + qy]. \tag{3.14}$$

One can think of $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ as $\mathbb{Z}/c\mathbb{Z}$ with a few additional points, which must be included to obtain a well-defined action. Indeed, any projective point $u \in \mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ can be written as $u = [x : y]$ for some $x \in \{1, \dots, c\}$ and $y \mid c$ with $(x, y) = 1$, and thus $|\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})| = c^{1+o(1)}$. In particular, one can embed $\mathbb{Z}/c\mathbb{Z} \subset \mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ by $x \mapsto [x : 1]$, and via this embedding, the generators from (3.13) act on elements of $\mathbb{Z}/c\mathbb{Z}$ by

$$Tx = x + 1, \qquad Sy = -\overline{y}, \qquad \text{for } x \in \mathbb{Z}/c\mathbb{Z},\ y \in (\mathbb{Z}/c\mathbb{Z})^\times.$$

We now briefly go over a few facts about the subgroups and representations of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$.

Notation 3.12 (Reduction mod d)**.** Given a positive integer d with $d \mid c$, we denote by

$$\pi_{c,d} : \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/d\mathbb{Z})$$

the natural epimorphism which 'reads' the entries of $g \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ modulo d. We write

$$\Gamma_c(d) := \ker \pi_{c,d}$$

for the congruence subgroup given by the kernel of this map (consisting of matrices of the form $I + dA$, where one may view the entries of A as elements of $\mathbb{Z}/\frac{c}{d}\mathbb{Z}$).

Lemma 3.13. $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ *acts transitively on* $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ *(i.e., there is only one orbit). In fact, for* $d \mid c$*, there is a bijection between* $\mathbb{P}^1(\mathbb{Z}/d\mathbb{Z})$ *and orbits of* $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ *under* $\Gamma_c(d)$*,*

$$\begin{aligned} \Gamma_c(d)\backslash\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z}) &\longrightarrow \mathbb{P}^1(\mathbb{Z}/d\mathbb{Z}), \\ \Gamma_c(d) \cdot u &\longmapsto u \bmod d. \end{aligned} \tag{3.15}$$

Proof. For any $[x:y] \in \mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$, there exist by definition $a, b \in \mathbb{Z}/c\mathbb{Z}$ with $ax+by \equiv 1 \pmod{c}$, so $[x:y] = \left(\begin{smallmatrix} x & -b \\ y & a \end{smallmatrix}\right)[1:0] \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \cdot [1:0]$. Thus the action of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ on $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ is transitive.

The map in (3.15) is well-defined since $(I+dA)u \pmod{d} = u \pmod{d}$ for any $I + dA \in \Gamma_c(d)$. It is surjective since the original map $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z}) \to \mathbb{P}^1(\mathbb{Z}/d\mathbb{Z})$ is surjective. To show that (3.15) is also injective, suppose $u \pmod{d} = v \pmod{d}$ for some $u, v \in \mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$, and we aim to show that $\Gamma_c(d) \cdot u = \Gamma_c(d) \cdot v$. By the transitivity of the action of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$, we can find $g \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ such that $gv = [1:0] \in \mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$, so

$$(gu) \bmod d = (gv) \bmod d = [1:0] \in \mathbb{P}^1(\mathbb{Z}/d\mathbb{Z}).$$

Write $gu = [xd+1 : yd]$ for some $x, y \in \mathbb{Z}/c\mathbb{Z}$. Since $gu \in \mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$, we have $1 = (xd+1, yd, c) = (xd+1, yd^2, c)$, so there exist $a, b \in \mathbb{Z}/c\mathbb{Z}$ with $a(xd+1) + byd^2 \equiv 1 \pmod{c}$, and in particular $a \equiv 1 \pmod{d}$. Then,

$$gu = \begin{pmatrix} xd+1 & -bd \\ yd & a \end{pmatrix} [1:0] \in \Gamma_c(d) \cdot gv = g\Gamma_c(d) \cdot v,$$

where the last equality is due to the normality of $\Gamma_c(d)$. Hence $u \in \Gamma_c(d) \cdot v$, as we wanted. $\square$

By the Chinese remainder theorem ($\mathbb{Z}/c\mathbb{Z} \cong \prod_{p^k \| c} \mathbb{Z}/p^k\mathbb{Z}$), combining the maps π_{c,p^k} for $p^k \| c$ produces isomorphisms

$$\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \cong \prod_{p^k \| c} \mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z}), \qquad \Gamma_c(d) \cong \prod_{\substack{p^k \| c \\ p^j \| d}} \Gamma_{p^k}(p^j), \tag{3.16}$$

for $d \mid c$ (in the products above, it is understood that only primes which divide c are included, so $k \geq 1$, but we allow $j = 0$). Since $|\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})| = p^{3k}(1 - \frac{1}{p^2})$, it follows that

$$|\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})| = c^3 \prod_{\text{prime } p \mid c} \left(1 - \frac{1}{p^2}\right) \asymp c^3 \qquad \Rightarrow \qquad |\Gamma_c(d)| = \frac{|\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})|}{|\mathrm{SL}_2(\mathbb{Z}/d\mathbb{Z})|} = \frac{c^3}{d^3}, \tag{3.17}$$

and that the irreducible representations of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ can be parametrized as

$$\widehat{\mathrm{SL}}_2(\mathbb{Z}/c\mathbb{Z}) = \left\{ \bigboxtimes_{p^k \| c} \rho_{p,k} : \rho_{p,k} \in \widehat{\mathrm{SL}}_2(\mathbb{Z}/p^k\mathbb{Z}) \right\}. \tag{3.18}$$

Now let p be a prime and $k \in \mathbb{Z}_+$, and let us focus on understanding $\widehat{\mathrm{SL}}_2(\mathbb{Z}/p^k\mathbb{Z})$.

Definition 3.14 (Primitive representations)**.** A representation $\rho : \mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z}) \to U(V)$ is called *primitive* iff its kernel does not contain $\Gamma_{p^k}(p^{k-1})$. Equivalently (by the first isomorphism theorem), ρ cannot be factored as $\rho' \circ \pi_{p^k, p^{k-1}}$ for some representation ρ' of $\mathrm{SL}_2(\mathbb{Z}/p^{k-1}\mathbb{Z})$. A primitive (resp., non-primitive) character is one induced by a primitive (resp., non-primitive) representation.

Thus the primitive irreducible representations of $\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$ are 'new' at level p^k, much like primitive Dirichlet characters or newforms in the theory of automorphic representations. We can easily isolate the 'maximal' non-primitive component of a representation using the following lemma.

Lemma 3.15. *Let $\rho : \mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z}) \to U(V)$ be a representation and*

$$V_f := \{v \in V : \rho(g)v = v, \ \forall g \in \Gamma_{p^k}(p^{k-1})\}.$$

Then $\rho|_{V_f}$ is non-primitive, and $\rho|_{V_f^\perp}$ is isomorphic to a direct sum of primitive irreducible representations.

Proof. The fact that V_f and thus $V_f^\perp$ are an invariant subspaces of V follows quickly from the fact that $\Gamma_{p^k \to p^{k-1}} \lhd G$, so $\rho|_{V_f}$ and $\rho|_{V_f^\perp}$ are well-defined. By definition, $\rho|_{V_f}(g) = \mathrm{Id}_{V_f}$ for all $g \in \Gamma_{p^k \to p^{k-1}}$, so the kernel of $\rho|_{V_f}$ includes $\Gamma_{p^k \to p^{k-1}}$, i.e., $\rho|_{V_f}$ is non-primitive.

Now let $\rho|_{V_0}$ be any irreducible subrepresentation of $\rho|_{V_f^\perp}$, where $V_0 \subset V_f^\perp$. Since $V_0 \neq \{0\}$ and $V_0 \cap V_f = \{0\}$, we can find some $v \in V_0 \setminus V_f$, and thus some $g \in \Gamma_{p^k \to p^{k-1}}$ such that $\rho(g)v \neq v$. But then $\rho|_{V_0}(g) \neq \mathrm{Id}_{V_0}$, so the kernel of $\rho|_{V_0}$ does not contain $\Gamma_{p^k \to p^{k-1}}$, i.e., ρ_{V_0} is primitive. ☐

The primitive irreducible representations of $\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$ are fairly complicated, but the following lemma will suffice for our purposes. This generalizes the classical spectral-gap result for $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$.

Lemma 3.16. *Any primitive irreducible representation ρ of $\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$ has $\dim \rho \gg p^k$.*

Proof. Complete tables with the dimensions of irreducible representations of $\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$, including the case $p = 2$, were given by Nobs–Wolfart [31, p. 525] (who refer to primitive representations in the sense of our Lemma 3.16 as having 'level k'). For odd primes p, these had been classified by Shalika [36, §4.3], Tanaka [41], and Kutzko [26].

For a more direct proof of the lower bound via Clifford theory, we refer to Bourgain–Gamburd's [5, Lemma 7.1] (this assumes p is odd, but an analogous argument applies if $p = 2$). To summarize their argument when k is even, Bourgain–Gamburd apply (a variant of) Lemma 3.9 with $G := \mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$ and $N := \Gamma_{p^k}(p^{k/2})$, to decompose $\rho|_N$ into irreducible representations $\sigma_1, \ldots, \sigma_L \in \widehat{N}$, all lying in the same orbit under G-conjugation. But N is abelian, so $\widehat{N} \cong N$, and $\sigma_1, \ldots, \sigma_L$ correspond to G-conjugate elements $g_1, \ldots, g_L \in N = \Gamma_{p^k}(p^{k/2})$. Moreover, the primitivity condition that $\ker \rho$ does not contain $\Gamma_{p^k}(p^{k-1})$ implies that $g_1, \ldots, g_L \notin \Gamma_{p^k}(p^{(k/2)+1})$. It follows that

$$\dim \rho \geq L = \frac{|G|}{|C_G(g_1)|} \gg \frac{p^{3k}}{|C_G(g_1)|},$$

where $C_G(g_1)$ is the centralizer of g_1 in G. It thus remains to bound $|C_G(g)| \ll p^{2k}$ for $g \in \Gamma_{p^k}(p^{k/2}) \setminus \Gamma_{p^k}(p^{(k/2)+1})$, which follows from an explicit matrix computation [5, Claim 7.1] (minor modifications are needed here if $p = 2$, but these only incur a constant-factor loss). ☐

4. Representations and Kloosterman matrices

Here we connect matrices of Kloosterman sums modulo c to Fourier analysis on $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$.

4.1. **The relevant representations.**

When digesting the notation below, the reader should keep in mind the informal outline from Section 2.2. We will first define the simpler representations (ρ_c, V_c) which are connected to matrices of Kloosterman sums $S(m, n; c)$, and then the more relevant subrepresentations $(\rho_c^\circ, V_c^\circ)$ which correspond to adding the restriction $(m, n, c) = 1$. In fact, the subspace $V_c^\circ \subset V_c$ will be constructed by sifting out 'old' subspaces isomorphic to V_d for $d \mid c$.

Definition 4.1 (Permutation representations of the projective action)**.** For $c \in \mathbb{Z}_+$, we denote the permutation representation of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ induced by the action (3.14) on $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ by

$$\rho_c : \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \to U(V_c), \qquad V_c := \mathbb{C}^{\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})}, \tag{4.1}$$

and its character by $\chi_c := \mathrm{Tr}\rho_c$. Hence V_c is the space of functions $f : \mathbb{P}^1(\mathbb{Z}/c\mathbb{Z}) \to \mathbb{C}$, equipped with the standard inner product, and $(\rho_c(g)f)(u) = f(g^{-1}u)$ for $g \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$, $u \in \mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$. In particular, for any $u \in \mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$, one has $\rho_c(g)\mathbb{1}_u = \mathbb{1}_{gu}$.

Definition 4.2 (Invariant subspaces)**.** For $c, d \in \mathbb{Z}_+$ with $d \mid c$, define

$$V_c(d) := \{f \in V_c : \rho_c(n)f = f \quad \forall n \in \Gamma_c(d)\} \subset V_c.$$

In particular, $V_c(c) = V_c$. Thus $V_c(d)$ is the space of complex-valued functions on $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ which are constant on orbits of $\Gamma_c(d)$, so by Lemma 3.13,

$$V_c(d) \cong \mathbb{C}^{\Gamma_c(d)\backslash\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})} \cong \mathbb{C}^{\mathbb{P}^1(\mathbb{Z}/d\mathbb{Z})} = V_d. \tag{4.2}$$

Lemma 4.3. *For $c, d \in \mathbb{Z}_+$ with $d \mid c$, $V_c(d)$ is an invariant subspace of ρ_c. In fact, using Notation 3.12, we have*

$$\rho_c|_{V_c(d)} \cong \rho_d \circ \pi_{c,d}.$$

Proof. The fact that $V_c(d)$ is an invariant subspace follows immediately from the normality of $\Gamma_c(d)$. Now let $\Phi : V_d \to V_c(d)$ be the invertible linear map from (4.2), which relies on the bijection from (3.15). Then for any $g \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$, one can easily check that $\rho_c(g)|_{V_c(d)} \circ \Phi = \Phi \circ \rho_d(\pi_{c,d}(g))$: both maps take the basis vector $\mathbb{1}_u \in V_d = \mathbb{C}^{\mathbb{P}^1(\mathbb{Z}/d\mathbb{Z})}$ to the L^2-normalized function in $V_c(d)$ which is only nonzero on the orbit $g\Gamma_c(d) \cdot u = \Gamma_c(d) \cdot gu$. ☐

In light of Lemma 4.3, we will need to remove the contribution of 'old' representations (ρ_d, V_d) to (ρ_c, V_c). To this end, it will be helpful to adopt the following convention for tensor products.

Notation 4.4 (Ordered tensor products)**.** If $c, c_1, c_2 \in \mathbb{Z}_+$ are such that $c = c_1c_2$ and $(c_1, c_2) = 1$, then the Chinese Remainder Theorem gives $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z}) \cong \mathbb{P}^1(\mathbb{Z}/c_1\mathbb{Z}) \times \mathbb{P}^1(\mathbb{Z}/c_2\mathbb{Z})$ by $u \mapsto (u \bmod c_1, u \bmod c_2)$, so $V_c \cong V_{c_1} \otimes V_{c_2}$. Since tensor products of vector spaces are defined up to isomorphism, it is not a great abuse of notation to write

$$V_c = V_{c_1} \otimes V_{c_2}.$$

In particular, given $f_1 \in V_{c_1}$ and $f_2 \in V_{c_2}$, we view $f_1 \otimes f_2$ as a function on $\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ with values $(f_1 \otimes f_2)(u) = f_1(u \bmod c_1) \cdot f_2(u \bmod c_2)$. This notation extends to tensor products of subspaces $W_1 \subset V_{c_1}$, $W_2 \subset V_{c_2}$ (so $W_1 \otimes W_2 \subset V_c$), and of linear transformations $T_1 : V_{c_1} \to V_{c_1}$, $T_2 : V_{c_2} \to V_{c_2}$.

We note that with the conventions from Notations 3.7 and 4.4, for $c = c_1c_2$ with $(c_1, c_2) = 1$, the product $\rho_{c_1} \boxtimes \rho_{c_2} : \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \to U(V_{c_1} \otimes V_{c_2}) = U(V_c)$ is precisely the permutation representation ρ_c (this is a genuine equality, not just an isomorphism). Moreover, a tensor product of invariant subspaces of ρ_{c_1} and ρ_{c_2} gives an invariant subspace of ρ_c, and in fact $V_{c_1}(d_1) \otimes V_{c_2}(d_2) = V_c(d)$ for $d_1 \mid c_1$, $d_2 \mid c_2$, and $d = d_1d_2$. Iterating this yields the factorizations

$$V_c = \bigotimes_{p^k \| c} V_{p^k}, \qquad \rho_c = \boxtimes_{p^k \| c} \rho_{p^k}. \tag{4.3}$$

and, more generally, for $d \mid c$,

$$V_c(d) = \bigotimes_{\substack{p^k \| c \\ p^j \| d}} V_{p^k}(p^j), \qquad \rho_c|_{V_c(d)} = \boxtimes_{\substack{p^k \| c \\ p^j \| d}} \rho_{p^k}|_{V_{p^k}(p^j)}. \tag{4.4}$$

Finally, we can define the representations $(\rho_c^\circ, V_c^\circ)$.

Definition 4.5 (Sifted representations)**.** For a prime power p^k, we let $V_{p^k}^\circ := V_{p^k}(p^{k-1})^\perp \subset V_{p^k}$ be the orthogonal complement of $V_{p^k}(p^{k-1})$ inside V_{p^k} (which is an invariant subspace of ρ_{p^k}). For $c \in \mathbb{Z}_+$, we define

$$V_c^\circ := \bigotimes_{p^k \| c} V_{p^k}^\circ, \qquad \rho_c^\circ := \rho_c|_{V_c^\circ}, \qquad \chi_c^\circ := \mathrm{Tr}\rho_c^\circ.$$

Proposition 4.6 (Decomposition of sifted representations)**.** *For any $c \in \mathbb{Z}_+$, one has*

$$\rho_c^\circ = \boxtimes_{p^k \| c} \rho_{p^k}^\circ. \tag{4.5}$$

Moreover, each $\rho_{p^k}^\circ$ is isomorphic to a nonempty direct sum of primitive irreducible representations, and ρ_c° is isomorphic to a direct sum of $c^{o(1)}$ irreducible representations of dimensions $c^{1-o(1)}$.

Proof. The factorization in (4.5) follows immediately from (4.3) and Definition 4.5. The fact that $\rho_{p^k}^\circ = \rho_{p^k}|_{V_{p^k}^\circ}$ is isomorphic to a direct sum of primitive irreducible representations is precisely the content of Lemma 3.15, wherein $V_f = V_{p^k}(p^{k-1})$ and $V_f^\perp = V_{p^k}^\circ$. One can easily construct a function on V_{p^k} which is not constant on orbits of $\Gamma_{p^k}(p^{k-1})$, so $V_{p^k}^\circ \neq \{0\}$, and thus $\rho_{p^k}^\circ \neq 0$.

Now write each $\rho_{p^k}^\circ$ as a direct sum of primitive irreducible representations $\rho_{p,k}$ of $\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$ (up to isomorphism), and expand the tensor product in (4.5). This expresses ρ_c° as a direct sum of representations (potentially with repetitions) of the shape

$$\rho = \boxtimes_{p^k \| c} \rho_{p,k},$$

which are irreducible, and have dimensions $\gg c^{1-o(1)}$ by Lemma 3.16 and the divisor bound. Since $\dim \rho_c^\circ \leq \dim \rho_c = |\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})| = c^{1+o(1)}$, the number of these representations is at most $c^{o(1)}$. $\square$

We now briefly analyze the orthogonal projections onto invariant subspaces of V_c. It will turn out that the projection onto V_c° can be obtained by a Möbius-inversion-type process.

Definition 4.7 (Special projections)**.** For $c, d \in \mathbb{Z}_+$ with $d \mid c$, we let $P_c(d), P_c^\circ : V_c \to V_c$ be the orthogonal projections onto $V_c(d)$, respectively V_c°. In particular, $P_c(c)$ is the identity map on V_c.

Lemma 4.8. *For $d \mid c$, one has*

$$P_c(d) = \bigotimes_{\substack{p^k \| c \\ p^j \| d}} P_{p^k}(p^j) = \frac{1}{|\Gamma_c(d)|} \sum_{n \in \Gamma_c(d)} \rho_c(n). \tag{4.6}$$

In particular, $P_c(d)$ commutes with $\rho_c(g)$ for any $g \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$. The matrix representation of this map with respect to the standard basis of $V_c = \mathbb{C}^{\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})}$ has entries

$$P_c(d)_{u,v} = \frac{d^2 \phi(c)}{c^2 \phi(d)} \mathbb{1}_{u \in \Gamma_c(d) \cdot v}, \qquad u, v \in \mathbb{P}^1(\mathbb{Z}/c\mathbb{Z}). \tag{4.7}$$

The proof of Lemma 4.8 is left to Appendix A.

Lemma 4.9. *For $c \in \mathbb{Z}_+$, one has*

$$P_c^\circ = \bigotimes_{p^k \| c} P_{p^k}^\circ = \sum_{d \mid c} \mu\left(\frac{c}{d}\right) P_c(d).$$

Proof. The factorization as a tensor product follows immediately from Definitions 4.5 and 4.7. Now for a prime power p^k, recall that $P_{p^k}(p^k)$ is the identity map on V_{p^k} and $P_{p^k}(p^{k-1})$ is the orthogonal projection onto $V_{p^k}(p^{k-1})$, so the orthogonal projection onto $V_{p^k}^\circ = V_{p^k}(p^{k-1})^\perp$ can be written as

$$P_{p^k}^\circ = P_{p^k}(p^k) - P_{p^k}(p^{k-1}).$$

It follows from this and (4.6) that

$$\bigotimes_{p^k \| c} P^\circ_{p^k} = \bigotimes_{p^k \| c} \Big(P_{p^k}(p^k) - P_{p^k}(p^{k-1}) \Big)$$
$$= \sum_{d|c} \mu\left(\frac{c}{d}\right) \bigotimes_{\substack{p^k \| c \\ p^j \| d}} P_{p^k}(p^j) \quad = \sum_{d|c} \mu\left(\frac{c}{d}\right) P_c(d),$$

as claimed. □

4.2. **The Kloosterman matrix.** Here we finally relate the abstract discussion in the preceding subsections to the classical Kloosterman sums.

