EFFECTIVE EQUIDISTRIBUTION OF PRIMITIVE RATIONAL POINTS ON EXPANDING HOROSPHERES

DANIEL EL-BAZ, MIN LEE, AND ANDREAS STRÖMBERGSSON

Abstract. We prove an effective version of a result due to Einsiedler, Mozes, Shah and Shapira on the asymptotic distribution of primitive rational points on expanding closed horospheres in the space of lattices. Key ingredients of our proof include recent bounds on matrix Kloosterman sums due to Erdélyi and Tóth, results by Clozel, Oh and Ullmo on the effective equidistribution of Hecke points, and Rogers' integration formula in the geometry of numbers. As an application of the main theorem, we also obtain a result on the limit distribution of the number of small solutions of a random system of linear congruences to a large modulus. Furthermore, as a by-product of our proofs, we obtain a sharp bound on the number of nonsquare matrices over a finite field $\mathbb{F}_p$ with small entries and of a given size and rank.

Contents

Date: July 18, 2023.

We are grateful to Árpád Tóth and Márton Erdélyi for sharing their preprint on the matrix Kloosterman sum early with us and several conversations. We are also grateful to Igor Shparlinski for making us aware of his paper with Ahmadi, [AS07]. We would further like to express our thanks for a Heilbronn Focused Research Grant and the hospitality of the Heilbronn institute in Bristol. D.E. is supported by the Austrian Science Fund (FWF), Projects P-34763 and Y-901. M.L. is supported by a Royal Society University Research Fellowship. A.S. is supported by the Knut and Alice Wallenberg Foundation.

1. Introduction

1.1. **Setup.** Let $1 \le n \le d$, $\mathrm{G} = \mathrm{SL}_{d+n}(\mathbb{R})$ and $\Gamma = \mathrm{SL}_{d+n}(\mathbb{Z})$. Our discussion will take place in the homogeneous space $\Gamma\backslash\mathrm{G}$. We will often view an element $g \in \mathrm{G}$ as a block matrix, $g = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$, where A, B, C, D are real matrices of dimensions $d \times d$, $d \times n$, $n \times d$ and $n \times n$, respectively. In particular, for $V \in \mathrm{M}_{n\times d}(\mathbb{R})$ (that is, V being a real matrix of dimension $n \times d$), let us write

$$n_+(V) := \begin{pmatrix} I_d & V \\ \mathbf{0} & I_n \end{pmatrix} \in \mathrm{G}. \tag{1.1}$$

Also, for $y > 0$, let

$$D(y) := \begin{pmatrix} y^{-\frac{n}{d}} I_d & \mathbf{0} \\ \mathbf{0} & y I_n \end{pmatrix} \in \mathrm{G}. \tag{1.2}$$

For each $V \in \mathrm{M}_{d\times n}(\mathbb{R})$, the point $\Gamma\, n_+(V)$ in $\Gamma\backslash\mathrm{G}$ depends only on V mod $\mathrm{M}_{d\times n}(\mathbb{Z})$; hence the map $V \mapsto \Gamma\, n_+(V)$ factors through a map

$$\widetilde{n}_+ : \mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z}) \to \Gamma\backslash\mathrm{G}. \tag{1.3}$$

In fact $\widetilde{n}_+$ is a smooth embedding of the dn-dimensional torus $\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})$; its image is a closed horosphere in $\Gamma\backslash\mathrm{G}$, which we call $\mathfrak{H}_1$. More generally, let $\mathfrak{H}_y$ be $\mathfrak{H}_1$ translated by $D(y)$:

$$\mathfrak{H}_y = \mathfrak{H}_1 D(y) = \big\{\widetilde{n}_+(V)D(y) \,:\, V \in \mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})\big\}.$$

These $\mathfrak{H}_y$ form a family of closed horospheres in $\Gamma\backslash\mathrm{G}$, which expand as y increases. It is well-known that as $y \to \infty$, the $\mathfrak{H}_y$ become equidistributed in $\Gamma\backslash\mathrm{G}$ with respect to the G-invariant probability measure.