Proposition 4.10 (From Kloosterman matrices to Fourier coefficients). *Let $c \in \mathbb{Z}_+$, $\psi_1, \psi_2 : \mathbb{Z}/c\mathbb{Z} \to \mathbb{C}$, and $K_c^{\psi_1,\psi_2} \in \mathbb{C}^{\mathbb{Z}/c\mathbb{Z} \times \mathbb{Z}/c\mathbb{Z}}$ be the $c \times c$ complex matrix with entries*

$$(K_c^{\psi_1,\psi_2})_{m,n} := \psi_1(m)\psi_2(n)\mathbb{1}_{(m,n,c)=1} S(m,n;c). \tag{4.8}$$

Consider the function $F_c^{\psi_1,\psi_2} : \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \to \mathbb{C}$ given by

$$F_c^{\psi_1,\psi_2} := \frac{1}{c^2} \sum_{h_1,h_2 \in \mathbb{Z}/c\mathbb{Z}} \widehat{\psi}_1(h_1)\widehat{\psi}_2(h_2)\, \mathbb{1}_{T^{h_1} S T^{h_2}}, \tag{4.9}$$

where T and S are as in (3.13). Then one has the inequality of operator norms

$$\|K_c^{\psi_1,\psi_2}\| \le c \|\widehat{F}_c^{\psi_1,\psi_2}(\rho_c^\circ)\|.$$

Remark. In $\widehat{\psi}_1$ and $\widehat{\psi}_2$, the Fourier transform is taken over $\mathbb{Z}/c\mathbb{Z}$, as in (3.11). In $\widehat{F}_c^{\psi_1,\psi_2}$, the Fourier transform is taken over the non-abelian group $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$, as in (3.10).

Proof of Proposition 4.10. Let U_c be the unitary $c \times c$ matrix with entries $(U_c)_{u,v} = c^{-1/2} e(\frac{uv}{c})$. By expanding the Kloosterman sums, we have

$$(U_c^* K_c^{\psi_1,\psi_2} U_c)_{u,v} = \frac{1}{c} \sum_{\substack{m,n \in \mathbb{Z}/c\mathbb{Z} \\ x \in (\mathbb{Z}/c\mathbb{Z})^\times}} \mathbb{1}_{(m,n,c)=1} \psi_1(m)\, e\left(\frac{m(x-u)}{c}\right) \psi_2(n)\, e\left(\frac{n(\overline{x}+v)}{c}\right),$$

for any $u, v \in \mathbb{Z}/c\mathbb{Z}$. We then expand the indicator $\mathbb{1}_{(m,n,c)=1}$ by Möbius inversion and Fourier analysis,

$$\mathbb{1}_{(m,n,c)=1} = \sum_{d|c} \mu(d) \mathbb{1}_{d|m} \mathbb{1}_{d|n} = \sum_{d|c} \frac{\mu(d)}{d^2} \sum_{a,b \in \mathbb{Z}/d\mathbb{Z}} e\left(\frac{am}{d}\right) e\left(\frac{bm}{d}\right),$$

and evaluate the sums over m, n to obtain

$$(U_c^* K_c^{\psi_1,\psi_2} U_c)_{u,v} = \frac{1}{c} \sum_{d|c} \frac{\mu(d)}{d^2} \sum_{x \in (\mathbb{Z}/c\mathbb{Z})^\times} \sum_{a,b \in \mathbb{Z}/d\mathbb{Z}} \widehat{\psi}_1\left(-x + u - \frac{ac}{d}\right) \widehat{\psi}_2\left(-\overline{x} - v - \frac{bc}{d}\right)$$
$$= \frac{1}{c} \sum_{d|c} \frac{\mu(d)}{d^2} \sum_{x \in (\mathbb{Z}/c\mathbb{Z})^\times} \sum_{h_1,h_2 \in \mathbb{Z}/c\mathbb{Z}} \widehat{\psi}_1(h_1)\widehat{\psi}_2(h_2) \mathbb{1}_{\substack{x \equiv u - h_1 \pmod{\frac{c}{d}} \\ -\overline{x} \equiv v + h_2 \pmod{\frac{c}{d}}}},$$

where we substituted $h_1 := -x + u + \frac{ac}{d}$, $h_2 = -\overline{x} - v - \frac{bc}{d}$. Switching divisors $d \mapsto \frac{c}{d}$, swapping sums, and evaluating the sum over x (which gives either 0 or $\phi(c)/\phi(d)$ solutions), we reach

$$(U_c^* K_c^{\psi_1,\psi_2} U_c)_{u,v} = \frac{1}{c} \sum_{h_1,h_2 \in \mathbb{Z}/c\mathbb{Z}} \widehat{\psi}_1(h_1)\widehat{\psi}_2(h_2) \sum_{d|c} \mu\left(\frac{c}{d}\right) \frac{d^2 \phi(c)}{c^2 \phi(d)} \mathbb{1}_{(u-h_1)(v+h_2) \equiv -1 \pmod{d}}. \tag{4.10}$$

Let us keep this in mind. Separately, by (4.9) and Definition 4.5, we have

$$\begin{aligned}\widehat{F}_c^{\psi_1,\psi_2}(\rho_c^\circ) &= \frac{1}{c^2}\sum_{h_1,h_2\in\mathbb{Z}/c\mathbb{Z}} \widehat{\psi}_1(h_1)\widehat{\psi}_2(h_2)\rho_c^\circ(T^{h_1}ST^{h_2}) \\ &= \left(\frac{1}{c^2}\sum_{h_1,h_2\in\mathbb{Z}/c\mathbb{Z}} \widehat{\psi}_1(h_1)\widehat{\psi}_2(h_2)\rho_c(T^{h_1}ST^{h_2})\right)\Bigg|_{V_c^\circ},\end{aligned}$$

and thus by Lemma 3.1,

$$\|\widehat{F}_c^{\psi_1,\psi_2}(\rho_c^\circ)\| = \|M_c^{\psi_1,\psi_2}\|, \tag{4.11}$$

where $M_c^{\psi_1,\psi_2} : V_c \to V_c$ is the map

$$M_c^{\psi_1,\psi_2} = \frac{1}{c^2}\sum_{h_1,h_2\in\mathbb{Z}/c\mathbb{Z}} \widehat{\psi}_1(h_1)\widehat{\psi}_2(h_2)\rho_c(T^{h_1}ST^{h_2})P_c^\circ.$$

By Lemma 4.9 and the commutativity claim in Lemma 4.8, we can further write

$$\begin{aligned}M_c^{\psi_1,\psi_2} &= \frac{1}{c^2}\sum_{h_1,h_2\in\mathbb{Z}/c\mathbb{Z}} \widehat{\psi}_1(h_1)\widehat{\psi}_2(h_2)\rho_c(T^{h_1}ST^{h_2})\sum_{d\mid c}\mu\left(\frac{c}{d}\right)P_c(d) \\ &= \frac{1}{c^2}\sum_{h_1,h_2\in\mathbb{Z}/c\mathbb{Z}} \widehat{\psi}_1(h_1)\widehat{\psi}_2(h_2)\sum_{d\mid c}\mu\left(\frac{c}{d}\right)\rho_c(T^{h_1})P_c(d)\rho_c(ST^{h_2}).\end{aligned}$$

By Definition 4.1 and (4.7), we can represent this map as a matrix in $\mathbb{C}^{\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})\times\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})}$ with entries

$$(M_c^{\psi_1,\psi_2})_{u,v} = \frac{1}{c^2}\sum_{h_1,h_2\in\mathbb{Z}/c\mathbb{Z}} \widehat{\psi}_1(h_1)\widehat{\psi}_2(h_2)\sum_{d\mid c}\mu\left(\frac{c}{d}\right)\frac{d^2\phi(c)}{c^2\phi(d)}\mathbb{1}_{T^{-h_1}u\in\Gamma_c(d)\cdot ST^{h_2}v} \tag{4.12}$$

for $u,v\in\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$; compare this to (4.10). We will show that restricting the matrix $M_c^{\psi_1,\psi_2}$ to those rows and columns indexed by $u,v\in\mathbb{Z}/c\mathbb{Z}\subset\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z})$ (by the canonical embedding $x\mapsto[x:1]$) yields precisely the matrix $U_c^*K_c^{\psi_1,\psi_2}U_c$. Indeed, using the notation above, if $u,v\in\mathbb{Z}/c\mathbb{Z}$, then $T^{-h_1}u = u-h_1 =: x\in\mathbb{Z}/c\mathbb{Z}$, $T^{h_2}v = v+h_2 =: y\in\mathbb{Z}/c\mathbb{Z}$, and we have $x\in\Gamma_c(d)\cdot Sy$ if and only if the equation

$$(I+dA)\begin{pmatrix}x\\1\end{pmatrix} = \alpha\begin{pmatrix}-1\\y\end{pmatrix}$$

has solutions in $I+dA\in\Gamma_c(d)$ and $\alpha\in(\mathbb{Z}/c\mathbb{Z})^\times$. On the one hand, the existence of such solutions implies that $\left(\begin{smallmatrix}x\\1\end{smallmatrix}\right)\equiv\left(\begin{smallmatrix}-\alpha\\\alpha y\end{smallmatrix}\right) \pmod d$, so $xy\equiv -1 \pmod d$. On the other hand, if $xy\equiv-1\pmod d$, then one can take $A=0$ and $\alpha=-x$ to obtain a solution. It follows that

$$\mathbb{1}_{T^{-h_1}u\in\Gamma_c(d)\cdot ST^{h_2}v} = \mathbb{1}_{(u-h_1)(v+h_2)\equiv-1 \pmod d}, \qquad u,v\in\mathbb{Z}/c\mathbb{Z}\subset\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z}),$$

and then by comparing (4.10) and (4.12), we find that

$$(U_c^*K_c^{\psi_1,\psi_2}U_c)_{u,v} = c(M_c^{\psi_1,\psi_2})_{u,v} \qquad u,v\in\mathbb{Z}/c\mathbb{Z}\subset\mathbb{P}^1(\mathbb{Z}/c\mathbb{Z}).$$

Since removing some rows and columns of a matrix can only decrease its spectral norm, we conclude that

$$\|K_c^{\psi_1,\psi_2}\| = \|U_c^*K_c^{\psi_1,\psi_2}U_c\| \le c\|M_c^{\psi_1,\psi_2}\|,$$

which, together with (4.11), completes our proof. ☐

Corollary 4.11. *Let $c,M,N\in\mathbb{Z}_+$ with $1\le M,N\le c$, $a\in(\mathbb{Z}/c\mathbb{Z})^\times$, and $\mathcal{I},\mathcal{J}\subset\mathbb{Z}$ be intervals of lengths $|\mathcal{I}|=M$, $|\mathcal{J}|=N$. Let $K_{c,a}^{\mathcal{I},\mathcal{J}}\in\mathbb{C}^{\mathcal{I}\times\mathcal{J}}$ be the $M\times N$ matrix indexed by $m\in\mathcal{I}$ and $n\in\mathcal{J}$, with entries*

$$(K_{c,a}^{\mathcal{I},\mathcal{J}})_{m,n} := S(am,n;c)\mathbb{1}_{(m,n,c)=1}. \tag{4.13}$$

Let $\varepsilon > 0$ and $H_1 := c^{1+\varepsilon} M^{-1}$, $H_2 := c^{1+\varepsilon} N^{-1}$. Then there exist complex weights $\alpha_h, \beta_h \ll 1$ such that for the function $F_{c,a}^{H_1,H_2} : \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \to \mathbb{C}$ given by

$$F_{c,a}^{H_1,H_2} := \frac{1}{H_1 H_2} \sum_{\substack{|h_1| \le H_1 \\ |h_2| \le H_2}} \alpha_{h_1} \beta_{h_2} \mathbb{1}_{T^{\bar{a} h_1} S T^{h_2}}, \tag{4.14}$$

one has

$$\|K_{c,a}^{\mathcal{I},\mathcal{J}}\| \le c^{1+2\varepsilon} \|\widehat{F}_{c,a}^{H_1,H_2}(\rho_c^\circ)\| + O_\varepsilon(c^{-100}). \tag{4.15}$$

Remark. Given (4.14), one can apply the triangle inequality for the operator norm to obtain

$$\begin{aligned} \|\widehat{F}_{c,a}^{H_1,H_2}(\rho_c^\circ)\| &= \left\| \frac{1}{H_1 H_2} \sum_{\substack{|h_1| \le H_1 \\ |h_2| \le H_2}} \alpha_{h_1} \beta_{h_2}\, \rho_c^\circ(T^{a h_1} S T^{h_2}) \right\| \\ &\le \frac{1}{H_1 H_2} \sum_{\substack{|h_1| \le H_1 \\ |h_2| \le H_2}} |\alpha_{h_1} \beta_{h_2}| \|\rho_c^\circ(T^{a h_1} S T^{h_2})\| \ll 1, \end{aligned} \tag{4.16}$$

since $\|\rho_c^\circ(T^{a h_1} S T^{h_2})\| = 1$ (as the norm of a unitary map). Plugging this into (4.15) recovers the trivial bound $\|K_c^{\mathcal{I},\mathcal{J}}\| \ll c^{1+o(1)}$ from (1.2). Our task in the later sections will therefore be to establish some power-saving spectral cancellation in the sum over h_1, h_2 from (4.16).

Proof of Corollary 4.11. Let us write $[M] := \{1, \ldots, M\}$, $[N] := \{1, \ldots, N\}$, and $\mathcal{I} = [M] + r$, $\mathcal{J} = [N] + s$ for some $r, s \in \mathbb{Z}$. Since $M, N \le c$, we may identify $\mathcal{I}$, $\mathcal{J}$ with their images in $\mathbb{Z}/c\mathbb{Z}$. Let $\Phi : \mathbb{R} \to \mathbb{C}$ be a smooth function supported in $[-1, 2]$, such that $\Phi \ge \mathbb{1}_{[0,1]}$ and $\Phi^{(j)} \ll_j 1$ for $j \ge 0$, and define $\psi_1, \psi_2 : \mathbb{Z}/c\mathbb{Z} \to \mathbb{C}$ by

$$\psi_1(m) := \sum_{\substack{m' \in \mathbb{Z} \\ a(m'+r) \equiv m \pmod{c}}} \Phi\left(\frac{m'}{M}\right), \qquad \psi_2(n) := \sum_{\substack{n' \in \mathbb{Z} \\ n'+s \equiv n \pmod{c}}} \Phi\left(\frac{n'}{N}\right). \tag{4.17}$$

Since $\Phi \ge \mathbb{1}_{[0,1]}$, we have $\psi_1 \ge \mathbb{1}_{a\mathcal{I}}$ and $\psi_2 \ge \mathbb{1}_{\mathcal{J}}$ (viewing these as functions on $\mathbb{Z}/c\mathbb{Z}$). But scaling a row or a column of a matrix by a constant in $[0, 1]$ can only decrease its spectral norm, so with the notation from (4.8) we get

$$\|K_c^{\psi_1,\psi_2}\| \ge \|K_{c,a}^{\mathcal{I},\mathcal{J}}\|.$$

So from Proposition 4.10, it follows that

$$\|K_{c,a}^{\mathcal{I},\mathcal{J}}\| \le c \|\widehat{F}_c^{\psi_1,\psi_2}(\rho_c^\circ)\|, \tag{4.18}$$

and it remains to compute $F_c^{\psi_1,\psi_2}$. For $h \in \mathbb{Z}/c\mathbb{Z}$, we obtain from (4.17), (3.11), (3.3), and (3.2) that

$$\begin{aligned} \widehat{\psi}_1(h) = \sum_{m \in \mathbb{Z}/c\mathbb{Z}} \psi_1(m)\, e\left(-\frac{hm}{c}\right) &= \sum_{m' \in \mathbb{Z}} \Phi\left(\frac{m'}{M}\right) e\left(-\frac{ah(m'+r)}{c}\right) \\ &= M e\left(-\frac{rah}{c}\right) \sum_{k \in \mathbb{Z}} \widehat{\Phi}\left(M\Big(k + \frac{ah}{c}\Big)\right) \\ &= M e\left(-\frac{rah}{c}\right) \sum_{\substack{h' \in \mathbb{Z} \\ h' \equiv ah \pmod{c}}} \widehat{\Phi}\left(\frac{h'}{c/M}\right), \end{aligned}$$

and similarly for $\widehat{\psi}_2(h)$ (with 1 in place of a). Plugging this into (4.9), we conclude that

$$F_c^{\psi_1,\psi_2} = \frac{MN}{c^2} \sum_{h_1,h_2\in\mathbb{Z}/c\mathbb{Z}} \sum_{\substack{h_1',h_2'\in\mathbb{Z} \\ h_1'\equiv ah_1 \pmod c \\ h_2'\equiv h_2 \pmod c}} \widehat{\Phi}\left(\frac{h_1'}{c/M}\right)\widehat{\Phi}\left(\frac{h_2'}{c/N}\right) e\left(\frac{-rah_1-sh_2}{c}\right)\mathbb{1}_{T^{h_1}ST^{h_2}}$$

$$= \frac{c^{2\varepsilon}}{H_1H_2}\sum_{h_1',h_2'\in\mathbb{Z}} \widehat{\Phi}\left(\frac{h_1'}{c/M}\right)\widehat{\Phi}\left(\frac{h_2'}{c/N}\right) e\left(\frac{-rah_1'-sh_2'}{c}\right)\mathbb{1}_{T^{\bar{a}h_1'}ST^{h_2'}}.$$

Using the Schwarz decay of $\widehat{\Phi}$, we can discard the contribution of the terms with $|h_1'| > H_1$ or $|h_2'| > H_2$ to $\|\widehat{F}_c^{\psi_1,\psi_2}(\rho_c^\circ)\|$, up to an error of $O_\varepsilon(c^{-100})$. Choosing

$$\alpha_h := \widehat{\Phi}\left(\frac{h}{c/M}\right) e\left(-\frac{rah}{c}\right), \qquad \beta_h := \widehat{\Phi}\left(\frac{h}{c/N}\right) e\left(-\frac{sh}{c}\right)$$

concludes our proof in light of (4.18) and (4.14). $\square$

5. The amplification argument

Recall that Proposition 4.10 and Corollary 4.11 reduce the problem of bounding bilinear forms with Kloosterman sums to that of bounding Fourier coefficients of certain functions on $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ at a certain representation. One can then reduce to irreducible subrepresentations via Lemma 3.10. To use Fourier analysis, we will pass to a sum over all irreducible representations of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$—but to avoid a critical loss, we insert an amplifier weight in this sum, as outlined in Section 2.3. Making this work in a non-abelian setting is the key step in our argument.

5.1. Introducing the amplifier.

We state the results in this subsection in fair generality, since they may be useful in other contexts. We recall the notation for Schatten norms from Section 3.1.

Proposition 5.1 (Non-abelian amplification). *Let G be a finite group, $N \triangleleft G$, $F : G \to \mathbb{C}$, $\rho \in \widehat{G}$, $\chi := \operatorname{Tr}\rho$, and q be an even positive integer. Then one has*

$$\|\widehat{F}(\rho)\|_{S^q}^q \le \frac{|G|}{\sum_{n\in N}|\chi(n)|^2} \sum_{\substack{g_1,\dots,g_q\in G \\ g_1\cdots g_q\in N}} F(g_1)\overline{F}(g_2^{-1})\cdots F(g_{q-1})\overline{F}(g_q^{-1})\chi(g_1\cdots g_q).$$

Remark. In comparison, expanding $\|\widehat{F}(\rho)\|_{S^q}^q$ by (3.5) and (3.10) yields

$$\|\widehat{F}(\rho)\|_{S^q}^q = \sum_{g_1,\dots,g_q\in G} F(g_1)\overline{F}(g_2^{-1})\cdots F(g_{q-1})\overline{F}(g_q^{-1})\chi(g_1\cdots g_q).$$

The upper bound from Proposition 5.1 replaces χ with a function proportional to $\chi\mathbb{1}_N$, normalized so that equality is attained when $F = \overline{\chi}$. The requirement that ρ be irreducible is crucial.

We first prove Proposition 5.1 using character orthogonality, which resembles more classical amplification arguments, and then give a more conceptual sketch of proof via induced representations. The first proof has the advantage that it may be generalized by considering other amplifiers.

Proof of Proposition 5.1 via character orthogonality. Write $F_0 := F$ and $F_1 : G \to \mathbb{C}$ for the function $F_1(g) := \overline{F}(g^{-1})$, so that $\widehat{F}_1(\rho) = \widehat{F}(\rho)^*$. By considering the function

$$F_{\mathrm{conv}} := F_0 * F_1 * F_0 * \cdots * F_{\frac{q}{2} \bmod 2},$$

where there are $\frac{q}{2}$ factors in the convolution, so that $\|\widehat{F}_{\mathrm{conv}}(\rho)\|_{S^2}^2 = \|\widehat{F}(\rho)\|_{S^q}^q$, we see that it suffices to prove the desired result when $q = 2$.

Let $\mathbb{1}_N : G \to \{0,1\}$ be the indicator function of N, and consider the amplifier $A : \widehat{G} \to [0,\infty)$ given at $\rho' \in \widehat{G}$ with $\chi' := \mathrm{Tr}\rho'$ by

$$A(\rho') := \frac{1}{|N|} \left\| \widehat{\mathbb{1}}_N \left(\overline{\rho}' \otimes \rho \right) \right\|_{S^2}^2 = \frac{1}{|N|} \left\| \sum_{n \in N} \overline{\rho}'(n) \otimes \rho(n) \right\|_{S^2}^2$$
$$= \frac{1}{|N|} \mathrm{Tr} \left(\sum_{n_1, n_2 \in N} \overline{\rho}'(n_1 n_2^{-1}) \otimes \rho(n_1 n_2^{-1}) \right) = \sum_{n \in N} \overline{\chi'}(n) \chi(n),$$

where we implicitly used that N is a group. In particular, $A(\rho) = \sum_{n \in N} |\chi(n)|^2$ is a positive integer multiple of N, due to (3.9). It follows from this and nonnegativity that

$$\left(\sum_{n \in N} |\chi(n)|^2 \right) \|\widehat{F}(\rho)\|_{S^2}^2 \leq \sum_{\rho' \in \widehat{G}} A(\rho') \|\widehat{F}(\rho')\|_{S^2}^2$$
$$= \sum_{\rho' \in \widehat{G}} \sum_{n \in N} \overline{\mathrm{Tr}\rho'}(n) \chi(n) \|\widehat{F}(\rho')\|_{S^2}^2 = \sum_{n \in N} \chi(n) \sum_{\rho' \in \widehat{G}} \overline{\mathrm{Tr}\rho'}(n) \|\widehat{F}(\rho')\|_{S^2}^2.$$

Expanding $\|\widehat{F}(\rho')\|_{S^2}^2$ via (3.5) and (3.10) and then using (3.8), the sum over ρ' above becomes

$$\sum_{\rho' \in \widehat{G}} \overline{\mathrm{Tr}\rho'}(n) \sum_{g_1, g_2 \in G} F(g_1) \overline{F}(g_2^{-1}) \mathrm{Tr}\rho'(g_1 g_2) = \sum_{g_1, g_2 \in G} F(g_1) \overline{F}(g_2^{-1}) \sum_{\chi' \in \mathrm{Irr}(G)} \overline{\chi'}(n) \chi'(g_1 g_2)$$
$$= \sum_{g_1, g_2 \in G} F(g_1) \overline{F}(g_2^{-1}) \frac{|G|}{|C_n|} \mathbb{1}_{g_1 g_2 \in C_n},$$

where C_n denotes the conjugacy class of n in G. Since N can be partitioned into such conjugacy classes by normality, and since characters are constant on conjugacy classes, it follows that

$$\left(\sum_{n \in N} |\chi(n)|^2 \right) \|\widehat{F}(\rho)\|_{S^2}^2 \leq |G| \sum_{g_1, g_2 \in G} F(g_1) \overline{F}(g_2^{-1}) \sum_{n \in N} \chi(n) \frac{\mathbb{1}_{g_1 g_2 \in C_n}}{|C_n|}$$
$$= |G| \sum_{\substack{g_1, g_2 \in G \\ g_1 g_2 \in N}} F(g_1) \overline{F}(g_2^{-1}) \chi(g_1 g_2).$$

This settles the case $q = 2$, thus completing our proof. □

Remark. After reducing to the case $q = 2$, one could also attempt to use the triangle inequality for Schatten norms and Cauchy–Schwarz, in the shape

$$\left\| \sum_{g \in G} F(g) \rho(g) \right\|_{S^2}^2 \leq \left(\sum_{N g_0 \in N \backslash G} \left\| \sum_{g \in N g_0} F(g) \rho(g) \right\|_{S^2} \right)^2$$
$$\leq [G : N] \sum_{N g_0 \in N \backslash G} \left\| \sum_{g \in N g_0} F(g) \rho(g) \right\|_{S^2}^2 = \frac{|G|}{|N|} \sum_{\substack{g_1, g_2 \in G \\ g_1 g_2 \in N}} F(g_1) \overline{F}(g_2^{-1}) \chi(g_1 g_2).$$

This argument does not use that N is normal or that ρ is irreducible, but it produces a weaker bound in general. Indeed, compared to Proposition 5.1, the bound above loses a factor of

$$\frac{1}{|N|} \sum_{n \in N} |\chi(n)|^2,$$

which is a (potentially large) positive integer by (3.9); note that although χ is irreducible on G, it is usually not irreducible on N (e.g., when $N = \{e\}$ is the trivial subgroup, the factor lost is $(\dim \rho)^2$). This is a discrepancy which does not arise in the abelian setting, when all irreducible characters are 1-dimensional—which is why classical amplification arguments with Dirichlet characters can often be re-formulated by applying Cauchy–Schwarz to a sum over residues (see, e.g., [28]).

Proof sketch of Proposition 5.1 using induced representations. Starting from $\rho : G \to U(V)$, consider the restricted representation $\rho|_N : N \to U(V)$, and then the *induced representation*

$$R := \mathrm{Ind}_N^G(\rho|_N).$$

This acts by translation on the space

$$W := \{f \in V^G : f(ng) = \rho(n)f(g), \forall n \in N, g \in G\} \cong V^{G/N},$$

so $\dim R = \dim W = [G : N] \dim V = \frac{|G|}{|N|} \dim \rho$. It can be shown using (3.9) that ρ is a subrepresentation of R, with

$$\mathrm{Mult}(\rho, R) = \frac{1}{|N|} \sum_{n \in N} |\chi(n)|^2,$$

and therefore, by Lemma 3.10,

$$\|\widehat{F}(R)\|_{S^q}^q = \sum_{\rho' \in \widehat{G}} \mathrm{Mult}(\rho', R) \|\widehat{F}(\rho')\|_{S^q}^q \geq \frac{1}{|N|} \left(\sum_{n \in N} |\chi(n)|^2 \right) \|\widehat{F}(\rho)\|_{S^q}^q.$$

To complete the proof, one can expand $\|\widehat{F}(R)\|_{S^q}^q$ via (3.5) and (3.10), and use the Frobenius formula

$$\mathrm{Tr} R(g) = \frac{1}{|N|} \sum_{\substack{x \in G \\ x^{-1}gx \in N}} \chi(x^{-1}gx) = \frac{|G|}{|N|} \chi(g) \mathbb{1}_N(g),$$

for all $g \in G$, where the last equality uses the normality of N. □

Remark. In the particular case when $N = \{e\}$ is the trivial subgroup, the induced representation R in the proof above is (isomorphic to) the regular representation R_G. In this case, the conclusion of Proposition 5.1 simply reads

$$\|\widehat{F}(\rho)\|_{S^q}^q \leq \frac{1}{\dim \rho} \|\widehat{F}(R_G)\|_{S^q}^q,$$

and similar ideas appear in [37, 30].

5.2. **Passing to a counting problem.** We now return to the setting when $G = \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$ and $N = \Gamma_c(d)$ for some $d \mid c$. The goal is to pass from the spectral norm $\|\widehat{F}_{c,a}^{H_1,H_2}(\rho_c^\circ)\|$ in (4.15) to a count of solutions to a certain equation in $\mathrm{PSL}_2(\mathbb{Z}/d\mathbb{Z})$. After applying Proposition 5.1, we will need an upper bound for $\chi(g_1 \cdots g_q)$, and a lower bound for the denominator $\sum_{n \in N} |\chi(n)|^2$. For the specific characters χ_c and χ_c° from Section 4, such bounds are given in the following results.

Lemma 5.2. *Let $c \in \mathbb{Z}_+$, $g \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$, and d be the largest divisor of c such that*

$$g \in \{\gamma \in \mathbb{Z}/c\mathbb{Z} : \gamma^2 = 1\} \cdot \Gamma_c(d).$$

Let $f \leq \sqrt{cd}$ be the largest positive integer such that $f^2 \mid cd$. Then one has

$$\chi_c(g) \ll c^{o(1)} f. \tag{5.1}$$

Remark. The bound in (5.1) is sharp in terms of c and d, as can be seen by taking $g = T^d$. In particular, if $c = p^k$ is a prime power, then $g = T^{p^j}$ has $p^{\lfloor (k+j)/2 \rfloor}$ fixed points of the shape $[1 : x]$, where $p^j x^2 \equiv 0 \pmod{p^k}$.

The proof of Lemma 5.2 is left to Appendix A.

Proposition 5.3 (Lower bound for squared character sums)**.** *Let $c \in \mathbb{Z}_+$ and χ be an irreducible character inside χ_c°. Let $d, d', e \in \mathbb{Z}_+$ be such that $d' \mid d$, $(d,e)=1$, and $c = dd'e$. Then one has*

$$\sum_{n \in \Gamma_c(d)} |\chi(n)|^2 \gg \frac{c^{3-o(1)}}{d}.$$

Remark. The lower bound in Proposition 5.3 wins a factor of about d^2 over the 'trivial' bound of $|\Gamma_c(d)| = \frac{c^3}{d^3}$ due to (3.9). This is because $|\chi(n)|$ typically has size $\gtrsim d$ when $n \in \Gamma_c(d)$ (note that when $d = c$, one has $\Gamma_c(c) = \{I\}$ and $\chi(I) = \dim \chi$).

The proof of Proposition 5.3 reduces to a local computation (i.e., for $c = p^k$ a prime power) given below, which builds on Lemma 3.16. Indeed, one can rephrase Lemma 3.16 as follows: if $\chi \in \mathrm{Irr}(\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z}))$ is primitive, then $|\chi(I)|^2 \gg p^{2k}$. Using a bit of Clifford theory, we can generalize this to averages of $|\chi|^2$ over the congruence subgroups $\Gamma_{p^k}(p^j)$, for some values of j.

Lemma 5.4. *Let p^k be a prime power and $j \in \mathbb{Z}$ be such that either $j = 0$ or $\frac{k}{2} \le j \le k$. Let χ be a primitive irreducible character of $\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$. Then*

$$\sum_{n \in \Gamma_{p^k}(p^j)} |\chi(n)|^2 \gg (k-j+1)^{-1} p^{3k-j}.$$

Remark. We expect the bound in Lemma 5.4 to be sharp, and to actually hold for all $0 \le j \le k$. This might follow from a more careful study of $\widehat{\Gamma}_{p^k}(p^j)$ for $1 \le j < \frac{k}{2}$, and it would imply Proposition 5.3 (as well as Theorem 1.2) for more flexible factorizations $c = de$.

The proof of Lemma 5.4 is also left to Appendix A, a key ingredient being the fact that $\Gamma_{p^k}(p^j)$ is abelian if $\frac{k}{2} \le j \le k$.

Proof of Proposition 5.3. By (the proof of) Proposition 4.6, ρ_c° is isomorphic to a direct sum of representations of the shape

$$\rho = \boxtimes_{p^k \| c} \rho_{p,k}, \tag{5.2}$$

where each $\rho_{p,k}$ is a primitive irreducible representation of $\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$. This gives the decomposition of ρ_c° into irreducible representations, so any irreducible representation occurring inside ρ_c° must be of the form in (5.2). Now let $\chi := \mathrm{Tr}\rho$ and $\chi_{p,k} := \mathrm{Tr}\rho_{p,k}$ for such a representation. From (3.16) and fact that $\chi(n) = \prod_{p^k \| n} \chi_{p,k}(\pi_{c,p^k}(n))$ for $n \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$, it follows that

$$\sum_{n \in \Gamma_c(d)} |\chi(n)|^2 = \prod_{\substack{p^k \| c \\ p^j \| d}} \sum_{n \in \Gamma_{p^k}(p^j)} |\chi_{p,k}(n)|^2.$$

One can then apply Lemma 5.4 to obtain

$$\sum_{n \in \Gamma_c(d)} |\chi(n)|^2 \gg \prod_{p^k \| c} (k+1)^{-1} p^{3k-j},$$

Note that the hypothesis on j from Lemma 5.4 is satisfied because whenever p is a prime dividing c with $p^k \| c$ and $p^j \| d$, one of the following holds:

(1) One has $p \mid e$ and $p \nmid d$, so $j = 0$, or
(2) One has $p \mid d$ and $p \nmid e$, so $k = v_p(dd') \le 2j$.

The desired conclusion then follows from the divisor bound. ∎

Finally, we can state the result of our amplification argument.

Proposition 5.5 (From Fourier coefficients to a counting problem)**.** *Let $c \in \mathbb{Z}_+$, $a \in (\mathbb{Z}/c\mathbb{Z})^\times$, $H_1, H_2 \gg 1$, and $F_{c,a}^{H_1,H_2}$ be as in (4.14). Let $d, d', e \in \mathbb{Z}_+$ be such that $(d,e) = 1$, $d' \mid d$, and $c = dd'e$. Then for any even positive integer q, one has*

$$\|\widehat{F}_{c,a}^{H_1,H_2}(\rho_c^\circ)\|_{S^q}^q \ll \frac{c^{o(1)} d}{(H_1H_2)^{q/2}} \max_{\substack{d|\tilde{d}|c \\ \tilde{f}^2|c\tilde{d}}} \tilde{f} \sum_{\substack{h_1,\ldots,h_q \in \mathbb{Z} \\ |h_i| \le 2H_j \\ \forall i \equiv j \pmod 2}} \mathbb{1}_{T^{\bar{a}h_1}ST^{h_2}S\cdots T^{\bar{a}h_{q-1}}ST^{h_q}S = I \text{ in } \mathrm{PSL}_2(\mathbb{Z}/\tilde{d}\mathbb{Z})},$$

where both variables $\tilde{d}, \tilde{f}$ in the maxima are understood to be integers (note that $\tilde{f} \le \sqrt{c\tilde{d}}$).

Proof. By Proposition 4.6, ρ_c° is a sum of $c^{o(1)}$ irreducible representations ρ. By Lemma 3.10, it suffices to prove the desired upper bound for each $\|\widehat{F}_{c,a}^{H_1,H_2}(\rho)\|_{S^q}^q$. The loss factor of $c^{o(1)}$ is acceptable here (but if one is only interested in the spectral norm, so $q = \infty$, there is actually no loss at this step).

For each irreducible representation ρ with $\mathrm{Mult}(\rho, \rho_c^\circ) > 0$, we apply Propositions 5.1 and 5.3 with $F := F_{c,a}^{H_1,H_2}$ to obtain

$$\|\widehat{F}(\rho)\|_{S^q}^q \ll \frac{c^3}{c^{3-o(1)}/d} \sum_{\substack{g_1,\ldots,g_q \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \\ g_1\cdots g_q \in \Gamma_c(d)}} F(g_1)\overline{F}(g_2^{-1})\cdots F(g_{q-1})\overline{F}(g_q^{-1})\chi(g_1\cdots g_q),$$

where $\chi := \mathrm{Tr}\rho$. In fact, by Proposition 5.1, the sum above is nonnegative if one replaces χ with any irreducible character of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$. Summing over all such characters $\chi' = \mathrm{Tr}\rho'$ with weight $\mathrm{Mult}(\rho', \rho_c)$ (which is at least 1 when $\rho' = \rho$), we find that

$$\|\widehat{F}(\rho)\|_{S^q}^q \ll c^{o(1)} d \sum_{\substack{g_1,\ldots,g_q \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z}) \\ g_1\cdots g_q \in \Gamma_c(d)}} F(g_1)\overline{F}(g_2^{-1})\cdots F(g_{q-1})\overline{F}(g_q^{-1})\chi_c(g_1\cdots g_q).$$

Here ρ_c is the original permutation representation from Definition 4.1. In light of Lemma 5.2, we ought to split the sum above based on the largest $\tilde{d} \mid c$ such that $g_1\cdots g_q \in \gamma\Gamma_c(\tilde{d})$ for some $\gamma \in \mathbb{Z}/c\mathbb{Z}$ with $\gamma^2 = 1$; note that by (3.12), this is equivalent to the equation $g_1\cdots g_q = I$ in $\mathrm{PSL}_2(\mathbb{Z}/\tilde{d}\mathbb{Z})$. Then from the triangle inequality, Lemma 5.2, and the divisor bound for $\tilde{d}$, we find that

$$\|\widehat{F}(\rho)\|_{S^q}^q \ll c^{o(1)} d \max_{\substack{d|\tilde{d}|c \\ \tilde{f}^2|c\tilde{d}}} \tilde{f} \sum_{g_1,\ldots,g_q \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})} |F(g_1)F(g_2^{-1})\cdots F(g_{q-1})F(g_q^{-1})| \mathbb{1}_{g_1\cdots g_q = I \text{ in } \mathrm{PSL}_2(\mathbb{Z}/\tilde{d}\mathbb{Z})}. \tag{5.3}$$

Now recalling (4.14), we can expand

$$|F(g)| = |F_{c,a}^{H_1,H_2}(g)| \ll \frac{1}{H_1H_2} \sum_{\substack{|h_1| \le H_1 \\ |h_2| \le H_2}} \mathbb{1}_{g = T^{\bar{a}h_1}ST^{h_2}}.$$

The conclusion follows by plugging this into (5.3), and noting that as h, h' vary in $[-H, H] \cap \mathbb{Z}$, the difference $h - h'$ varies in $[-2H, 2H] \cap \mathbb{Z}$, each value being attained $O(H)$ times. ☐

6. Counting solutions in $\mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})$

We now develop the final ingredient towards Theorem 1.2, as outlined in Section 2.4. Given $c, q \in \mathbb{Z}_+$ with q even, $a_1, a_2 \in (\mathbb{Z}/c\mathbb{Z})^\times$, and $1 \le H_1 \le H_2 \ll c$, we will count solutions in $(h_1, \ldots, h_q) \in \mathbb{Z}$

to the system

$$\begin{cases} T^{a_1h_1}ST^{a_2h_2}S\cdots T^{a_1h_{q-1}}T^{a_2h_q}S = I \text{ in } \mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z}), \\ |h_i| \le H_j, \text{ for all } i,j \text{ with } i \equiv j \pmod 2). \end{cases} \tag{6.1}$$

In particular, the ranges of h_i alone produce the trivial bound $\ll_q (H_1H_2)^{q/2}$ for the number of solutions. Focusing on the case $a_1 = a_2 = 1$, below are some classes of solutions to (6.1):

(i). *Integer solutions (1).* If $h_1 = h_{\frac{q}{2}+1} = 0$ (and similarly for cyclic permutations of this case), noting that $S^2 = I$ in $\mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})$, (6.1) becomes

$$T^{h_2}S\cdots ST^{h_{\frac{q}{2}}} = T^{-h_q}ST^{-h_{q-1}}S\cdots ST^{-h_{\frac{q}{2}+2}}.$$

This has $\asymp_q H_1^{\lfloor (q-2)/4\rfloor}H_2^{\lceil (q-2)/4\rceil}$ diagonal solutions with $h_i = h_{q-i+2}$, which actually give solutions to (6.1) in $\mathrm{PSL}_2(\mathbb{Z})$.

(ii). *Integer solutions (2).* If $h_1 = h_3 = \cdots = h_{q-1} = 0$, (6.1) becomes

$$T^{h_2+h_4+\cdots+h_q} = I \qquad \iff \qquad h_2 + h_4 + \cdots + h_q \equiv 0 \pmod c.$$

This has $\asymp_q H_2^{(q-2)/2}$ solutions, which supersedes the diagonal contribution from (i).

(iii). *Generic terms.* The product $T^{h_1}ST^{h_2}S\cdots T^{h_q}S$ can take $\approx c^3$ values in $\mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})$. If each matrix is attained roughly the same number of times (and such equidistribution ought to happen for large enough q), this gives an expected number of $\approx c^{-3}(H_1H_2)^{q/2}$ solutions.

These heuristics can be formalized to produce a lower bound, imposing a limitation on our methods.

Lemma 6.1. *Let $c, q \in \mathbb{Z}_+$ with q even, $a_1, a_2 \in (\mathbb{Z}/c\mathbb{Z})^\times$, and $1 \le H_1 \le H_2 \ll c$. The number of solutions $(h_1, \ldots, h_q) \in \mathbb{Z}^q$ to (6.1) is at least*

$$\gg_q H_2^{(q-2)/2} + \frac{(H_1H_2)^{q/2}}{c^3}. \tag{6.2}$$

Proof. The lower bound by the first term in (6.2) follows by considering the aforementioned integer solutions with $h_1 = h_3 = \cdots = h_{q-1} = 0$ and $h_2 + h_4 + \cdots + h_q = 0$. To obtain a lower bound by the second term in (6.2), we let $k := \frac{q+2}{2}$, write (H_j, a_j) for (H_1, a_1) or (H_2, a_2) depending on whether j is odd or even, and apply Cauchy–Schwarz to obtain

$$\begin{aligned}
\sum_{\substack{h_1,\ldots,h_k\in\mathbb{Z} \\ |h_i|\le\frac12 H_j \\ \forall i\equiv j \pmod 2}} 1 &= \sum_{g\in\mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})} \sum_{\substack{h_1,\ldots,h_k\in\mathbb{Z} \\ |h_i|\le\frac12 H_j \\ \forall i\equiv j \pmod 2}} \mathbb{1}_{T^{a_1h_1}S\cdots T^{a_kh_k}S=g} \\
&\ll c^{3/2}\left(\sum_{g\in\mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})}\left(\sum_{\substack{h_1,\ldots,h_k\in\mathbb{Z} \\ |h_i|\le\frac12 H_j \\ \forall i\equiv j \pmod 2}} \mathbb{1}_{T^{a_1h_1}S\cdots T^{a_kh_k}S=g}\right)^2\right)^{1/2} \\
&= c^{3/2}\left(\sum_{\substack{h_1,\ldots,h_k\in\mathbb{Z} \\ h_1',\ldots,h_k'\in\mathbb{Z} \\ |h_i|,|h_i'|\le\frac12 H_j \\ \forall i\equiv j \pmod 2}} \mathbb{1}_{T^{a_1h_1}S\cdots T^{a_kh_k}S=T^{a_1h_1'}S\cdots T^{a_kh_k'}S \text{ in } \mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})}\right)^{1/2}.
\end{aligned}$$

One can rewrite the last equation $T^{a_1h_1}S\cdots T^{a_kh_k}S = T^{a_1h_1'}S\cdots T^{a_kh_k'}S$ in $\mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})$ as

$$T^{a_1(h_1-h_1')}ST^{a_2h_2}S\cdots T^{a_{k-1}h_{k-1}}ST^{a_k(h_k-h_k')}ST^{-a_{k-1}h_{k-1}'}S\cdots T^{-a_1h_1'}S = I.$$

Comparing this with (6.1), recalling that $k = \frac{q+2}{2}$, and noting that $h_1 - h_1'$ takes each value in $\mathbb{Z} \cap [-H_1, H_1]$ at most $O(H_1)$ times (and similarly for $h_k - h_k'$), we conclude that the desired count of solutions is at least

$$\gg_q \frac{1}{c^3 H_1 H_k} \left(\sum_{\substack{h_1, \ldots, h_k \in \mathbb{Z} \\ |h_i| \le \frac{1}{2} H_j \\ \forall i \equiv j \pmod 2}} 1 \right)^2 \gg_q \frac{1}{c^3 H_1 H_k} H_1^{2\lceil k/2 \rceil} H_2^{2\lfloor k/2 \rfloor} = \frac{H_1^{k-1} H_2^{k-1}}{c^3}.$$

Since $k - 1 = \frac{q}{2}$, this completes our proof. □

Optimistically, we may expect the lower bound in Lemma 6.1 to be essentially sharp.