Our main object of study is a very special finite subset of the closed horosphere $\mathfrak{H}_y$, appearing when y is an integer. To describe this set, let H be the following subgroup of G:

$$\mathrm{H} = \left\{ \begin{pmatrix} A & \mathbf{0} \\ U & I_n \end{pmatrix} : A \in \mathrm{SL}_d(\mathbb{R}),\, U \in \mathrm{M}_{n\times d}(\mathbb{R}),\, A = I_n \text{ if } n = d \right\}. \tag{1.4}$$

Then $\Gamma\backslash\Gamma\mathrm{H}$ is a closed embedded submanifold of $\Gamma\backslash\mathrm{G}$, and $\Gamma\backslash\Gamma\mathrm{H}$ has the structure of a torus fiber bundle over $\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})$. Let $\mathfrak{S}_y$ be the *intersection* of $\Gamma\backslash\Gamma\mathrm{H}$ and $\mathfrak{H}_y$.

Lemma 1.1. *The set $\mathfrak{S}_y$ is empty unless y is an integer. For $y = q$ a positive integer, the set $\mathfrak{S}_q$ consists exactly of the points $\widetilde{n}_+(q^{-1}R)D(q)$ where R runs through all matrices in $\mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})$ with the property that the rows of R generate $(\mathbb{Z}/q\mathbb{Z})^n$.*

(Here, naturally, $\mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})$ denotes the group of $d \times n$ matrices with entries in $\mathbb{Z}/q\mathbb{Z}$; note also that for any $R \in \mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})$, $q^{-1}R$ is a well-defined point in the torus $\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})$.)

We prove Lemma 1.1 in Section 2 (see also [EMSS16, Sec. 2]). As in [EMSS16, Definition 1.1], for a positive integer q, let us call a matrix $R \in \mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})$ *(q-)primitive* if the rows of R generate $(\mathbb{Z}/q\mathbb{Z})^n$. We will also say that a matrix $R \in \mathrm{M}_{d\times n}(\mathbb{Z})$ is q-primitive if its reduction mod q is q-primitive. Let $\mathcal{R}_q$ be the set of all primitive matrices in $\mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})$. Then Lemma 1.1 says that

$$\mathfrak{S}_q = \big\{\widetilde{n}_+(q^{-1}R)D(q) \,:\, R \in \mathcal{R}_q\big\}. \tag{1.5}$$

We call $\mathfrak{S}_q$ the set of primitive rational points on $\mathfrak{H}_q$.

We are interested in the behavior of this point set $\mathfrak{S}_q$ for q large. It was proved by Einsiedler, Mozes, Shah and Shapira [EMSS16] that $\mathfrak{S}_q$ becomes equidistributed in $\Gamma\backslash\Gamma\mathrm{H}$ with respect to the H-invariant probability measure, as $q \to \infty$. In fact, confirming a conjecture by Marklof, they proved the much stronger fact that the point set

$$\widetilde{\mathfrak{S}}_q := \big\{ \big(q^{-1}R, \widetilde{n}_+(q^{-1}R)D(q)\big) \ : \ R \in \mathcal{R}_q \big\} \tag{1.6}$$

becomes (jointly) equidistributed in the product space $(\mathbb{R}/\mathbb{Z})^{dn} \times \Gamma\backslash\Gamma\mathrm{H}$, where we have identified the torus $\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})$ with $(\mathbb{R}/\mathbb{Z})^{dn}$.

In the present paper we give a new proof of this equidistribution result which relies on harmonic analysis and number theory, spectral theory of automorphic forms, the newly studied object of matrix Kloosterman sums, and Rogers' integration formula in the geometry of numbers. Our proof leads to an *effective* version of the equidistribution result, that is, we obtain explicit information on how quickly the equidistribution takes place as $q \to \infty$.

1.2. **Informal statement of the main result.** Given a function $f : (\mathbb{R}/\mathbb{Z})^{dn} \times \Gamma\backslash\Gamma\mathrm{H} \to \mathbb{R}$, set

$$\mathcal{A}_q(f) = \frac{1}{\#\mathcal{R}_q} \sum_{R\in\mathcal{R}_q} f(q^{-1}R, \widetilde{n}_+(q^{-1}R)D(q)). \tag{1.7}$$

Then the statement of Einsiedler–Mozes–Shah–Shapira's theorem is precisely that whenever f is bounded and continuous, $\mathcal{A}_q(f)$ converges to the integral of f as $q \to \infty$. By standard approximation arguments, it is equivalent to state that this convergence holds whenever f is smooth and compactly supported.