Conjecture 6.2. *Let $c, q \in \mathbb{Z}_+$ with q even. For all $a_1, a_2 \in (\mathbb{Z}/c\mathbb{Z})^\times$ and $1 \le H_1 \le H_2 \ll c$, the number of solutions $(h_1, \ldots, h_q) \in \mathbb{Z}^q$ to (6.1) is at most*

$$\ll_q c^{o(1)} \left(H_2^{(q-2)/2} + \frac{(H_1 H_2)^{q/2}}{c^3} \right). \tag{6.3}$$

Remark. When c is large enough in terms of H_1, H_2, q (so in particular, the first term in (6.3) dominates), Conjecture 6.2 becomes a statement about $\mathrm{PSL}_2(\mathbb{Z})$, which follows from a somewhat tedious combinatorial computation. When q is large enough in terms of H_1, H_2, c (so in particular, the second term in (6.3) dominates), Conjecture 6.2 can be attacked using expansion methods; see [37, Lemma 53 and Theorem 50] for the case when c is prime. However, note that (6.3) saves at best c^3 over the trivial bound of $(H_1 H_2)^{q/2}$, and this saving becomes $c^{3/q}$ in our final bounds; therefore, using a large value of q ultimately produces a quantitatively-weak power saving. On the other hand, if q is too small, then combining Conjecture 6.2 with Proposition 5.5 gives information about a small moment of singular values, which produces a weak bound for the top singular value.

Because of this, Conjecture 6.2 is most relevant in the median range when $q \asymp 1$, say $q \in \{6, 8, 10\}$, and $H_1, H_2 \in [\sqrt{c}, c]$. It seems very difficult to fully establish Conjecture 6.2 in these cases, but we can nevertheless make some partial progress towards it. When c is prime, the cases $q \in \{4, 6\}$ are also related to some computations of Shkredov [38, Lemma 15].

Lemma 6.3. *Conjecture 6.2 holds if $q = 2$ or $q = 4$.*

Proof. When $q = 2$, the equation in (6.1) reads $T^{a_1 h_1} S = S T^{-a_2 h_2}$, which implies the entry-wise congruence

$$\begin{pmatrix} a_1 h_1 & -1 \\ 1 & 0 \end{pmatrix} \equiv \gamma \begin{pmatrix} 0 & -1 \\ 1 & -a_2 h_2 \end{pmatrix} \pmod{c},$$

for some $\gamma \in \mathbb{Z}/c\mathbb{Z}$ with $\gamma^2 = 1$. This actually forces $\gamma = 1$, and $h_1, h_2 \equiv 0 \pmod{c}$. Since $H_1, H_2 \ll c$, we obtain $O(1)$ choices of h_1, h_2, which matches the bound from (6.3).

When $q = 4$, the equation in (6.1) reads $T^{a_1 h_1} S T^{a_2 h_2} S = S T^{-a_2 h_4} S T^{-a_1 h_3}$, which translates to

$$\begin{pmatrix} a_1 a_2 h_1 h_2 - 1 & -a_1 h_1 \\ a_2 h_2 & -1 \end{pmatrix} \equiv \gamma \begin{pmatrix} -1 & a_1 h_3 \\ -a_1 h_4 & a_1 a_2 h_3 h_4 - 1 \end{pmatrix} \pmod{c},$$

for some $\gamma \in \mathbb{Z}/c\mathbb{Z}$ with $\gamma^2 = 1$. Since there are $c^{o(1)}$ such values[5] of γ, we may fix γ up to an acceptable loss. To establish the desired bound of $O(c^{o(1)} H_2)$ for the number of solutions (h_1, h_2, h_3, h_4) with $|h_1|, |h_3| \le H_1$ and $|h_2|, |h_4| \le H_2$, we split into three cases.

[5] This follows from a local computation via Hensel's lemma, and the divisor bound.

Case 1: $h_1 = 0$. This forces $h_3 \equiv 0 \pmod{c}$, and each choice of h_2 induces a unique residue of $h_4 \pmod{c}$. Since $H_1, H_2 \ll c$, this gives a total of $O(c^{o(1)} H_2)$ solutions.

Case 2: $h_2 = 0$. This forces $h_4 \equiv 0 \pmod{c}$, and each choice of h_1 induces a unique residue of $h_3 \pmod{c}$. Since $H_1, H_2 \ll c$, this gives a total of $O(c^{o(1)} H_1)$ solutions, and recall $H_1 \le H_2$.

Case 3: $h_1 h_2 \neq 0$. Then the congruence $a_1 a_2 h_1 h_2 - 1 \equiv -\gamma \pmod{c}$ fixes the residue of $h_1 h_2 \pmod{c}$, leaving $O(1 + \frac{H_1 H_2}{c})$ possible values of $h_1 h_2$, each of which gives $O(c^{o(1)})$ choices of h_1, h_2 by the divisor bound. This gives a total of $\ll c^{o(1)}(1 + \frac{H_1 H_2}{c}) \ll c^{o(1)} H_2$ solutions. $\square$

Proposition 6.4 (Combinatorial count for $q = 6$)**.** *Let $c \in \mathbb{Z}_+$, $a_1, a_2 \in (\mathbb{Z}/c\mathbb{Z})^\times$, and $1 \le H_1 \le H_2 \ll c$. The number of solutions $(h_1, \dots, h_6) \in \mathbb{Z}^6$ to (6.1) with $q = 6$ is at most*

$$\ll c^{o(1)} \left(H_2^2 + \frac{(H_1 H_2)^2}{c} \right). \tag{6.4}$$

Remark. Conjecture 6.2 would replace the second term in (6.4) with $\frac{(H_1 H_2)^3}{c^3}$. In particular, Proposition 6.4 establishes Conjecture 6.2 when $q = 6$ and either $H_1^2 \ll c$ or $H_1, H_2 \asymp c$.

Proof of Proposition 6.4. To simplify the exposition, we focus on the case $a_1 = a_2 = 1$. The proof is almost completely unchanged when $a_1, a_2 \in (\mathbb{Z}/c\mathbb{Z})^\times$ are arbitrary. We may then write the equation in (6.1) (with $q = 6$) as

$$T^{h_1} S T^{h_2} S T^{h_3} S T^{h_4} S = S T^{-h_6} S T^{-h_5}.$$

A short computation brings this to the entry-wise congruence

$$\begin{pmatrix} h_1 h_2 h_3 h_4 - h_1 h_4 - h_3 h_4 - h_1 h_2 + 1 & -h_1 h_2 h_3 + h_1 + h_3 \\ h_2 h_3 h_4 - h_2 - h_4 & -h_2 h_3 + 1 \end{pmatrix} \equiv \gamma \begin{pmatrix} -1 & h_5 \\ -h_6 & h_5 h_6 - 1 \end{pmatrix} \pmod{c}$$

for some $\gamma \in \mathbb{Z}/c\mathbb{Z}$ with $\gamma^2 = 1$. As before, since there are $c^{o(1)}$ possible values of γ, we may as well regard γ as fixed. Since both sides have determinant 1, this is actually a system of three congruences:

$$\begin{cases} 1 - h_2 h_3 \equiv \gamma(h_5 h_6 - 1) \pmod{c}, \\ h_1(1 - h_2 h_3) + h_3 \equiv \gamma h_5 \pmod{c}, \\ h_4(1 - h_2 h_3) + h_2 \equiv \gamma h_6 \pmod{c}. \end{cases} \tag{6.5}$$

Our argument now requires some casework.

Case 1: One has $h_j = 0$ for some $j \in \{1, \dots, 6\}$. Since the original equation can also be written as $T^{h_{j-1}} S T^{h_j} S T^{h_{j+1}} S T^{h_{j+2}} S T^{h_{j+3}} S T^{h_{j+4}} S = I$ in $\mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z})$ (viewing indices modulo 6), we may assume without loss of generality that $j = 2$, up to potentially swapping H_1 and H_2 in the final bound (so we momentarily forget that $H_1 \le H_2$). So let us say $h_2 = 0$, which reduces (6.5) to

$$\begin{cases} 1 \equiv \gamma(h_5 h_6 - 1) \pmod{c}, \\ h_1 + h_3 \equiv \gamma h_5 \pmod{c}, \\ h_4 \equiv \gamma h_6 \pmod{c}. \end{cases} \tag{6.6}$$

Subcase 1.1: One has $h_5 = 0$. Then for any values of h_1 and h_6, the system in (6.6) leaves only $O(1)$ possibilities for h_3, h_4 (since $H_1, H_2 \ll c$). This gives $O(H_1 H_2)$ solutions.

Subcase 1.2: One has $h_6 = 0$. Then for any values of h_1 and h_5, the system in (6.6) leaves only $O(1)$ possibilities for h_3, h_4. This gives $O(H_1^2)$ solutions.

Subcase 1.3: One has $h_5 h_6 \neq 0$. Then the first congruence in (6.6) fixes $h_5 h_6 \pmod{c}$, leaving $1 + \frac{H_1 H_2}{c}$ possibilities for the nonzero integer $h_5 h_6$, each of which gives $O(c^{o(1)})$ possible values for h_5, h_6 by the divisor bound. Once h_5 and h_6 are fixed, each value of h_1 produces $O(1)$ final solutions. This gives $(1+\frac{H_1H_2}{c})H_1$ solutions.

From Case 1, we obtain a total number of solutions of

$$\ll c^{o(1)}\left(H_1^2+H_2^2+\left(1+\frac{H_1H_2}{c}\right)(H_1+H_2)\right),$$

which is $O(c^{o(1)}H_2^2)$ once we remember that $H_1 \le H_2$ and $H_1 \ll c$. This is acceptable in (6.4).

Case 2: One has $h_j \neq 0$ for all $j \in \{1,\dots,6\}$. We fix $d := (1-h_2h_3, c) = (h_5h_6-1, c)$ up to an acceptable $O(c^{o(1)})$ loss. Since $h_5h_6 \equiv 1 \pmod d$, there are $O(1+\frac{H_1H_2}{c})$ ways to pick the nonzero integer h_5h_6, each of which gives $O(c^{o(1)})$ ways to pick h_5, h_6 by the divisor bound.

Once h_5, h_6 are fixed, we pick h_2, h_3 subject to the system

$$\begin{cases} 1-h_2h_3 \equiv \gamma(h_5h_6-1) \pmod c, \\ h_2 \equiv \gamma h_6 =: r \pmod d, \end{cases} \tag{6.7}$$

which follows from (6.5). We can do this in either of the following ways:

- Pick the nonzero integer h_2h_3 subject to its residue mod c (due to the first congruence in (6.7)) in $O(1+\frac{H_1H_2}{c})$ ways, and then h_2, h_3 in $O(c^{o(1)})$ ways by the divisor bound.
- For each choice of h_2 with $|h_2| \le H_2$ and $h_2 \equiv r \pmod d$, pick h_3 subject to its residue mod $\frac{c}{(h_2,c)}$ (again, due to the first congruence in (6.7)) in $O(1+\frac{H_1}{c}(h_2,c))$ ways.

Finally, once h_5, h_6, h_2, h_3 are fixed, (6.5) determines the residues of h_1 and h_4 modulo $\frac{c}{d}$, so there are $(1+\frac{H_1}{c}d)(1+\frac{H_2}{c}d)$ choices of h_1, h_4. From Case 2, we obtain a total number of solutions of

$$\begin{aligned} \ll c^{o(1)} \max_{d\mid c} \underbrace{\left(1+\frac{H_1H_2}{d}\right)}_{\text{from picking } h_5,h_6} & \underbrace{\min\left(1+\frac{H_1H_2}{c}, \max_{r\in(\mathbb{Z}/d\mathbb{Z})^\times} \sum_{\substack{|h_2|\le H_2 \\ h_2\equiv r \pmod d}} \left(1+\frac{H_1}{c}(h_2,c)\right)\right)}_{\text{from picking } h_2,h_3} \\ & \times \underbrace{\left(1+\frac{H_1}{c}d\right)}_{\text{from picking } h_1} \underbrace{\left(1+\frac{H_2}{c}d\right)}_{\text{from picking } h_4}. \end{aligned} \tag{6.8}$$

To bound the sum over h_2, we write

$$\begin{aligned} \sum_{\substack{|h_2|\le H_2 \\ h_2\equiv r \pmod d}} (h_2,c) &\le \sum_{\substack{g\mid c \\ (g,d)=1}} g \sum_{\substack{|h_2|\le H_2 \\ g\mid h_2\equiv r \pmod d}} 1 \\ &= \sum_{\substack{g\mid \frac{c}{d} \\ (g,d)=1}} g \sum_{\substack{|h_2'|\le \frac{H_2}{g} \\ h_2'\equiv \bar{g}r \pmod d}} 1 \\ &\ll \sum_{\substack{|g|\le \frac{c}{d} \\ (g,d)=1}} g\left(1+\frac{H_2}{gd}\right) \ll c^{o(1)}\left(\frac{c}{d}+\frac{H_2}{d}\right), \end{aligned}$$

and the term $\frac{H_2}{d}$ can be omitted since $H_2 \ll c$. Plugging this into (6.8) gives a total count of

$$\ll c^{o(1)} \max_{d\mid c}\left(1+\frac{H_1H_2}{d}\right)\left(1+\frac{H_2}{c}d\right)\left(1+\frac{H_1}{c}d\right)\min\left(1+\frac{H_1H_2}{c}, 1+\frac{H_2}{d}+\frac{H_1}{c}\cdot\frac{c}{d}\right)$$

Since $H_1 \le H_2$, the final term of $\frac{H_1}{c} \cdot \frac{c}{d} = \frac{H_1}{d}$ can be omitted. The bound above now becomes

$$\begin{aligned}
&\ll c^{o(1)} \max_{d|c} \left(1 + \frac{H_1 H_2}{d}\right)\left(1 + \frac{H_2}{c} d\right) \max\left(\frac{H_1}{c} d, 1\right)\left(1 + \frac{H_2}{d} \min\left(\frac{H_1}{c} d, 1\right)\right) \\
&= c^{o(1)} \max_{d|c} \left(1 + \frac{H_1 H_2}{d}\right)\left(1 + \frac{H_2}{c} d\right)\left(\max\left(\frac{H_1}{c} d, 1\right) + \frac{H_2}{d}\left(\frac{H_1}{c} d \cdot 1\right)\right) \\
&\le c^{o(1)} \max_{d|c} \left(1 + \frac{H_1 H_2}{d}\right)\left(1 + \frac{H_2}{c} d\right)\left(1 + \frac{H_1}{c} d + \frac{H_1 H_2}{c}\right).
\end{aligned}$$

After expanding the expression inside the maximum, each term is either strictly increasing, constant, or strictly decreasing in $d \in [1, c]$, so each term is maximized when $d = 1$ or $d = c$. It follows that we can bound the maximum over $d \mid c$, up to a constant, by looking only at the extreme points $d = 1$ and $d = c$. This gives a total count of

$$\begin{aligned}
\ll c^{o(1)} \max\left(H_1 H_2 \left(1 + \frac{H_2}{c}\right)\left(1 + \frac{H_1}{c} + \frac{H_1 H_2}{c}\right), \left(1 + \frac{H_1 H_2}{c}\right) H_2 \left(H_1 + \frac{H_1 H_2}{c}\right)\right) \\
\ll c^{o(1)} \max\left(H_1 H_2 \left(1 + \frac{H_1 H_2}{c}\right), \left(1 + \frac{H_1 H_2}{c}\right) H_2 H_1\right),
\end{aligned}$$

where, to reach the last line, we used that $H_1, H_2 \ll c$. This establishes the desired bound. $\square$

Finally, we may remove the restriction that $H_1, H_2 \ll c$ from the counting results in this section up to some additional factors.

Corollary 6.5. *Let $c \in \mathbb{Z}_+$, $a_1, a_2 \in (\mathbb{Z}/c\mathbb{Z})^\times$, and $1 \le H_1 \le H_2$. Then the number of solutions to (6.1) with $q = 4$ is at most*

$$\ll c^{o(1)} \left(1 + \frac{H_1^2}{c^2}\right)\left(1 + \frac{H_2}{c}\right) H_2, \tag{6.9}$$

and the number of solutions to (6.1) with $q = 6$ is at most

$$\ll c^{o(1)} \left(1 + \frac{H_1 H_2}{c} + \frac{H_1^3 H_2}{c^3}\right) H_2^2. \tag{6.10}$$

Proof. Since the equation in (6.1) only depends on the residues of $h_1, \ldots, h_q \pmod c$, we may as well count solutions to the system

$$\begin{cases} T^{a_1 h_1'} S T^{a_2 h_2'} S \cdots T^{a_1 h_{q-1}'} T^{a_2 h_q'} S = I \text{ in } \mathrm{PSL}_2(\mathbb{Z}/c\mathbb{Z}), \\ |h_i'| \le \min(H_j, c), \text{ for all } i, j \text{ with } i \equiv j \pmod 2, \end{cases} \tag{6.11}$$

and multiply the final count by a factor of $\ll_q (1 + \frac{H_1}{c})^{q/2}(1 + \frac{H_2}{c})^{q/2}$ (indeed, each solution $(h_1', \ldots, h_q')$ to (6.11) induces at most this many solutions $(h_1, \ldots, h_q)$ to (6.1) with the same residues modulo c, and all solutions to (6.1) can be obtained this way).

If $q = 4$, Lemma 6.3 applied for $\min(H_1, c)$ and $\min(H_2, c)$ gives a total number of solutions of

$$\ll c^{o(1)} \left(1 + \frac{H_1}{c}\right)^2 \left(1 + \frac{H_2}{c}\right)^2 \min(H_2, c),$$

and the bound in (6.9) follows by noting that $\min(H_2, c) \asymp H_2 \left(1 + \frac{H_2}{c}\right)^{-1}$.

Similarly, if $q = 6$, Proposition 6.4 applied for $\min(H_1, c)$ and $\min(H_2, c)$ gives a total count of

$$\begin{aligned}
&\ll c^{o(1)} \left(1 + \frac{H_1}{c}\right)^3 \left(1 + \frac{H_2}{c}\right)^3 \left(\min(H_2, c)^2 + \frac{\min(H_1, c)^2 \min(H_2, c)^2}{c}\right) \\
&\ll c^{o(1)} \left(1 + \frac{H_1^3}{c^3}\right) \left(1 + \frac{H_2}{c}\right) H_2^2 + c^{o(1)} \left(1 + \frac{H_1}{c}\right) \left(1 + \frac{H_2}{c}\right) \frac{H_1^2 H_2^2}{c} \\
&\ll c^{o(1)} \left(1 + \frac{H_2}{c} + \frac{H_1^3}{c^3} + \frac{H_1^3 H_2}{c^4}\right) H_2^2 + c^{o(1)} \left(1 + \frac{H_2}{c} + \frac{H_1 H_2}{c^2}\right) \frac{H_1^2 H_2^2}{c}.
\end{aligned}$$

We note that the third and fourth terms in the first parenthesis above can be ignored: their contribution to the final bound is $\frac{H_1^3 H_2^2}{c^3} + \frac{H_1^3 H_2^3}{c^4}$, which is superseded by the contribution of $\frac{H_1^3 H_2^3}{c^3}$ from the third term in the second parenthesis. This gives a total count of

$$\ll c^{o(1)} \left(1 + \frac{H_2}{c} + \frac{H_1^2}{c} + \frac{H_1^2 H_2}{c^2} + \frac{H_1^3 H_2}{c^3}\right) H_2^2. \tag{6.12}$$

This bounded by (6.10), noting that $\frac{H_1^2 H_2}{c^2}$ is the geometric mean of $\frac{H_1 H_2}{c}$ and $\frac{H_1^3 H_2}{c^3}$. □

7. Bilinear forms with Kloosterman sums

We now combine the work in Sections 4 to 6, to deduce our main results from Theorems 1.1 and 1.2.