Our main result is an effective version of that result, with a power-saving error term, meaning that we prove, for every $1 \le n \le d$, and for any sufficiently smooth f,

$$\mathcal{A}_q(f) = \int_{(\mathbb{R}/\mathbb{Z})^{dn}\times\Gamma\backslash\Gamma\mathrm{H}} f \, dT \, d\mu_{\mathrm{H}} + O_d\big(S(f)\, q^{-\delta}\big) \tag{1.8}$$

as $q \to \infty$, where dT is the usual Lebesgue measure on $(\mathbb{R}/\mathbb{Z})^{dn}$, μ_{H} is the H-invariant probability measure on $\Gamma\backslash\Gamma\mathrm{H}$, and $S(f)$ is a certain Sobolev norm of f, defined in terms of the L^2 and L^∞ norms of f and its first several derivatives (see §1.3), while $\delta > 0$ is a fixed constant.

In the special case $n = 1$ such an effective equidistribution result was obtained in [LM17] (for $d = 2$) and [EBHL22] (for general d). Our main theorem, stated more precisely in the next section, finally provides an effective version of the general case of the Einsiedler–Mozes–Shah–Shapira theorem.

1.3. **Formal statement of the main result.** In order to state our result we need to introduce certain Sobolev norms of functions on homogeneous spaces (compare [Ven10, Sec. 2.9.2]). Suppose Λ is a lattice in a connected Lie group L, and let μ be the L-invariant probability measure on $\Lambda\backslash L$. Fix, once and for all, a linear basis $\mathcal{B}$ for the Lie algebra of L. Let $k \ge 0$ be an integer. For $f \in \mathrm{C}^k(\Lambda\backslash L)$ and $1 \le p \le \infty$ (in fact we will only consider $p = 2$ and $p = \infty$), we define the Sobolev norm of f

$$S_{p,k}(f) = \sum_{\mathrm{ord}(\mathcal{D})\le k} \|\mathcal{D}f\|_{\mathrm{L}^p(\Lambda\backslash L,\mu)}, \tag{1.9}$$

where $\mathcal{D}$ runs through all monomials in $\mathcal{B}$ of order $\le k$. Here $\mathcal{D}$ acts on f by right differentiation:

$$Xf(g) = \frac{d}{dt} f(g\exp(tX))\big|_{t=0} \text{ for any } X \in \mathcal{B}. \tag{1.10}$$

It should be noted that changing the basis $\mathcal{B}$ only distorts $S_{p,k}$ by a bounded factor.

We write $\mathrm{C}_b^k(\Lambda\backslash L)$ for the space of functions in $\mathrm{C}^k(\Lambda\backslash L)$ which have all derivatives of order $\le k$ bounded, i.e.,

$$\mathrm{C}_b^k(\Lambda\backslash L) = \big\{ f \in \mathrm{C}^k(\Lambda\backslash L) \,:\, S_{\infty,k}(f) < \infty \big\}.$$

It will be convenient to also introduce, in a non-standard but elementary way, *fractional* Sobolev norms (cf. [SV05, Lemma 2]): For any real number $k < \kappa < k+1$ and $f \in \mathrm{C}^{k+1}(\Lambda\backslash L)$, we set

$$S_{p,\kappa}(f) = S_{p,k}(f)^{k+1-\kappa} S_{p,k+1}(f)^{\kappa-k}. \tag{1.11}$$

In the statement of the following theorem, the above formalism is applied for the homogeneous space $(\mathbb{R}/\mathbb{Z})^{dn} \times \Gamma\backslash\Gamma\mathrm{H}$, that is, with $\Lambda = \mathbb{Z}^{dn} \times (\Gamma \cap \mathrm{H})$ and $L = \mathbb{R}^{dn} \times \mathrm{H}$.

Let θ be the constant towards the Ramanujan conjecture for Maass wave forms on $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$, which asserts $\theta = 0$. The current best bound is $\theta \le 7/64$, due to Kim and Sarnak [Kim03, Appendix 2].