7.1. **Composite moduli.** Here we prove a generalization of Theorem 1.2, which allows for larger values of M, N. We state our upper bound in two ways, to facilitate comparison with (1.2).

Theorem 7.1. *Let $c = dd'e$ for some $d, d', e \in \mathbb{Z}_+$ with $d' \mid d$ and $(d, e) = 1$, and $f \le \sqrt{cd}$ be the largest integer with $f^2 \mid cd$. Let $\mathcal{I}, \mathcal{J} \subset \mathbb{Z}$ be intervals of lengths $|\mathcal{I}| = M$, $|\mathcal{J}| = N$, with*[6] *$1 \le N \le M \le c$. Then for any complex sequences $(\alpha_m)_{m \in \mathcal{I}}$, $(\beta_n)_{n \in \mathcal{J}}$ and $a \in (\mathbb{Z}/c\mathbb{Z})^\times$, one has*

$$\begin{aligned}
\sum_{\substack{m \in \mathcal{I}, n \in \mathcal{J} \\ (m,n,c)=1}}\sum \alpha_m \beta_n S(am, n; c) &\ll \|\alpha\| \|\beta\| c^{1+o(1)} \left(\frac{dM^3 N}{c^3} + \frac{fM^2}{c^2} + \frac{f}{d^2}\right)^{\frac{1}{6}} \\
&= \|\alpha\| \|\beta\| c^{o(1)} \sqrt{MNc} \left(\frac{d}{N^2} + \frac{fc}{MN^3} + \frac{fc^3}{d^2 M^3 N^3}\right)^{\frac{1}{6}}.
\end{aligned}$$

Example 7.2. *Suppose $M \asymp N$. The smallest value of N for which Theorem 7.1 can beat the Weil bound in (1.2) is $N = c^{2/5+o(1)}$, attained, e.g., when $c = pq$ where p, q are distinct primes with $p \asymp q^{3/2}$. The largest value of N for which Theorem 7.1 can beat the Fourier-theoretic bound in (1.2) is $N = c^{3/4-o(1)}$, attained when c has a divisor $d = c^{o(1)}$ such that c/d is square-free.*

Remark. Additional savings are possible in Theorem 7.1 in the unbalanced range $M > N$. Firstly, the bound in (6.10) can be refined to (6.12), but we omit this optimization for the sake of simplicity. Secondly and more substantially, bounding the largest singular value of an $M \times N$ matrix by the sixth moment of its singular values (as we do) can be particularly lossy if $M > N$, since then the singular values often exhibit concentration near their maximum; one can try to amend this by subtracting a suitable main term from the sixth moment, as in [18, Lemma 4.2].

Proof of Theorem 7.1. Let $\varepsilon > 0$ and $H_1 := c^{1+\varepsilon} M^{-1}$, $H_2 := c^{1+\varepsilon} N^{-1}$. Since $M \ge N$, we have $H_1 \le H_2$. Let q be an even positive integer.

[6] The assumption that $N \le M$ is only included to shorten the statement of Theorem 7.1; one can of course swap m and n in the bilinear sum, up to swapping M and N in the upper bound.

Then by the characterization of operator norms from (3.4), Corollary 4.11 (using the notation from (4.13) and (4.14)), the fact that $\|A\| \le \|A\|_{S^q}$ for any linear map A, and then Proposition 5.5, we have that

$$\begin{aligned}\Bigg|\sum_{\substack{m\in\mathcal{I},n\in\mathcal{J}\\(m,n,c)=1}}\sum \alpha_m\beta_n S(am,n;c)\Bigg| &\le \|\alpha\|\|\beta\|\|K_{c,a}^{\mathcal{I},\mathcal{J}}\| \\ &\le \|\alpha\|\|\beta\|\left(c^{1+2\varepsilon}\|\widehat{F}_{c,a}^{H_1,H_2}(\rho_c^\circ)\| + O_\varepsilon(c^{-100})\right) \\ &\ll_\varepsilon \|\alpha\|\|\beta\|\left(c^{1+3\varepsilon}\frac{d^{1/q}}{(H_1H_2)^{1/2}}\mathscr{S}^{1/q} + O_\varepsilon(c^{-100})\right),\end{aligned} \tag{7.1}$$

where

$$\mathscr{S} := \max_{\substack{d|\tilde{d}|c\\ \tilde{f}^2|c\tilde{d}}} \tilde{f} \sum_{\substack{h_1,\dots,h_q\in\mathbb{Z}\\ |h_i|\le 2H_j\\ \forall i\equiv j \pmod 2}} \mathbb{1}_{T^{\bar{a}h_1}ST^{h_2}S\cdots T^{\bar{a}h_{q-1}}ST^{h_q}S=I \text{ in } \mathrm{PSL}_2(\mathbb{Z}/\tilde{d}\mathbb{Z})}.$$

The inner sum is a count of solutions to an equation of the type (6.1) with c replaced by $\tilde{d}$, so we can estimate $\mathscr{S}$ using Corollary 6.5. We will make use of the following quick fact. Recalling that f is the maximal positive integer with $f^2 \mid cd$, we claim that for all integers $k \ge 0$, one has

$$\max_{\substack{d|\tilde{d}|c\\ \tilde{f}^2|c\tilde{d}}} \frac{\tilde{f}}{\tilde{d}^k} = \begin{cases} \frac{f}{d^k}, & k \ge 1, \\ c, & k = 0. \end{cases} \tag{7.2}$$

Indeed, when one appends a prime p to $\tilde{d}$, the numerator $\tilde{f}$ can increase by at most p, so the expression $\tilde{f}/\tilde{d}^k$ cannot increase if $k \ge 1$, and the maximum is attained when $\tilde{d} = d$. On the other hand, if $k = 0$, then clearly $\tilde{f} \le \sqrt{c \cdot c} = c$, and the maximum is attained when $\tilde{d} = c$.

Now set $q = 6$. From (6.10) and (7.2), we obtain that

$$\mathscr{S} \ll_\varepsilon c^\varepsilon \max_{\substack{d|\tilde{d}|c\\ \tilde{f}^2|c\tilde{d}}} \tilde{f}\left(1 + \frac{H_1H_2}{\tilde{d}} + \frac{H_1^3H_2}{\tilde{d}^3}\right)H_2^2 \le c^\varepsilon\left(c + \frac{fH_1H_2}{d} + \frac{fH_1^3H_2}{d^3}\right)H_2^2.$$

Plugging this into (7.1), we conclude that

$$\begin{aligned}\sum_{\substack{m\in\mathcal{I},n\in\mathcal{J}\\(m,n,c)=1}}\sum \alpha_m\beta_n S(am,n;c) &\ll_\varepsilon \|\alpha\|\|\beta\|c^{1+4\varepsilon}\frac{d^{1/6}}{(H_1H_2)^{1/2}}\left(cH_2^2 + \frac{fH_1H_2^3}{d} + \frac{fH_1^3H_2^3}{d^3}\right)^{1/6}. \\ &= \|\alpha\|\|\beta\|c^{1+4\varepsilon}\left(\frac{cd}{H_1^3H_2} + \frac{f}{H_1^2} + \frac{f}{d^2}\right)^{1/6}.\end{aligned}$$

The desired bound follows by recalling that $H_1 = c^{1+\varepsilon}M^{-1}$, $H_2 = c^{1+\varepsilon}N^{-1}$. □

Remark. Using (6.9) with $q = 4$ instead of (6.10) with $q = 6$ in the proof above leads to a final bound of

$$\begin{aligned}\sum_{\substack{m\in\mathcal{I},n\in\mathcal{J}\\(m,n,c)=1}}\sum \alpha_m\beta_n S(am,n;c) &\ll c^{o(1)}\|\alpha\|\|\beta\|c\left(\frac{dM^2N}{c^2} + \frac{fM^2}{c^2} + \frac{fN}{dc} + \frac{f}{d^2}\right)^{1/4} \\ &= c^{o(1)}\|\alpha\|\|\beta\|\sqrt{MNc}\left(\frac{d}{N} + \frac{f}{N^2} + \frac{fc}{dM^2N} + \frac{fc^2}{d^2M^2N^2}\right)^{1/4},\end{aligned}$$

which is weaker than Theorem 7.1 in the main ranges of interest.

Proof of Theorem 1.2. Note that the result holds trivially if $c = O(1)$. Since $M, N \ll c^{1/2+o(1)}$ and the result is symmetric in M, N, we can assume without loss of generality that $N \le M \le c$. One can then apply Theorem 7.1, and since $M, N \ll c^{1/2+o(1)}$, the upper bound becomes

$$\|\alpha\|\|\beta\| c^{1+o(1)} \left(\frac{d}{c} + \frac{f}{c} + \frac{f}{d^2} \right)^{\frac{1}{6}}.$$

The first term can be omitted in light of the bound $d \le f$ (since $d^2 \mid cd$). $\square$

7.2. **Near-prime moduli.** Building towards an unconditional result for general moduli, we need to slightly develop the result of Kowalski–Michel–Sawin [24] from Theorem 3.4.

Corollary 7.3 (Kowalski–Michel–Sawin bounds for near-primes)**.** *Let $c = pq$ where p is a prime, $q \in \mathbb{Z}_+$, and $p \nmid q$. Let $M, N \in \mathbb{Z}$ be integers such that $1 \le N \le M \le c$. Then for any complex sequences $(\alpha_m)_{m \le M}$ and $(\beta_n)_{n \le N}$, and any $a \in (\mathbb{Z}/c\mathbb{Z})^\times$, one has*

$$\sum_{\substack{m=1 \\ (m,n,c)=1}}^{M} \sum_{n=1}^{N} \alpha_m \beta_n S(am, n; c) \ll \|\alpha\|\|\beta\| c^{o(1)} \sqrt{MNc} \left(N^{-\frac{1}{2}} q + (MN)^{-\frac{3}{16}} c^{\frac{11}{64}} q^{\frac{53}{64}} \right).$$

Proof. Using the twisted multiplicativity of Kloosterman sums and splitting the sums over m, n according to their residues modulo q, we can write

$$\begin{aligned} \sum_{\substack{m=1 \\ (m,n,pq)=1}}^{M} \sum_{n=1}^{N} \alpha_m \beta_n S(am, n; pq) &= \sum_{\substack{m=1 \\ (m,n,pq)=1}}^{M} \sum_{n=1}^{N} \alpha_m \beta_n S(a\overline{q}^2 m, n; p) S(a\overline{p}^2 m, n; q) \\ &= \sum_{\substack{r=1 \\ (r,s,q)=1}}^{q} \sum_{s=1}^{q} S(a\overline{p}^2 r, s; q) \sum_{\substack{m=1 \\ (m,n,p)=1}}^{M} \sum_{n=1}^{N} \alpha_{m,r} \beta_{n,s} S(a\overline{q}^2 m, n; p), \end{aligned} \tag{7.3}$$

where

$$\alpha_{m,r} := \alpha_m \mathbb{1}_{m \equiv r \pmod{q}}, \qquad \beta_{n,s} := \beta_n \mathbb{1}_{n \equiv s \pmod{q}}.$$

First, we consider the contribution to (7.3) of those m, n with $p \mid m$ (and $p \nmid n$) or $p \mid n$ (and $p \nmid m$). By the Weil and Ramanujan bounds from Lemmas 3.2 and 3.3, this contribution is

$$\begin{aligned} &\ll \sum_{\substack{r=1 \\ (r,s,q)=1}}^{q} \sum_{s=1}^{q} q^{\frac{1}{2}+o(1)} \sum_{\substack{m=1 \\ (m,n,p)=1}}^{M} \sum_{n=1}^{N} |\alpha_{m,r} \beta_{n,s}| \ll q^{\frac{1}{2}+o(1)} \left(\sum_{m=1}^{M} |\alpha_m| \right) \left(\sum_{n=1}^{N} |\beta_n| \right) \\ &\ll \|\alpha\|\|\beta\| q^{o(1)} \sqrt{MNq}, \end{aligned}$$

Plugging this into (7.3) and applying the Weil bound from Lemma 3.3, we find that

$$\begin{aligned} \sum_{\substack{m=1 \\ (m,n,pq)=1}}^{M} \sum_{n=1}^{N} \alpha_m \beta_n S(am, n; pq) \ll \sum_{\substack{r=1 \\ (r,s,q)=1}}^{q} \sum_{s=1}^{q} q^{\frac{1}{2}+o(1)} &\left| \sum_{\substack{m=1 \\ p \nmid mn}}^{M} \sum_{n=1}^{N} \alpha_{m,r} \beta_{n,s} S(a\overline{q}^2 m, n; p) \right| \\ &+ \|\alpha\|\|\beta\| q^{o(1)} \sqrt{MNq}. \end{aligned} \tag{7.4}$$

The last bilinear sum over m, n is now almost in the correct shape to apply Theorem 3.4. Indeed, splitting it according to the residues of m and n modulo p, we can rewrite it as

$$\sum_{\substack{m=1 \\ p\nmid mn}}^{M}\sum_{n=1}^{N} \alpha_{m,r}\beta_{n,s} S(a\overline{q}^2 m, n; p) = \sum_{m'=1}^{\min(M,p-1)} \sum_{n'=1}^{\min(N,p-1)} \alpha_{m',r}\beta_{n',s} S(a\overline{q}^2 m', n'; p), \tag{7.5}$$

where

$$\alpha_{m',r} := \sum_{\substack{m\le M \\ m\equiv m' \pmod p}} \alpha_{m,r} = \sum_{\substack{m\le M \\ m\equiv m' \pmod p \\ m\equiv r \pmod q}} \alpha_m,$$

and $\beta_{n',s}$ is defined similarly. Note that the last sum above contains at most one term, so

$$\sum_{m'=1}^{\min(M,p-1)} |\alpha_{m',r}|^2 = \sum_{m'=1}^{\min(M,p-1)} \sum_{\substack{m\le M \\ m\equiv m' \pmod p \\ m\equiv r \pmod q}} |\alpha_m|^2 \le \sum_{\substack{m\le M \\ m\equiv r \pmod q}} |\alpha_m|^2,$$

and similarly for $\sum_{n'=1}^{\min(N,p-1)} |\beta_{n',s}|^2$. Applying Theorem 3.4 (and the remark that follows it) for a bilinear sum with lengths $\min(M, p-1)$ and $\min(N, p-1)$, and using the monotonicity of the bound from Theorem 3.4 in M and N, we obtain

$$\begin{aligned}\sum_{m'=1}^{\min(M,p-1)} \sum_{n'=1}^{\min(N,p-1)} \alpha_{m',r}\beta_{n',s} S(a\overline{q}^2 m', n'; p) &\ll \sqrt{\sum_{\substack{m\le M \\ m\equiv r \pmod q}} |\alpha_m|^2 \sum_{\substack{n\le N \\ n\equiv s \pmod q}} |\beta_n|^2} \\ &\quad \times p^{o(1)}\sqrt{MNp}\left(N^{-\frac12} + (MN)^{-\frac{3}{16}} p^{\frac{11}{64}}\right).\end{aligned}$$

Plugging this into (7.5) and (7.4), and applying Cauchy–Schwarz to the sum over r, s, we find that

$$\begin{aligned}\sum_{\substack{m=1 \\ (m,n,c)=1}}^{M}\sum_{n=1}^{N} \alpha_m\beta_n S(am, n; c) &\ll c^{o(1)} q^{\frac32} \|\alpha\|\|\beta\|\sqrt{MNp}\left(N^{-\frac12} + (MN)^{-\frac{3}{16}} p^{\frac{11}{64}}\right) \\ &\quad + \|\alpha\|\|\beta\| c^{o(1)}\sqrt{MNq}.\end{aligned}$$

Finally, recalling that $pq = c$ and $M, N \le c$ (which imply $\sqrt{MNq} \le \sqrt{Mcq} \le q^{3/2}\sqrt{Mp}$), the last term can be omitted, and we arrive at the desired bound. □

7.3. **General moduli.** Finally, we prove a generalization of Theorem 1.1.

Theorem 7.4. *Let $\delta \in [0, \frac{1}{24}]$, $c, M, N \in \mathbb{Z}_+$ with $M, N \le c$. Let $\tilde M := \max(M, N)$ and $\tilde N := \min(M, N)$. Then for any complex sequences $(\alpha_m)_{m\le M}$, $(\beta_n)_{n\le N}$ and any $a \in (\mathbb{Z}/c\mathbb{Z})^\times$, one has*

$$\begin{aligned}\sum_{\substack{m=1 \\ (m,n,c)=1}}^{M}\sum_{n=1}^{N} \alpha_m\beta_n S(am, n; c) &\ll \|\alpha\|\|\beta\| c^{o(1)}\sqrt{MNc} \\ &\quad \times \left(\frac{c^{\frac{11+53\delta}{64}}}{(MN)^{\frac{3}{16}}} + \frac{c^{\frac{1-\delta}{6}}}{\tilde N^{\frac13}} + \frac{c^{\frac{4-\delta}{12}}}{\tilde M^{\frac16}\tilde N^{\frac12}} + \frac{c^{\frac{11}{24}}}{(MN)^{\frac12}}\right).\end{aligned} \tag{7.6}$$

Moreover, if $|\alpha_m| \le 1$ for all m (so $\|\alpha\| \le \sqrt{M}$), then

$$\sum_{\substack{m=1}}^{M}\sum_{\substack{n=1\\(n,c)=1}}^{N} \alpha_m\beta_n S(am,n;c) \ll \sqrt{M}\|\beta\|c^{o(1)}\sqrt{MNc} \times \left(\frac{c^{\frac{11+53\delta}{64}}}{(MN)^{\frac{3}{16}}} + \frac{c^{\frac{1-\delta}{4}}}{\tilde{N}^{\frac{1}{2}}} + \frac{1}{c^{\frac{3}{16}}} + \frac{c^{\frac{1}{8}}}{\tilde{N}^{\frac{1}{3}}} + \frac{c^{\frac{11}{24}}}{(MN)^{\frac{1}{2}}}\right). \tag{7.7}$$

Remark. The bound (7.7) actually holds without the assumption that $M, N \le c$. Indeed, Theorem 7.4 covers the case $M, N \le c$. If $M, N > c$, then applying the (first) bound from (1.2) for the sequences $(\alpha'_{m'})_{m' \le c}$, $(\beta'_n)_{n \le N}$ given by $\alpha'_{m'} := \sum_{m \equiv m' \pmod c} \alpha_m$ and $\beta'_n := \mathbb{1}_{(n,c)=1}\sum_{n \equiv n' \pmod c} \beta_n$ leads to the bound

$$\begin{aligned}\sum_{\substack{m=1}}^{M}\sum_{\substack{n=1\\(n,c)=1}}^{N} \alpha_m\beta_n S(am,n;c) &\ll \|\alpha'\|\|\beta'\|c^{1+o(1)} \ll \frac{M}{c}\sqrt{c}\sqrt{\frac{N}{c}}\|\beta\|c^{1+o(1)} \\ &\ll \sqrt{M}\|\beta\|c^{o(1)}\sqrt{MNc}\cdot c^{-3/16},\end{aligned}$$

so (7.7) still holds. If $N \le c < M$, then applying the (first) trivial bound from (1.2) for the sequences $(\alpha'_{m'})_{m' \le c}$, $(\beta'_n)_{n \le N}$ given by $\alpha'_{m'} := \sum_{m \equiv m' \pmod c} \alpha_m$ and $\beta'_n := \beta_n\mathbb{1}_{(n,c)=1}$ leads to the bound

$$\begin{aligned}\sum_{\substack{m=1}}^{M}\sum_{\substack{n=1\\(n,c)=1}}^{N} \alpha_m\beta_n S(am,n;c) &\ll \|\alpha'\|\|\beta'\|c^{1+o(1)} \ll \frac{M}{c}\sqrt{c}\|\beta\|c^{1+o(1)} \\ &\ll \sqrt{M}\|\beta\|c^{o(1)}\sqrt{MNc}\cdot \frac{c^{\frac{1-\delta}{4}}}{N^{\frac{1}{2}}},\end{aligned}$$

so (7.7) still holds. An analogous argument covers the remaining case $M \le c < N$.

Proof of Theorem 7.4. We begin by noting a quick consequence of Theorem 7.1. Suppose c has a factorization $c = dd'e$ with $d' \mid d$ and $(d,e) = 1$ such that $d \ge c^{1/2}$. Then by combining Theorem 7.1 (applied for $\tilde{M}$, $\tilde{N}$ instead of M, N) with the bounds $f \le \sqrt{cd}$ and then $d \ge c^{1/2}$, we get

$$\begin{aligned}&\sum_{\substack{m=1}}^{M}\sum_{\substack{n=1\\(m,n,c)=1}}^{N} \alpha_m\beta_n S(am,n;c) \\ &\ll \|\alpha\|\|\beta\|c^{o(1)}\sqrt{MNc}\left(\frac{d}{\tilde{N}^2} + \frac{c^{\frac{3}{2}}d^{\frac{1}{2}}}{\tilde{M}\tilde{N}^3} + \frac{c^{\frac{11}{4}}}{M^3N^3}\right)^{\frac{1}{6}} \\ &\ll \|\alpha\|\|\beta\|c^{o(1)}\sqrt{MNc}\left(\frac{d^{\frac{1}{6}}}{\tilde{N}^{\frac{1}{3}}} + \frac{c^{\frac{1}{4}}d^{\frac{1}{12}}}{\tilde{M}^{\frac{1}{6}}\tilde{N}^{\frac{1}{2}}} + \frac{c^{\frac{11}{24}}}{(MN)^{\frac{1}{2}}}\right) =: \mathcal{B}(d).\end{aligned} \tag{7.8}$$

Note that this bound $\mathcal{B}(d)$ is increasing with $d \in [c^{1/2}, c]$, and that the right-hand side of (7.6) supersedes $\mathcal{B}(c^{1-\delta})$. In particular,

$$\mathcal{B}(c^{3/4}) = \|\alpha\|\|\beta\|c^{o(1)}\sqrt{MNc}\left(\frac{c^{\frac{1}{8}}}{\tilde{N}^{\frac{1}{3}}} + \frac{c^{\frac{5}{16}}}{\tilde{M}^{\frac{1}{6}}\tilde{N}^{\frac{1}{2}}} + \frac{c^{\frac{11}{24}}}{(MN)^{\frac{1}{2}}}\right).$$

A quick computation shows that since $\delta \le \frac{1}{24}$, one has

$$\frac{c^{\frac{5}{16}}}{\tilde{M}^{\frac{1}{6}}\tilde{N}^{\frac{1}{2}}} \le \max\left(\frac{c^{\frac{11}{24}}}{(\tilde{M}\tilde{N})^{\frac{1}{2}}}, \frac{c^{\frac{1-\delta}{4}}}{\tilde{N}^{\frac{1}{2}}}\right).$$

Therefore, the right-hand side of (7.7) supersedes $\mathcal{B}(c^{3/4})$. We now split into cases depending on the factorization of the modulus c.

Case 1: c is divisible by a maximal prime power $p^k \ge c^{1-\delta}$. Then let us write $c = p^k q$, where q is not necessarily a prime, but $q \le c^\delta$ and $(p, q) = 1$.

Subcase 1.1: One has $k = 1$. Then we can apply Corollary 7.3 (with M, N replaced by $\tilde{M}, \tilde{N}$), which gives the bound

$$\sum_{\substack{m=1 \\ (m,n,c)=1}}^{M} \sum_{n=1}^{N} \alpha_m \beta_n S(am, n; c) \ll \|\alpha\|\|\beta\| c^{o(1)} \sqrt{MNc}\left(\tilde{N}^{-\frac{1}{2}} c^\delta + (MN)^{-\frac{3}{16}} c^{\frac{11+53\delta}{64}}\right).$$

The first term here is superseded by the third term in (7.6) and the second term in (7.7), since $\tilde{M} \le c$ and $\delta \le \frac{1}{24}$. The second term here appears directly in both (7.6) and (7.7).

Subcase 1.2: One has $k \ge 2$. Then we let $d := p^{\lceil k/2 \rceil} q$, $d' := p^{\lfloor k/2 \rfloor}$, and $e := 1$, which gives a valid decomposition $c = dd'e$ to use in our Theorem 7.1. Moreover, since $k \ge 2$, we have $\lceil k/2 \rceil \le 2k/3$, so $d \le p^{2k/3} q = c^{2/3} q^{1/3}$, and thus

$$d \in [c^{1/2}, c^{(2+\delta)/3}].$$

From (7.8) we thus obtain an upper bound of $\mathcal{B}(c^{(2+\delta)/3})$, which is acceptable in both (7.6) and (7.7) since $\frac{2+\delta}{3} \le \frac{3}{4}$.

Case 2: All prime powers $p^k \mid c$ have $p^k < c^{1-\delta}$.

Subcase 2.1: All prime powers $p^k \mid c$ have $p^k < c^{1/2}$. Then we set $d' := 1$, and construct d, e by a greedy algorithm. Initially, we take $d = e := 1$. For each prime power $p^k \| c$, we append p^k to the smaller of d and e. Note that throughout this process, d and e cannot differ by a factor larger than $c^{1/2}$. In the end, if $d < e$, we swap d and e. This produces a factorization $c = de$ with $(d, e) = 1$ and

$$d \in [c^{1/2}, c^{3/4}].$$

Then (7.8) gives a bound of $\mathcal{B}(c^{3/4})$, which is acceptable in both (7.6) and (7.7).

Subcase 2.2: The largest prime power dividing c is some $p^k \in [c^{1/2}, c^{3/4})$. Then we let $d := p^k$, $d' := 1$, and $e := cp^{-k}$, and (7.8) gives an acceptable bound of $\mathcal{B}(c^{3/4})$ once again.

Subcase 2.3: The largest prime power dividing c is some $p^k \in [c^{3/4}, c^{1-\delta})$. On the one hand, (7.8) gives a bound of $\mathcal{B}(c^{1-\delta})$, which is acceptable in (7.6); this completes the proof of (7.6).

Now assume (still within Subcase 2.3) that $|\alpha_m| \le 1$ for all m, and we aim to establish (7.7).

- If $p = 2$, then writing $c = 2^k q$, we can factorize $c = dd'e$ with $d := 2^{\lceil k/2 \rceil} q$, $d' := 2^{\lfloor k/2 \rfloor}$, and $e := 1$. Here $d \ll \sqrt{2^k q^2} = \sqrt{cq} \le c^{3/4}$, since $2^k \ge c^{1/2}$ implies $q \le c^{1/2}$. But then (7.8) gives an acceptable bound of $\mathcal{B}(c^{3/4})$.
- If $p > 2$, then we can use $d := p^k$ in Theorem 3.5. Since $d \in [c^{3/4}, c^{1-\delta})$, this gives the bound

$$\sum_{m=1}^{M} \sum_{\substack{n=1 \\ (n,c)=1}}^{N} \alpha_m \beta_n S(am, n; c) \ll \sqrt{M}\|\beta\| c^{o(1)} \sqrt{MNc}\left(\frac{c^{\frac{1}{8}}}{M^{\frac{1}{2}}} + \frac{1}{c^{\frac{3}{16}}} + \frac{c^{\frac{1-\delta}{4}}}{N^{\frac{1}{2}}}\right).$$

Since $M, N \geq \tilde{N}$ and $\frac{1}{8} \leq \frac{1-\delta}{4}$, the first and the last terms in the parenthesis above are superseded by the term $c^{(1-\delta)/4}\tilde{N}^{-1/2}$ from (7.7). The second term appears directly in (7.7).

This covers all cases. $\square$

Proof of Theorem 1.1. As before, we can assume without loss of generality that $M, N \leq c$ since the result is trivial when $c = O(1)$. Then the result follows by applying Theorem 7.4 with $M, N \ll c^{1/2+o(1)}$, and using the optimal choices $\delta = \frac{3}{175}$ in (7.6), respectively $\delta = \frac{1}{69}$ in (7.7). $\square$

7.4. **Averaging over moduli.** Finally, let us prove a generalization of Corollary 1.4.

Corollary 7.5. *Let $q = dd'e$ for some $d, d', e \in \mathbb{Z}_+$ with $d' \mid d$ and $(d, e) = 1$, and $f \leq \sqrt{qd}$ be the largest integer with $f^2 \mid qd$. Let $C \geq \frac{1}{2}$ and $\mathcal{I}, \mathcal{J} \subset \mathbb{Z}$ be intervals of lengths $|\mathcal{I}| = M$, $|\mathcal{J}| = N$, with $1 \leq N \leq M \leq C$. Let $(\alpha_m)_{m\in\mathcal{I}}, (\beta_n)_{n\in\mathcal{J}}$ be complex sequences, and for each $c \sim C$, let $(\alpha_m(c))_{m\in\mathcal{I}}$, $(\beta_n(c))_{n\in\mathcal{J}}$ be such that $|\alpha_m(c)| \leq |\alpha_m|$, $|\beta_n(c)| \leq |\beta_n|$ for all $m \in \mathcal{I}, n \in \mathcal{J}$. Then one has*

$$\sum_{\substack{c\sim C\\ q\mid c}} \left| \sum_{\substack{m\in\mathcal{I}, n\in\mathcal{J}\\ (m,n,q)=1}}\sum \alpha_m(c)\beta_n(c)S(m,n;c) \right| \ll \|\alpha\|\|\beta\| \frac{C^{2+o(1)}}{q} \min \begin{cases} \left(\frac{dM^3N}{C^3} + \frac{fM^2}{C^2} + \frac{f}{d^2}\right)^{\frac{1}{6}}, \\ \left(\frac{dM^3N}{qC^2} + \frac{fM^2}{qC} + \frac{fq}{d^2C}\right)^{\frac{1}{6}}. \end{cases}$$

Proof of Corollary 7.5. Throughout this proof, we will use the notation

$$f_a := \max_{\tilde{f}^2\mid a} \tilde{f}, \qquad \|\alpha_{a*}\| := \sqrt{\sum_{\substack{m\in\mathcal{I}\\ a\mid m}} |\alpha_m|^2}, \qquad \|\beta_{a*}\| := \sqrt{\sum_{\substack{n\in\mathcal{J}\\ a\mid n}} |\beta_n|^2},$$

for any $a \in \mathbb{Z}_+$. In particular, the assumption of the present Corollary 7.5 takes $f = f_{qd}$. Note that $f_a \mid f_{ab}$ and $f_{a^2b} = af_b$ for any $a, b \in \mathbb{Z}_+$, and that $f_{ab} = f_af_b$ when $(a,b) = 1$.

We can of course assume without loss of generality that $C \gg q$, since otherwise the sum over c is empty. For each $c \sim C$ with $q \mid c$, we consider the sum

$$\begin{aligned} \mathcal{S}(c) &:= \sum_{\substack{m\in\mathcal{I}, n\in\mathcal{J}\\ (m,n,q)=1}}\sum \alpha_m(c)\beta_n(c)S(m,n;c) \\ &= \sum_{\substack{g\mid c\\ (g,q)=1}} \sum_{\substack{m\in\mathcal{I}, n\in\mathcal{J}\\ (m,n,c)=g}}\sum \alpha_m(c)\beta_n(c)S(m,n;c) = \sum_{\substack{g\mid c\\ (g,q)=1}} \frac{\phi(c)}{\phi(c/g)} \sum_{\substack{m\in\mathcal{I}, n\in\mathcal{J}\\ (m,n,c)=g}}\sum \alpha_m(c)\beta_n(c)S(\tfrac{m}{g}, \tfrac{n}{g}; \tfrac{c}{g}), \end{aligned}$$

where the last equality follows from the identity $S(m,n;c) = \frac{\phi(c)}{\phi(c/g)}S(\frac{m}{g}, \frac{n}{g}; \frac{c}{g})$. From the triangle inequality and the bound $\frac{\phi(c)}{\phi(c/g)} \leq g$, we find that

$$\sum_{\substack{c\sim C\\ q\mid c}} |\mathcal{S}(c)| \leq \sum_{\substack{g\leq 2C/q\\ (g,q)=1}} g \sum_{\substack{c\sim C\\ gq\mid c}} |\mathcal{S}(c;g)|, \tag{7.9}$$

where

$$\mathcal{S}(c;g) := \sum_{\substack{m\in\mathcal{I}, n\in\mathcal{J}\\ g\mid(m,n)\\ (\frac{m}{g}, \frac{n}{g}, \frac{c}{g})=1}}\sum \alpha_m(c)\beta_n(c)S(\tfrac{m}{g}, \tfrac{n}{g}; \tfrac{c}{g}).$$

We aim to apply Theorem 7.1 (with $M, N, c \leftarrow \frac{M}{g}, \frac{N}{g}, \frac{c}{g}$) to bound each sum $\mathcal{S}(c;g)$, and this requires a suitable factorization of the modulus $\frac{c}{g}$. There are two ways to construct this from the assumed factorization $q = dd'e$, which correspond to placing 'most' of the factor $\frac{c}{gq}$ into e or into d.

Method 1. For each $c \sim C$ with $gq \mid c$, consider the factorization

$$\frac{c}{g} =: c'q = \tilde{d}d'\tilde{e}, \qquad \tilde{d} := (c', d^\infty)d, \qquad \tilde{e} := \frac{ec'}{(c', d^\infty)},$$

which has $d' \mid \tilde{d}$ and $(\tilde{d}, \tilde{e}) = 1$. We find that

$$f_{(c/g)\tilde{d}}^2 \mid c'q(c', q^\infty)d = (c', q^\infty)^2 \frac{c'}{(c', q^\infty)} qd \qquad \Rightarrow \qquad f_{(c/g)\tilde{d}} \le (c', q^\infty) f_{c'} f_{qd} = (c', q^\infty) f_{c'} f,$$

so Theorem 7.1 gives

$$\begin{aligned}\mathcal{S}(c; g) = \mathcal{S}(c'gq; g) &\ll \|\alpha_{g*}\|\|\beta_{g*}\| \frac{c^{1+o(1)}}{g} \left(\frac{\tilde{d}M^3N}{c^3} + \frac{f_{(c/g)\tilde{d}}M^2}{c^2} + \frac{f_{(c/g)\tilde{d}}}{\tilde{d}^2} \right)^{\frac{1}{6}} g^{\frac{1}{6}} \\ &\ll \|\alpha_{g*}\|\|\beta_{g*}\| C^{1+o(1)} \left(\frac{dM^3N}{C^3} + \frac{fM^2}{C^2} + \frac{f}{d^2} \right)^{\frac{1}{6}} \left((c', q^\infty) f_{c'} \right)^{\frac{1}{6}}.\end{aligned}$$

Therefore,

$$\begin{aligned}\sum_{\substack{c \sim C \\ gq \mid c}} |\mathcal{S}(c; g)| &= \sum_{c' \sim \frac{C}{gq}} |\mathcal{S}(c; g)| \\ &\ll \|\alpha_{g*}\|\|\beta_{g*}\| C^{1+o(1)} \left(\frac{dM^3N}{C^3} + \frac{fM^2}{C^2} + \frac{f}{d^2} \right)^{\frac{1}{6}} \sum_{c' \sim \frac{C}{gq}} \left((c', q^\infty) f_{c'} \right)^{\frac{1}{6}}.\end{aligned}$$

After applying Cauchy–Schwarz to the last sum, it remains to bound the sums $\sum_{c' \sim C/(gq)} (c', q^\infty)$ and $\sum_{c' \sim C/(gq)} f_{c'}$, both of which are $O(\frac{C^{1+o(1)}}{gq})$. In particular, for the second sum, we can write

$$\sum_{c' \sim \frac{C}{gq}} f_{c'} \le \sum_{f \ll \sqrt{\frac{C}{gq}}} f \sum_{c' \sim \frac{C}{gq}} \mathbb{1}_{f^2 \mid c'} \ll \sum_{f \le \sqrt{\frac{C}{gq}}} \frac{C}{gqf} \ll \frac{C^{1+o(1)}}{gq}.$$

From this and (7.9), we conclude that

$$\sum_{\substack{c \sim C \\ q \mid c}} |\mathcal{S}(c)| \ll \frac{C^{2+o(1)}}{q} \left(\frac{dM^3N}{C^3} + \frac{fM^2}{C^2} + \frac{f}{d^2} \right)^{\frac{1}{6}} \sum_{\substack{g \le 2C/q \\ (g,q)=1}} \|\alpha_{g*}\|\|\beta_{g*}\|.$$

Finally, the last sum is easily bounded by $C^{o(1)}\|\alpha\|\|\beta\|$ using Cauchy–Schwarz and the divisor bound. This establishes the bound from Corollary 7.5 with the first term from the minimum.

Method 2. For each $c \sim C$ with $gq \mid c$, consider the factorization

$$\frac{c}{g} =: c'q = \tilde{d}d'\tilde{e}, \qquad \tilde{d} := \frac{c'd}{(c', e^\infty)}, \qquad \tilde{e} := e(c', e^\infty),$$

which satisfies $d' \mid \tilde{d}$ and $(\tilde{d}, \tilde{e}) = 1$. We find that

$$f_{(c/g)\tilde{d}}^2 \mid (c')^2 qd \qquad \Rightarrow \qquad f_{(c/g)\tilde{d}} \le c' f_{qd} \ll \frac{fC}{q},$$

so Theorem 7.1 gives

$$\mathcal{S}(c; g) = \mathcal{S}(c'gq; g) \ll \|\alpha_{g*}\|\|\beta_{g*}\| C^{1+o(1)} \left(\frac{dM^3N}{qC^2} + \frac{fM^2}{qC} + \frac{fq}{d^2C} \right)^{\frac{1}{6}} (c', e^\infty)^{\frac{1}{3}}.$$

The second bound from Corollary 7.5 now follows similarly as before from (7.9), since the sum over $c' \sim \frac{C}{gq}$ 'washes out' the factor (c', e^∞). $\square$

Proof of Corollary 1.4. This follows from Corollary 7.5 analogously to how Theorem 1.2 follows from Theorem 7.1. $\square$

8. Moments of twisted modular L-functions

Here we prove Theorem 1.5, by inserting our bounds for bilinear forms with Kloosterman sums into the proofs from [3]. We begin by restating (7.7) in a shape more similar to [3, Theorem 5].

Corollary 8.1. *Let $r, q \in \mathbb{Z}_+$ with $r \mid q$. Let $K, M \geq 1$, $\tilde{K} := \max(K, M)$, $\tilde{M} := \min(K, M)$, and $(\lambda_k)_{K \leq k \leq 2K}$ be a sequence with $|\lambda_k| \leq 1$ for all k. Then one has*

$$\sum_{\substack{M \leq m \leq 2M \\ (m,q)=1}} \left| \sum_{K \leq k \leq 2K} \lambda_k S(k, m; r) \right|^2 \ll (qKM)^{o(1)} K^2 M r$$
$$\times \left(\frac{r^{\frac{11+53\delta}{32}}}{(KM)^{\frac{3}{8}}} + \frac{r^{\frac{1-\delta}{2}}}{\tilde{M}} + \frac{1}{r^{\frac{3}{8}}} + \frac{r^{\frac{1}{4}}}{\tilde{M}^{\frac{2}{3}}} + \frac{r^{\frac{11}{12}}}{KM} \right).$$

Proof. One can of course assume without loss of generality that $M, K \in \mathbb{Z}_+$, and extend the sum over m to include all $m \in [M, 2M]$ with $(m, r) = 1$. By duality, it suffices to establish the bound

$$\sum_{\substack{M \leq m \leq 2M \\ (m,r)=1}} \beta_m \sum_{K \leq k \leq 2K} \lambda_k S(k, m; r) \ll (qKM)^{o(1)} \|\beta\| K \sqrt{Mr}$$
$$\times \left(\frac{r^{\frac{11+53\delta}{64}}}{(KM)^{\frac{3}{16}}} + \frac{r^{\frac{1-\delta}{4}}}{\tilde{M}^{\frac{1}{2}}} + \frac{1}{r^{\frac{3}{16}}} + \frac{r^{\frac{1}{8}}}{\tilde{M}^{\frac{1}{3}}} + \frac{r^{\frac{11}{24}}}{(KM)^{\frac{1}{2}}} \right),$$

for any sequence $(\beta_m)_{M \leq m \leq 2M}$. But this is precisely the content of (7.7) with (M, N, c) replaced by (K, M, r); the remark after Theorem 7.4 allows us to ignore the constraint $K, M \leq c$. $\square$

We can now prove an analogue of [3, Proposition 7]. We use the same normalization as in [3, (2.3)] for the Hecke eigenvalues $\lambda_f(n)$ of a holomorphic cuspidal newform f for $\mathrm{SL}_2(\mathbb{Z})$, so that

$$\lambda_f(n)\rho_f(1) = \sqrt{n}\rho_f(n), \qquad \text{where} \qquad f(z) = \sum_{n=1}^{\infty} \rho_f(n)(4\pi n)^{k/2} e(nz). \tag{8.1}$$

In particular, the Deligne bound [8] reads

$$\lambda_f(n) \ll n^{o(1)}. \tag{8.2}$$

Proposition 8.2. *Let $\varepsilon > 0$, $q, d \in \mathbb{Z}_+$ with $d \mid q$, $\frac{1}{20}N \geq M \geq 1$ with $MN \leq q^{2+\varepsilon}$, and let $\lambda_1(m)$, $\lambda_2(n)$ be the Hecke eigenvalues of two (fixed) holomorphic cuspidal newforms for $\mathrm{SL}_2(\mathbb{Z})$. Let $V_1, V_2 : \mathbb{R} \to \mathbb{C}$ be functions supported in $[1, 2]$ with derivatives $V_i^{(j)} \ll_{j,\varepsilon} q^\varepsilon$, and denote*

$$S_{N,M,d,q} := \frac{d}{(NM)^{1/2}} \sum_{r=1}^{2N/d} \sum_{\substack{n \equiv m \pmod d \\ (nm,q)=1 \\ n \neq m}} \lambda_1(m)\lambda_2(n) V_1\left(\frac{m}{M}\right) V_2\left(\frac{n}{N}\right). \tag{8.3}$$

Then for any $\delta \in [0, \frac{1}{24}]$, one has

$$S_{N,M,d,q} \ll_\varepsilon q^{O(\varepsilon)} \left(\frac{M^{\frac{5}{16}} q^{\frac{83+53\delta}{64}}}{N^{\frac{5}{16}}} + \frac{q^{\frac{7-\delta}{4}}}{\sqrt{N}} + \frac{M^{\frac{1}{2}} q^{\frac{21}{16}}}{N^{\frac{1}{2}}} + \frac{q^{\frac{13}{8}} M^{\frac{1}{6}}}{N^{\frac{1}{2}}} + q^{\frac{23}{24}} \right). \tag{8.4}$$