Theorem 1.2. *For the given positive integers $1 \le n \le d$, let*

$$\kappa = 2dn; \quad \vartheta = \begin{cases} n-1 & (\text{if } n > 1), \\ \frac{1}{2} & (\text{if } n = 1); \end{cases} \quad \text{and} \quad \begin{cases} \kappa' = \frac{1}{2}(d^2-1); \quad \vartheta' = \frac{1}{2}\min(n, d-n) & \\ \qquad\qquad\qquad\qquad\qquad\qquad (\text{if } n < d \text{ and } d \ge 3); & \\ \kappa' = \frac{3}{2}; \quad \vartheta' = \frac{1}{2} - \theta \qquad (\text{if } n = 1 \text{ and } d = 2); & \\ \kappa' = \kappa; \quad \vartheta' = \vartheta \qquad\quad (\text{if } n = d). & \end{cases}$$

Also let k be the smallest integer greater than both κ and κ'.

Then for any $0 < \varepsilon < \frac{1}{2}$, $f \in \mathrm{C}_b^k((\mathbb{R}/\mathbb{Z})^{dn} \times \Gamma\backslash\Gamma\mathrm{H})$, and any positive integer q,

$$\mathcal{A}_q(f) = \int_{(\mathbb{R}/\mathbb{Z})^{dn}} \int_{\Gamma\backslash\Gamma\mathrm{H}} f(T,g)\, d\mu_{\mathrm{H}}(g)\, dT + O\Big(S_{\infty,\kappa+\varepsilon}(f)\, q^{-\vartheta+\varepsilon} + S_{2,\kappa'+\varepsilon}(f)\, q^{-\vartheta'+\varepsilon} \Big), \tag{1.12}$$

where the implied constant only depends on d and ε.

Remark 1.3. In the statement of Theorem 1.2, it should be noted that we always have $\kappa + \varepsilon < k$ and $\kappa' + \varepsilon < k$, and thus both the Sobolev norms $S_{\infty,\kappa+\varepsilon}(f)$ and $S_{2,\kappa'+\varepsilon}(f)$ are defined and finite. It should also be noted that the introduction of κ' and ϑ' in the case $n = d$ is only a notational convenience, allowing a simple comprehensive statement of (1.12). Indeed, in that case the error term in (1.12) reduces to $O\big(S_{\infty,\kappa+\varepsilon}(f)\, q^{-\vartheta+\varepsilon}\big)$, since $S_{2,\kappa+\varepsilon}(f) \le S_{\infty,\kappa+\varepsilon}(f)$.

1.4. **Discussion of the result and layout of the proof.** As we have already mentioned, the problem of studying the limiting distribution of the primitive rational points (1.5) on the expanding closed horosphes $\mathfrak{H}_q$, was raised by Marklof, specifically in [Mar10a] when $n = 1$. Marklof proved an averaged version of the equidistribution of primitive rational points on expanding horospheres and used it to obtain a limiting distribution result for Frobenius numbers. His work was made effective, using estimates on the decay of matrix coefficients, by Li [Li15].

The proof of Marklof's conjecture by Einsiedler, Mozes, Shah and Shapira [EMSS16] uses techniques from homogeneous dynamics and relies in particular on measure-classification theorems due to Ratner [Rat91], extended by Shah [Sha98], which are inherently ineffective.

For $n \ge 2$, the result of Theorem 1.2, with *any* effective rate of equidistribution, is new. It is also worth noticing that in the special case $n = 1$, our error bound is stronger than those in [LM17] (for $n = 1$ and $d = 2$) and in [EBHL22] (for $n = 1$ and $d \ge 2$). More precisely, for $n = 1$ and $d \ge 3$, the error bound in Theorem 1.2 is $\big(S_{\infty,\kappa+\varepsilon}(f) + S_{2,\kappa'+\varepsilon}(f)\big) \cdot q^{-\frac{1}{2}+\varepsilon}$ with $\kappa = 2d$ and $\kappa' = \frac{1}{2}(d^2-1)$; this is stronger than the error term in [EBHL22, Theorem 1.1], both in terms of the Sobolev norm and the power of q.[1] For $n = 1$ and $d = 2$, the error bound in

[1] One may note that the q-exponent in [EBHL22, Theorem 1.1] tends to our exponent $-\frac{1}{2}+\varepsilon$ if one lets the order of the Sobolev norm tend to $+\infty$.