Proof. We closely follow the proof in [3, §4]. In particular, we decompose $q = q_d q'$ where q' is maximal with $(q', d) = 1$. The bound [3, (4.2)] reads

$$S_{N,M,d,q} \ll \sqrt{N} \sum_{\substack{g|f|q' \\ r|d}} \frac{\mu^2(f)|\lambda_2(f/g)|}{fgr} \left(\sum_{\substack{m \asymp M \\ (m,q)=1}} \Big| \sum_n S(\overline{fg}m, n; r) \lambda_2(n) V_2^\circ \left(\frac{nN}{fgr^2} \right) \Big|^2 \right)^{1/2},$$

where V_2° is a transform of V_2 as in [3, (2.10)] (coming from an application of the Voronoi summation formula). Using the rapid decay of V_2°, we may truncate the sum over n at

$$n \le K_{f,g,r} := q^\varepsilon \frac{fgr^2}{N},$$

up to an acceptable loss. Note that the resulting sum over n vanishes unless $K_{f,g,r} \ge 1$. From Corollary 8.1, (8.2), and the divisor bound, we conclude that

$$S_{N,M,d,q} \ll_\varepsilon q^{O(\varepsilon)} \sqrt{N} \max_{\substack{g|f|q' \\ r|d}} \frac{1}{fgr} K_{f,g,r} \sqrt{Mr} \left(\frac{r^{\frac{11+53\delta}{64}}}{(K_{f,g,r}M)^{\frac{3}{16}}} + \frac{r^{\frac{1-\delta}{4}}}{\min(K_{f,g,r}, M)^{\frac{1}{2}}} + \frac{1}{r^{\frac{3}{16}}} \right.$$
$$\left. + \frac{r^{\frac{1}{8}}}{\min(K_{f,g,r}, M)^{\frac{1}{3}}} + \frac{r^{\frac{11}{24}}}{(K_{f,g,r}M)^{\frac{1}{2}}} \right).$$

Plugging in the definition of $K_{f,g,r}$, we see that the expression inside the maximum is non-decreasing in r and non-increasing in f, g. Writing

$$K := K_{1,1,q} = \frac{q^{2+\varepsilon}}{N} \ge M,$$

we find that

$$S_{N,M,d,q} \ll_\varepsilon q^{O(\varepsilon)} \sqrt{N} \frac{1}{q} K \sqrt{Mq} \left(\frac{q^{\frac{11+53\delta}{64}}}{(KM)^{\frac{3}{16}}} + \frac{q^{\frac{1-\delta}{4}}}{M^{\frac{1}{2}}} + \frac{1}{q^{\frac{3}{16}}} + \frac{q^{\frac{1}{8}}}{M^{\frac{1}{3}}} + \frac{q^{\frac{11}{24}}}{(KM)^{\frac{1}{2}}} \right)$$
$$\ll_\varepsilon q^{O(\varepsilon)} \frac{q^{\frac{3}{2}} M^{\frac{1}{2}}}{N^{\frac{1}{2}}} \left(\frac{q^{\frac{11+53\delta}{64}}}{(q^2M/N)^{\frac{3}{16}}} + \frac{q^{\frac{1-\delta}{4}}}{M^{\frac{1}{2}}} + \frac{1}{q^{\frac{3}{16}}} + \frac{q^{\frac{1}{8}}}{M^{\frac{1}{3}}} + \frac{q^{\frac{11}{24}}}{(q^2M/N)^{\frac{1}{2}}} \right),$$

which reduces to the desired bound. □

We can now prove the desired asymptotic for twisted moments of modular L-functions.

Proof of Theorem 1.5. Let $\varepsilon > 0$ and $\gamma := \frac{1}{674}$. We closely follow the proof in [3, §3], making no changes to the main term analysis from [3, §3.1]. Treating the off-diagonal term as in [3, §3.2], it remains to establish the bound

$$S_{N,M,d,q} \overset{?}{\ll}_\varepsilon q^{1-\gamma+O(\varepsilon)}, \tag{8.5}$$

for all $d \mid q$ and $N \ge M \ge 1$ with $MN \le q^{2+\varepsilon}$, using the notation from (8.3). As in [3, §3.3], we can easily discount the contribution of the range $M \le N < 20M$ using [3, (3.12)], so let us assume that $N \ge 20M$. We will rely on the bounds

$$S_{N,M,d,q} \ll_\varepsilon q^{O(\varepsilon)} (MN)^{\frac{1}{2}}, \tag{8.6}$$

$$S_{N,M,d,q} \ll_\varepsilon q^{O(\varepsilon)} \left(\frac{(Nq)^{\frac{1}{2}}}{M^{\frac{1}{2}}} + \frac{N^{\frac{3}{4}}}{M^{\frac{1}{4}}} + \frac{N^{\frac{1}{4}} q^{\frac{3}{4}}}{M^{\frac{1}{4}}} + N^{\frac{1}{2}} q^{\frac{1}{4}} \right), \tag{8.7}$$

from [3, (3.6) and (3.11)], as well as on our Proposition 8.2 (instead of [3, Proposition 7]). First, the trivial bound (8.6) establishes (8.5) unless

$$M > \frac{q^{2-2\gamma}}{N}, \tag{8.8}$$

so let us assume that we are in this range. We now split into cases depending on the size of N.

Case 1. One has $N \le q^{3/2-3\gamma}$. Then by plugging (8.8) into (8.7), we obtain (8.5).

Case 2. One has $N \in (q^{3/2-3\gamma}, q^{3/2-2\gamma}]$. Then by plugging (8.8) into (8.7), we find that

$$S_{N,M,d,q} \ll_\varepsilon q^{O(\varepsilon)} \left(q^{1-\gamma} + \frac{N^{\frac{1}{4}} q^{\frac{3}{4}}}{M^{\frac{1}{4}}} \right),$$

which is acceptable in (8.5) unless

$$\frac{N}{M} > q^{1-4\gamma}.$$

Plugging this and $N > q^{3/2-3\gamma}$ into (8.4), we find that

$$S_{N,M,d,q} \ll_\varepsilon q^{O(\varepsilon)} \left(q^{\frac{63+53\delta}{64} + \frac{5\gamma}{4}} + q^{1-\frac{\delta}{4}+\frac{3\gamma}{2}} + q^{\frac{23}{24}+2\gamma} \right), \tag{8.9}$$

which is acceptable in (8.5) provided that

$$10\gamma \le \delta \le \frac{1-144\gamma}{53}.$$

This is precisely attained for our choice of $\gamma = \frac{1}{674}$ by taking $\delta := \frac{10}{674}$ in Proposition 8.2.

Case 3. One has $N \in (q^{3/2-2\gamma}, q^{3/2+\gamma})$. Then (8.7) is useless because of the last term. We plug in $M \le q^{2+\varepsilon}/N$ and then $N \ge q^{3/2-2\gamma}$ into (8.4) to find that

$$S_{N,M,d,q} \ll_\varepsilon q^{O(\varepsilon)} \left(q^{\frac{63+53\delta}{64} + \frac{5\gamma}{4}} + q^{1-\frac{\delta}{4}+\gamma} + q^{\frac{23}{24}+2\gamma} \right),$$

which is a stronger bound than (8.9). This completes our proof. $\square$

9. Large sieve for exceptional cusp forms

Here we prove a generalization of Corollary 1.6, which requires some background from the spectral theory of automorphic forms. We recall [11] that for $q \in \mathbb{Z}_+$, the congruence subgroup $\Gamma_0(q)$ contains those matrices in $\mathrm{SL}_2(\mathbb{Z})$ with bottom-left entries divisible by q. Each cusp $\mathfrak{a}$ of the the fundamental domain $\Gamma_0(q)\backslash\mathbb{H}$ is equivalent to a fraction of the form $\frac{u}{w}$, where $u, w \in \mathbb{Z}_+$, $w \mid q$, $(u,w) = 1$, and $u \le (w, \frac{q}{w})$; in particular, the cusp at ∞ is equivalent to $\frac{1}{q}$. To such a cusp, one can associate a scaling matrix $\sigma_\mathfrak{a} \in \mathrm{PSL}_2(\mathbb{R})$ with $\sigma_\mathfrak{a}\infty = \mathfrak{a}$, and via these scaling matrices, functions on $\Gamma_0(q)\backslash\mathbb{H}$ can be Fourier expanded around $\mathfrak{a}$.

The discrete spectrum of the hyperbolic Laplacian $\Delta = -y^2(\partial_x^2 + \partial_y^2)$ is parametrized by Maass cusp forms: these are smooth functions $f : \Gamma_0(q)\backslash\mathbb{H} \to \mathbb{C}$ which are eigenfunctions of Δ, vanish at all cusps of $\Gamma_0(q)\backslash\mathbb{H}$, and are square-integrable with respect to the Petersson inner product. Following the normalization of Deshouillers–Iwaniec [11], we write the Fourier expansion of f at $z = x+iy \in \mathbb{H}$ around a cusp $\mathfrak{a}$ (with scaling matrix $\sigma_\mathfrak{a}$) as

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

where K is a Whittaker function as in [11, p. 264]. Altering the choice of scaling matrix $\sigma_{\mathfrak{a}}$ results in multiplying the Fourier coefficients $\rho_{\mathfrak{a}}(n)$ by an exponential phase $e(n\omega)$, for some uniform $\omega \in \mathbb{R}/\mathbb{Z}$.

The Kuznetsov trace formula [11, 27], as well as the large sieve inequalities that derive from it, involve an orthonormal basis of Maass cusp forms. The following notation will therefore be useful.

Notation 9.1. Let $q \in \mathbb{Z}_+$, $\mathfrak{a}$ be a cusp of $\Gamma_0(q)$ equivalent[7] to $\frac{1}{s}$ for some $s \mid q$ with $(s, \frac{q}{s}) = 1$, and $\sigma_{\mathfrak{a}} \in \mathrm{PSL}_2(\mathbb{R})$ be any scaling matrix for $\mathfrak{a}$. Consider an orthonormal basis $(f_j)_{j \geq 1}$ of Maass cusp forms for $\Gamma_0(q)$, with:

(i). Laplacian eigenvalues λ_j and spectral parameters $\theta_j := \max(0, \frac{1}{4} - \lambda_j)^{1/2}$;
(ii). Fourier coefficients $(\rho_{j\mathfrak{a}}(n))_{n \in \mathbb{Z}}$ around the cusp $\mathfrak{a}$, using the scaling matrix $\sigma_{\mathfrak{a}}$.

Proposition 9.2. *Assume Notation 9.1, let $X, N \geq 1/2$, and let $(\alpha_n)_{n \sim N}$ be a complex sequence. Let $\Phi : \mathbb{R} \to [0, \infty)$ be a smooth function supported in $[\Omega(1), O(1)]$, with $\int \Phi(t)\, dt \gg 1$ and $\Phi^{(j)}(t) \ll_j 1$. Then there exists $\omega \in \mathbb{R}/\mathbb{Z}$ (depending only on $\mathfrak{a}$, $\sigma_{\mathfrak{a}}$) such that*

$$\begin{aligned} \sum_{\lambda_j < 1/4} X^{2\theta_j} \left| \sum_{n \sim N} \alpha_n\, \rho_{j\mathfrak{a}}(n) \right|^2 &\ll (qN)^{o(1)} \left(1 + \frac{N}{q}\right) \|\alpha_n\|^2 \\ &\quad + \left| \sum_{c \in q\mathbb{Z}_+} \frac{1}{c} \sum_{m,n \sim N} \overline{\alpha_m e(m\omega)}\, \alpha_n e(n\omega)\, S(m,n;c)\, \Phi\left(\frac{\sqrt{mn}}{c} X\right) \right|. \end{aligned} \tag{9.1}$$

Proof. This is [32, Corollary I], which follows from the Kuznetsov trace formula and the regular-spectrum large sieve inequalities of Deshouillers–Iwaniec [11, Theorem 2]. We have implicitly used [32, Lemma B] to write down the Kloosterman sums and c-supports for cusps $\mathfrak{a}$ equivalent to $\frac{1}{s}$ for some $s \mid q$ (the latter condition is written as $\mu(\mathfrak{a}) = q^{-1}$ in loc. cit.). Note that we incur factors of $e(m\omega)$ and $e(n\omega)$ since we do not assume a special scaling matrix $\sigma_{\mathfrak{a}}$ (as we may), but this will be irrelevant in our computations since the sequence (α_n) is arbitrary. ☐

In the right-hand side of (9.1), the sum over c is really supported on $c \asymp NX$ due to the Φ-weight, and it vanishes if $q \gg NX$ with a large enough implied constant. Deshouillers–Iwaniec used this simple observation to deduce the following result, which combines [11, Theorems 2 and 5].

Theorem 9.3 (Deshouillers–Iwaniec [11])**.** *Assume Notation 9.1, let $N \geq \frac{1}{2}$, and let $(\alpha_n)_{n \sim N}$ be a complex sequence. Then one has*

$$\sum_{\lambda_j < 1/4} X^{2\theta_j} \left| \sum_{n \sim N} \alpha_n\, \rho_{j\mathfrak{a}}(n) \right|^2 \ll (qN)^{o(1)} \left(1 + \frac{N}{q}\right) \|\alpha\|^2, \tag{9.2}$$

for any positive $X \ll 1 + \frac{q}{N}$.

Until now, if $\sqrt{q} \ll N \ll q$, Theorem 9.3 has been the state-of-the-art exceptional-spectrum large sieve bound for general sequences (α_n) and a single group $\Gamma_0(q)$; the same is true if one averages over levels $q \sim Q$ and allows the sequence (α_n) to depend on q.

We can now achieve an improvement of Theorem 9.3 when q has a factorization as in Theorem 7.1, and similar results can be deduced for arbitrary levels q using Theorem 7.4. We require a coprimality constraint $(n, q) = 1$ for technical reasons, but this is usually harmless in applications. The resulting power savings are relatively small, but serve as a proof of concept that Theorem 9.3 is not a fundamental barrier.

[7]The assumption that $\mathfrak{a}$ is equivalent to $\frac{1}{s}$ is true in most applications (note that this includes the cusp at ∞), and only made for convenience; one can prove similar results at arbitrary cusps with small adjustments.

Theorem 9.4 (Large sieve for composite levels)**.** *Assume Notation 9.1, let $N \geq \frac{1}{2}$, and let $(\alpha_n)_{n\sim N}$ be a complex sequence supported on $(n,q)=1$. Suppose that $q = dd'e$ with $d' \mid d$ and $(d,e)=1$, and let $f \leq \sqrt{qd}$ be the largest integer with $f^2 \mid qd$. Then (9.2) holds for any positive*

$$X \ll 1 + \frac{q}{N} + \min\left(\frac{q^2}{d^{1/3}N^{7/3}}, \frac{q^{3/2}}{f^{1/4}N^{3/2}}, \frac{qd^{1/3}}{f^{1/6}N}\right) + \min\left(\frac{q^{7/4}}{d^{1/4}N^2}, \frac{q^{7/5}}{f^{1/5}N^{7/5}}, \frac{qd^{2/5}}{f^{1/5}N}\right). \tag{9.3}$$

Proof of Theorem 9.4. We may assume without loss of generality that

$$1 + \frac{q}{N} < X \ll \min\left(\frac{q^2}{d^{1/3}N^{7/3}}, \frac{q^{3/2}}{f^{1/4}N^{3/2}}, \frac{qd^{1/3}}{f^{1/6}N}\right) + \min\left(\frac{q^{7/4}}{d^{1/4}N^2}, \frac{q^{7/5}}{f^{1/5}N^{7/5}}, \frac{qd^{2/5}}{f^{1/5}N}\right), \tag{9.4}$$

since otherwise the result follows from Theorem 9.3. We apply Proposition 9.2 with a choice of Φ supported on $[2,4]$, then separate variables in the smooth weight $\Phi(\cdot)$ via two-dimensional Fourier inversion, as in [32, Proof of Theorem 13], to arrive at

$$\sum_{\lambda_j < 1/4} X^{2\theta_j} \left|\sum_{n\sim N} \alpha_n\, \rho_{j\mathfrak{a}}(n)\right|^2 \ll (qN)^{o(1)} \left(1 + \frac{N}{q}\right) \|\alpha\|^2$$
$$+ \sum_{\substack{\frac{NX}{4} < x \leq NX \\ q\mid c}} \frac{1}{c} \sup_{\substack{(\beta_n)_{n\sim N} \\ |\beta_n| = |\alpha_n|}} \sup_{\substack{(\gamma_n)_{n\sim N} \\ |\gamma_n| = |\alpha_n|}} \left|\sum_{m,n\sim N} \beta_m \gamma_n S(m,n;c)\right|.$$

The sequences (β_n), (γ_n) in the supremum arise by incorporating exponential phases $e(n\omega)$ into (α_n), partly from the choice of the scaling matrix $\sigma_{\mathfrak{a}}$, and partly due to the separation of variables. The suprema are of course attained by some sequences (β_n), (γ_n) supported on $(n,q)=1$, so we can apply Corollary 7.5 with $M = N$ and $C \asymp NX$, to obtain

$$\sum_{\lambda_j < 1/4} X^{2\theta_j} \left|\sum_{n\sim N} \alpha_n\, \rho_{j\mathfrak{a}}(n)\right|^2 \ll (qN)^{o(1)} \left(1 + \frac{N}{q}\right) \|\alpha\|^2$$
$$+ \|\alpha\|^2 \frac{(NX)^{1+o(1)}}{q} \min \begin{cases} \left(\frac{dN}{X^3} + \frac{f}{X^2} + \frac{f}{d^2}\right)^{\frac{1}{6}}, \\ \left(\frac{dN^2}{qX^2} + \frac{fN}{qX} + \frac{fq}{d^2NX}\right)^{\frac{1}{6}}. \end{cases}$$

We conclude by noting that

$$\frac{NX}{q}\left(\frac{dN}{X^3} + \frac{f}{X^2} + \frac{f}{d^2}\right)^{\frac{1}{6}} \ll 1 \qquad \text{for} \qquad X \ll \min\left(\frac{q^2}{d^{1/3}N^{7/3}}, \frac{q^{3/2}}{f^{1/4}N^{3/2}}, \frac{qd^{1/3}}{f^{1/6}N}\right),$$

and

$$\frac{NX}{q}\left(\frac{dN^2}{qX^2} + \frac{fN}{qX} + \frac{fq}{d^2NX}\right)^{\frac{1}{6}} \ll 1 \qquad \text{for} \qquad X \ll \min\left(\frac{q^{7/4}}{d^{1/4}N^2}, \frac{q^{7/5}}{f^{1/5}N^{7/5}}, \frac{qd^{2/5}}{f^{1/5}N}\right).$$

This covers the range in (9.4). $\square$

Proof of Corollary 1.6. If q has a divisor $d \asymp \sqrt{q}$ such that $\frac{q}{d}$ is square-free, then we can take $d' = (d, \frac{q}{d})$ and $f = d \asymp \sqrt{q}$ in Theorem 9.4, so (9.2) holds for any positive

$$X \ll 1 + \frac{q}{N} + \min\left(\frac{q^{11/6}}{N^{7/3}}, \frac{q^{11/8}}{N^{3/2}}, \frac{q^{13/12}}{N}\right) + \min\left(\frac{q^{13/8}}{N^2}, \frac{q^{13/10}}{N^{7/5}}, \frac{q^{11/10}}{N}\right).$$

If additionally $N \ll q^{1/2+o(1)}$ (as Corollary 1.6 assumes), then we can take $X = q^{3/5}$, since this is only larger by a factor of $q^{o(1)}$ than the second minimum above. $\square$

Appendix A. Some necessary computations in $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$

Here we fill in some details involving explicit matrix computations, subgroups, and characters of $\mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$. In particular, we prove Lemmas 4.8, 5.2 and 5.4.

Proof of Lemma 4.8. The first equality from (4.6) follows immediately from (4.4). Now let $T_c(d) : V_c \to V_c$ denote the map in the (extreme) right-hand side of (4.6); we will show that $T_c(d) = P_c(d)$.

The fact that $\Gamma_c(d)$ is a subgroup quickly implies that $T_c(d)$ is self-adjoint and that $T_c(d)^2 = T_c(d)$, so $T_c(d)$ is an orthogonal projection. Moreover, one has $\rho_c(n)T_c(d) = T_c(d)$ for any $n \in \Gamma_c(d)$, so any $T_c(d)f \in T_c(d)V_c$ has $\rho_c(n)T_c(d)f = T_c(d)f$, which shows $T_c(d)V_c \subset V_c(d)$. Conversely, if $f \in V_c(d)$, so $\rho_c(n)f = f$ for all $n \in \Gamma_c(d)$, then clearly $f = T_c(d)f$, which shows $V_c(d) \subset T_c(d)V_c$. Thus $T_c(d)$ is the orthogonal projection onto $T_c(d)V_c = V_c(d)$, i.e., $T_c(d) = P_c(d)$.

The claim about commutativity follows directly from (4.6) and the normality of $\Gamma_c(d)$.

Finally, let us prove (4.7). It follows from (4.6) and (3.17) that

$$P_c(d)_{u,v} = \frac{1}{|\Gamma_c(d)|} \sum_{n \in \Gamma_c(d)} \mathbb{1}_{u=nv} = \frac{d^3}{c^3} \mathbb{1}_{u \in \Gamma_c(d) \cdot v} |\Gamma_c(d)_u|, \tag{A.1}$$

where $|\Gamma_c(d)_u|$ is the stabilizer of u inside $\Gamma_c(d)$ (indeed, once $u = n_0 v$ for some $n_0 \in \Gamma_c(d)$, all other solutions to $u = nv$ satisfy $n_0 n^{-1} u = u$, so $n \in \Gamma_c(d)_u n_0$). By the normality of $\Gamma_c(d)$, we see that for any $g \in \mathrm{SL}_2(\mathbb{Z}/c\mathbb{Z})$,

$$\begin{aligned} |\Gamma_c(d)_{gu}| &= \{n \in \Gamma_c(d) : ngu = gu\} \\ &= \{n \in \Gamma_c(d) : (g^{-1}ng)u = u\} = |(g^{-1}\Gamma_c(d)g)_u| = |\Gamma_c(d)_u|, \end{aligned}$$

so by the transitivity of the projective action,

$$|\Gamma_c(d)_u| = |\Gamma_c(d)_{[1:0]}| = \left\{ \begin{pmatrix} q & r \\ s & t \end{pmatrix} \in \Gamma_c(d) : [q:s] = [1:0] \right\}.$$

Note that $[q : s] = [1 : 0]$ simply means $(q, s) = (\alpha, 0)$ for some unit $\alpha \in (\mathbb{Z}/c\mathbb{Z})^\times$. This forces $s = 0$, $q \in (\mathbb{Z}/c\mathbb{Z})^\times$, and $t = \overline{q}$; moreover, given such a choice of q, s, t, any $r \in d\mathbb{Z}/c\mathbb{Z}$ gives a solution. Since there are $\phi(c)/\phi(d)$ choices of q in the kernel of $\mathbb{Z}/c\mathbb{Z} \to \mathbb{Z}/d\mathbb{Z}$, and $\frac{c}{d}$ choices of r in $d\mathbb{Z}/c\mathbb{Z} \cong \mathbb{Z}/\frac{c}{d}\mathbb{Z}$, we find that

$$|\Gamma_c(d)_u| = \frac{\phi(c)}{\phi(d)} \cdot \frac{c}{d},$$

and plugging this into (A.1) proves (4.7). $\square$

Proof of Lemma 5.4. Set $G := \mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$ and $N := \Gamma_{p^k}(p^j)$, so $N \triangleleft G$. Say $\chi = \mathrm{Tr}\rho$ where $\rho \in \widehat{G}$ is primitive. By (3.9), we have

$$\frac{1}{|N|} \sum_{n \in N} |\chi(n)|^2 = \sum_{\rho_0 \in \widehat{N}} \mathrm{Mult}(\rho_0, \rho|_N)^2.$$

By Lemma 3.9, $\rho|_N$ contains L irreducible representations of N, each with multiplicity m, for some positive integers L, m. Thus

$$\sum_{\rho_0 \in \widehat{N}} \mathrm{Mult}(\rho_0, \rho|_N)^2 = Lm^2,$$

and in light of (3.17), it remains to show that

$$Lm^2 \gg (k-j+1)^{-1} p^{2j}. \tag{A.2}$$

If $j = 0$, this is a trivial statement. Suppose now that $\frac{k}{2} \le j \le k$. Then

$$\begin{aligned} N = \Gamma_{p^k}(p^j) &= \left\{ I + p^j A : A \in (\mathbb{Z}/p^{k-j}\mathbb{Z})^{2\times 2},\ \det(I + p^j A) \equiv 1 \pmod{p^k} \right\} \\ &= \left\{ I + p^j A : A \in (\mathbb{Z}/p^{k-j}\mathbb{Z})^{2\times 2},\ \mathrm{Tr}(A) \equiv 0 \pmod{p^{k-j}} \right\} \end{aligned}$$

is abelian, since $(I + p^j A)(I + p^j B) = I + p^j(A + B)$ for $j \ge \frac{k}{2}$. In fact, this shows that

$$(N, \cdot) \cong \left(\left\{ A \in (\mathbb{Z}/p^{k-j}\mathbb{Z})^{2\times 2} : \mathrm{Tr}(A) = 0 \right\}, + \right) \cong (\widehat{N}, \cdot).$$

In particular, all irreducible representations of N are 1-dimensional, and can be expressed as

$$\sigma_B(I + p^k A) := e\left(\frac{\mathrm{Tr}(AB)}{p^{k-j}} \right), \qquad B \in (\mathbb{Z}/p^{k-j}\mathbb{Z})^{2\times 2},\ \mathrm{Tr}(B) = 0. \tag{A.3}$$

Now equating dimensions and using Lemma 3.16, we find that

$$Lm = \dim \rho \gg p^k \qquad \iff \qquad Lm^2 \gg \frac{p^{2k}}{L}. \tag{A.4}$$

We will finish by finding an upper bound on L. By the conclusion of Lemma 3.9, all L non-isomorphic representations in the decomposition of $\rho|_N$ lie in the same orbit of G's action by conjugation. For $g \in G$ and B, σ_B as in (A.3), we have

$$\begin{aligned} \sigma_B(g(I + p^k A)g^{-1}) = \sigma_B(I + p^k gAg^{-1}) &= e\left(\frac{\mathrm{Tr}(gAg^{-1}B)}{p^{k-j}} \right) \\ &= e\left(\frac{\mathrm{Tr}(Ag^{-1}Bg)}{p^{k-j}} \right) = \sigma_{g^{-1}Bg}(I + p^k A). \end{aligned}$$

In other words, the action of G by conjugation on irreducible representations of N corresponds to conjugation of the underlying matrices B. It follows that L is at most the maximal size of an orbit in the set

$$\left\{ B \in (\mathbb{Z}/p^{k-j}\mathbb{Z})^{2\times 2} : \mathrm{Tr}(B) = 0 \right\}$$

under conjugation by $\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$, or equivalently by $\mathrm{SL}_2(\mathbb{Z}/p^{k-j}\mathbb{Z})$. Since conjugation preserves the determinant $\Delta = \det(B)$, we find that

$$L \le \max_{\Delta \in \mathbb{Z}/p^{k-j}\mathbb{Z}} \# \left\{ B \in (\mathbb{Z}/p^{k-j}\mathbb{Z})^{2\times 2} : \mathrm{Tr}(B) = 0,\ \det(B) = \Delta \right\}.$$

Writing $B = \left(\begin{smallmatrix} x & y \\ z & -x \end{smallmatrix}\right)$, we further get

$$L \le \max_{\Delta \in \mathbb{Z}/p^{k-j}\mathbb{Z}} \sum_{x,y,z \in \mathbb{Z}/p^{k-j}\mathbb{Z}} \mathbb{1}_{-x^2 - yz = \Delta} \le p^{k-j} \max_{a \in \mathbb{Z}/p^{k-j}\mathbb{Z}} \sum_{y,z \in \mathbb{Z}/p^{k-j}\mathbb{Z}} \mathbb{1}_{yz = a},$$

where we substituted $a := -x^2 - \Delta$. Now given $a \in \mathbb{Z}/p^{k-j}\mathbb{Z}$, write $a = p^\ell a'$ for some $0 \le \ell \le k-j$ and $a' \in (\mathbb{Z}/p^{k-j-\ell}\mathbb{Z})^\times$. The equation $yz = a$ then implies

$$y = p^{\ell_y} y', \qquad z = p^{\ell_z} z', \qquad y'z' = a'.$$

for some $\ell_y, \ell_z \ge 0$ with $\ell_y + \ell_z = \ell$, and some $y' \in (\mathbb{Z}/p^{k-j-\ell_y}\mathbb{Z})^\times$, $z' \in (\mathbb{Z}/p^{k-j-\ell_z}\mathbb{Z})^\times$. There are $\ell + 1 \le k - j + 1$ choices of (ℓ_y, ℓ_z), and for every choice of (ℓ_y, ℓ_z, y'), there are at most $p^{k-j-\ell_z}/p^{k-j-\ell} = p^{\ell_y}$ choices of z' (since $z' \pmod{p^{k-j-\ell}}$ is fixed). Putting these counts together, we obtain

$$L \le p^{k-j}(k-j+1) \max_{\ell_y + \ell_z = \ell \le k} p^{k-j-\ell_y} p^{\ell_y} \ll (k-j+1) p^{2k-2j}.$$

Combining this with (A.4) establishes the desired bound from (A.2). $\square$

Proof of Lemma 5.2. From (4.3), it follows that $\chi_c(g) = \prod_{p^k \| c} \chi_{p^k}(\pi_{c,p^k}(g))$. Working locally at a prime $p|c$, with say $p^k \| c$ and $p^j \| d$, we will establish the bound

$$\chi_{p^k}(g) \ll p^{\lfloor \frac{k+j}{2} \rfloor}, \tag{A.5}$$

for all $g \in \mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$ such that p^j is the largest p-power for which $g \in \{\gamma \in \mathbb{Z}/p^k\mathbb{Z} : \gamma^2 = 1\} \cdot \Gamma_{p^k}(p^j)$. Given (A.5), the desired bound in (5.1) follows from the divisor bound.

Since $\rho_{p^k}(g)$ is a permutation map, $\chi_{p^k}(g) = \mathrm{Tr}\rho_{p^k}(g)$ equals the number of fixed points of g in $\mathbb{P}^1(\mathbb{Z}/p^k\mathbb{Z})$, i.e., the number of solutions in $u \in \mathbb{P}^1(\mathbb{Z}/p^k\mathbb{Z})$ to $gu = u$. Let us write $g = \left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right)$ and $u = [x : y]$ for some integers q, r, s, t, x, y with $qt - rs \equiv 1 \pmod{p^k}$ and $(x, y, p) = 1$. Scaling both entries of u by a unit in $(\mathbb{Z}/p^k\mathbb{Z})^\times$, we can assume without loss of generality that $x = 1$ or $y = 1$; in fact, replacing $\left(\begin{smallmatrix} q & r \\ s & t \end{smallmatrix}\right) \leftrightarrow \left(\begin{smallmatrix} t & s \\ r & q \end{smallmatrix}\right)$ if necessary, we may assume that $y = 1$. Then the equality $gu = u$ means that for some $\alpha \in \mathbb{Z}$, one has

$$\begin{pmatrix} q & r \\ s & t \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} \equiv \alpha \begin{pmatrix} x \\ 1 \end{pmatrix} \pmod{p^k} \qquad \Rightarrow \qquad qx + r \equiv (sx + t)x \pmod{p^k}.$$

This gives the quadratic congruence

$$sx^2 + (t - q)x - r \equiv 0 \pmod{p^k}.$$

Now from our assumption that $g \in \gamma\Gamma_{p^k}(p^j)$ for some $\gamma \in \mathbb{Z}/p^k\mathbb{Z}$ with $\gamma^2 = 1$, we know that $p^j \mid s$, $p^j \mid r$, and $p^j \mid t - q$ (since $t \equiv \gamma \equiv q \pmod{p^j}$). In fact, p^j is the largest p-power with this property (otherwise, we could pick some $\gamma \equiv q \pmod{p^{j+1}}$ such that $g \in \gamma\Gamma_{p^k}(p^{j+1})$). Therefore, letting $a_2 := sp^{-j}$, $a_1 := (t - q)p^{-j}$ and $a_0 := -rp^{-j}$, we find that

$$a_2x^2 + a_1x + a_0 \equiv 0 \pmod{p^{k-j}}, \tag{A.6}$$

where a_0, a_1, a_2 are not all divisible by p. It now remains to show that this equation has

$$O\left(p^{\lfloor \frac{k-j}{2} \rfloor}\right)$$

solutions in $x \pmod{p^{k-j}}$; every such solution will have p^j lifts to $\mathbb{Z}/p^k\mathbb{Z}$, inducing a total of $O(p^{\lfloor (k-j)/2 \rfloor + j}) = O(p^{\lfloor (k+j)/2 \rfloor})$ fixed points $u = [x : 1]$ of g.

Case 1. $p \nmid a_2$. Then given any two solutions x_0, x of (A.6), we can subtract the two equalities to obtain

$$p^{k-j} \mid a_2(x^2 - x_0^2) + a_1(x - x_0) = (x - x_0)(a_2(x + x_0) + a_1). \tag{A.7}$$

Let $\ell := \lceil (k - j)/2 \rceil$. By the pigeonhole principle, we must have $p^\ell \mid x - x_0$ or $p^\ell \mid a_2(x + x_0) + a_1$. Since $p \nmid a_2$, either option uniquely determines $x \pmod{p^\ell}$ in terms of x_0. So there can be at most

$$\frac{p^{k-j}}{p^\ell} = p^{(k-j) - \lceil \frac{k-j}{2} \rceil} = p^{\lfloor \frac{k-j}{2} \rfloor}$$

solutions in $x \pmod{p^{k-j}}$.

Case 2. $p \nmid a_1$. Given the previous case, we can assume $p \mid a_2$. Then $p \nmid a_2(x + x_0) + a_1$, so from (A.7) we find that $p^{k-j} \mid x - x_0$, forcing only one solution in $x \pmod{p^{k-j}}$.

Case 3. $p \nmid a_0$. Then (A.6) implies $p \nmid x$, and by substituting $x \leftrightarrow \overline{x} \pmod{p^{k-j}}$, we reduce to the case $p \nmid a_2$. $\square$

References

[1] Valentin Blomer, Étienne Fouvry, Emmanuel Kowalski, Philippe Michel, and Djordje Milićević. On moments of twisted L-functions. *Amer. J. Math.*, 139(3):707–768, 2017.

[2] Valentin Blomer, Étienne Fouvry, Emmanuel Kowalski, Philippe Michel, Djordje Milićević, and Will Sawin. The second moment theory of families of L-functions—the case of twisted Hecke L-functions. *Mem. Amer. Math. Soc.*, 282(1394):v+148, 2023.

[3] Valentin Blomer and Djordje Milićević. The second moment of twisted modular L-functions. *Geom. Funct. Anal.*, 25(2):453–516, 2015.

[4] Enrico Bombieri, John B. Friedlander, and Henryk Iwaniec. Primes in arithmetic progressions to large moduli. *Acta Math.*, 156(3-4):203–251, 1986.

[5] Jean Bourgain and Alex Gamburd. Expansion and random walks in $\mathrm{SL}_d(\mathbb{Z}/p^n\mathbb{Z})$. I. *J. Eur. Math. Soc. (JEMS)*, 10(4):987–1011, 2008.

[6] A. H. Clifford. Representations induced in an invariant subgroup. *Ann. of Math. (2)*, 38(3):533–550, 1937.

[7] 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):1577–1624, 2020.

[8] Pierre Deligne. La conjecture de Weil. I. *Inst. Hautes Études Sci. Publ. Math.*, (43):273–307, 1974.

[9] J.-M. Deshouillers and H. Iwaniec. Power mean values of the Riemann zeta function. *Mathematika*, 29(2):202–212, 1982.

[10] J.-M. Deshouillers and H. Iwaniec. Power mean-values for Dirichlet's polynomials and the Riemann zeta-function. II. *Acta Arith.*, 43(3):305–312, 1984.

[11] Jean-Marc Deshouillers and Henryk Iwaniec. Kloosterman sums and Fourier coefficients of cusp forms. *Invent. Math.*, 70(2):219–288, 1982.

[12] Sary Drappeau, Kyle Pratt, and Maksym Radziwiłł. One-level density estimates for Dirichlet L-functions with extended support. *Algebra Number Theory*, 17(4):805–830, 2023.

[13] William Duke, John Friedlander, and Henryk Iwaniec. Bilinear forms with Kloosterman fractions. *Invent. Math.*, 128(1):23–43, 1997.

[14] Étienne Fouvry, Emmanuel Kowalski, and Philippe Michel. Algebraic trace functions over the primes. *Duke Math. J.*, 163(9):1683–1736, 2014.

[15] Étienne Fouvry, Emmanuel Kowalski, Philippe Michel, and Will Sawin. Lectures on applied ℓ-adic cohomology. In *Analytic methods in arithmetic geometry*, volume 740 of *Contemp. Math.*, pages 113–195. Amer. Math. Soc., [Providence], RI, [2019] ©2019.

[16] William Fulton and Joe Harris. *Representation theory*, volume 129 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1991. A first course, Readings in Mathematics.

[17] Lasse Grimmelt and Jori Merikoski. On the greatest prime factor and uniform equidistribution of quadratic polynomials. *Preprint, arXiv:2505.00493*, 2025.

[18] Larry Guth and James Maynard. New large value estimates for Dirichlet polynomials. *Ann. of Math., to appear. Preprint, arXiv:2405.20552*, 2024.

[19] H. A. Helfgott. Growth and generation in $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. *Ann. of Math. (2)*, 167(2):601–623, 2008.

[20] I. Martin Isaacs. *Character theory of finite groups*. AMS Chelsea Publishing, Providence, RI, 2006. Corrected reprint of the 1976 original [Academic Press, New York; MR0460423].

[21] Henryk Iwaniec and Emmanuel Kowalski. *Analytic number theory*, volume 53. American Mathematical Society, Providence, RI, 2021.

[22] 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):578–621, 2023.

[23] Henry H. Kim. Functoriality for the exterior square of GL_4 and the symmetric fourth of GL_2. *J. Amer. Math. Soc.*, 16(1):139–183, 2003. With Appendix 1 by Dinakar Ramakrishnan and Appendix 2 by Kim and Peter Sarnak.

[24] Emmanuel Kowalski, Philippe Michel, and Will Sawin. Bilinear forms with Kloosterman sums and applications. *Ann. of Math. (2)*, 186(2):413–500, 2017.

[25] 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)*, 21:1453–1530, 2020.

[26] Philip C. Kutzko. The characters of the binary modular congruence group. *Bull. Amer. Math. Soc.*, 79:702–704, 1973.

[27] 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):334–383, 479, 1980.

[28] James Maynard. Primes in Arithmetic Progressions to Large Moduli I: Fixed Residue Classes. *Mem. Amer. Math. Soc.*, 306(1542), 2025.

[29] Djordje Milićević, Xinhua Qin, and Xiaosheng Wu. Bilinear forms with Kloosterman sums and moments of twisted L-functions. *arXiv preprint*, November 2025.
[30] Nikolay G. Moshchevitin and Ilya D. Shkredov. On a modular form of Zaremba's conjecture. *Pacific J. Math.*, 309(1):195–211, 2020.
[31] Alexandre Nobs and Jürgen Wolfart. Die irreduziblen Darstellungen der Gruppen $SL_2(Z_p)$, insbesondere $SL_2(Z_p)$. II. *Comment. Math. Helv.*, 51(4):491–526, 1976.
[32] 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.
[33] Alexandru Pascadi. On the exponents of distribution of primes and smooth numbers. *Preprint, arXiv:2505.00653*, 2025.
[34] Atle Selberg. On the estimation of Fourier coefficients of modular forms. In *Proc. Sympos. Pure Math., Vol. VIII*, pages 1–15. Amer. Math. Soc., Providence, RI, 1965.
[35] Jean-Pierre Serre. *Linear representations of finite groups*, volume Vol. 42 of *Graduate Texts in Mathematics*. Springer-Verlag, New York-Heidelberg, french edition, 1977.
[36] Joseph A. Shalika. Representation of the two by two unimodular group over local fields. In *Contributions to automorphic forms, geometry, and number theory*, pages 1–38. Johns Hopkins Univ. Press, Baltimore, MD, 2004.
[37] I. D. Shkredov. On asymptotic formulae in some sum-product questions. *Trans. Moscow Math. Soc.*, 79:231–281, 2018.
[38] I. D. Shkredov. Modular hyperbolas and bilinear forms of Kloosterman sums. *J. Number Theory*, 220:182–211, 2021.
[39] Igor E. Shparlinski. On sums of Kloosterman and Gauss sums. *Trans. Amer. Math. Soc.*, 371(12):8679–8697, 2019.
[40] Igor E. Shparlinski and Tianping Zhang. Cancellations amongst Kloosterman sums. *Acta Arith.*, 176(3):201–210, 2016.
[41] Shunichi Tanaka. Irreducible representations of the binary modular congruence groups mod p^λ. *J. Math. Kyoto Univ.*, 7:123–132, 1967.
[42] Audrey Terras. *Fourier analysis on finite groups and applications*, volume 43 of *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 1999.
[43] Berke Topacogullari. The shifted convolution of generalized divisor functions. *Int. Math. Res. Not. IMRN*, (24):7681–7724, 2018.
[44] Jie Wu and Ping Xi. Arithmetic exponent pairs for algebraic trace functions and applications. *Algebra Number Theory*, 15(9):2123–2172, 2021. With an appendix by Will Sawin.
[45] Xiaosheng Wu. The fourth moment of Dirichlet L-functions at the central value. *Math. Ann.*, 387(3-4):1199–1248, 2023.
[46] Ping Xi. Ternary divisor functions in arithmetic progressions to smooth moduli. *Mathematika*, 64(3):701–729, 2018.
[47] Matthew P. Young. The fourth moment of Dirichlet L-functions. *Ann. of Math. (2)*, 173(1):1–50, 2011.

Mathematisches Institut, Endenicher Allee 60, 53115 Bonn, Germany

Email address: alexpascadi@gmail.com