Theorem 1.2 is $S_{\infty,4+\varepsilon}(f)\, q^{-\frac{1}{2}+\varepsilon} + S_{2,\frac{3}{2}+\varepsilon}(f)\, q^{-\frac{1}{2}+\theta+\varepsilon}$, which is stronger than the bound in both [LM17, Theorem 1.3] and [EBHL22, Remark 1.2]. Finally for $n = d = 1$ the error bound in Theorem 1.2 is $S_{\infty,2+\varepsilon}(f)\, q^{-\frac{1}{2}+\varepsilon}$. That case is quite easy; see [Mar10b] and [EMSS16, Sec. 2.1] (neither of those include the precise error term, but that is not at all difficult).

The basic set-up of the proof of Theorem 1.2 is similar to the one in both [LM17] and [EBHL22]: In Lemma 2.2 we give a parametrization of the set $\mathcal{R}_q$ of primitive matrices in terms of $\Gamma^0(q)\backslash \mathrm{SL}_d(\mathbb{Z})$ and $\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$, where $\Gamma^0(q)$ is a certain congruence subgroup of $\mathrm{SL}_d(\mathbb{Z})$ (for $n = 1$ this was done in [EBHL22, Lemma 2.2]). Furthermore, our first step is to Fourier expand the given test function on $(\mathbb{R}/\mathbb{Z})^{dn} \times \Gamma\backslash\Gamma\mathrm{H}$, both with respect to the variable in the torus $(\mathbb{R}/\mathbb{Z})^{dn}$ and with respect to the torus fiber variable in $\Gamma\backslash\Gamma\mathrm{H}$; see Section 3. Then the main term in (1.12) is obtained by using the asymptotic equidistribution of certain Hecke orbits in $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$, and for this an optimal error term is provided by the results of Clozel, Oh and Ullmo [COU01]; see Section 4.

However, the task of bounding the contribution from the remaining sums is significantly more challenging in the present paper where we deal with general $n \geq 1$. Here our first step is to apply bounds on the newly studied "matrix Kloosterman sums". For prime moduli, key bounds on these matrix Kloosterman sums have been proved by Erdélyi and Tóth [ET21]; for the case of higher prime power moduli we prove non-trivial bounds in Section 5.3, by elementary but somewhat complicated computations. Similar bounds have also, independently, been obtained by Erdélyi, Tóth and Zábrádi in the recent paper [ETZ22]. The majorizing sum which arises from the application of the bounds on matrix Kloosterman sums is still non-trivial to control. At this point we make use of a Hecke operator interpretation followed by an application of an integration formula by Rogers [Rog55] in the geometry of numbers, to arrive at a satisfactory final bound. This is carried out in Section 7.5. The usage of Rogers' integration formula in the present method is also the reason for our improvement of the error bounds in [LM17] and [EBHL22] in the case $n = 1$.

1.5. **Consequences of our main theorem and its proof.** The case $n = 1$ of the equidistribution result of Einsiedler, Mozes, Shah and Shapira is known to have applications to the distribution of Frobenius numbers [Mar10a], the distribution of shapes of lattices [EMSS16], and to the distribution of metric parameters of random Cayley graphs of cyclic groups [MS13]. Naturally, an effective version of this equidistribution result can be expected to lead to information on the rate of convergence in these applications; in [EBHL22, Cor. 5.1] this was carried out for the case of the diameter of random Cayley graphs of cyclic groups. (Our improved error bound in Theorem 1.2 should lead to an improved exponent η_d in [EBHL22, Cor. 5.1].)

In the present article, in Section 8.1, we give a new application of the equidistribution result, this time for arbitrary $1 \leq n \leq d$: We obtain the limit distribution of the number of small solutions of a random system of linear congruences to a large modulus. This can be seen as a variation, and in a sense a refinement, of results by Strömbergsson and Venkatesh [SV05] (see Remark 8.2).

Furthermore, while first attempting to follow the strategy deployed in [EBHL22], we came across an elementary counting problem in linear algebra, for which we were however unable to find an elementary solution. That resulted in the technique we instead follow in Section 7.5, using a Hecke operator interpretation followed by an application of Rogers' integration formula. As a by-product of our proof, we are able to satisfactorily solve the linear algebra problem, whose statement is as follows:

Problem 1.4. *For integers $1 \leq r < n < d$, a prime $p \geq 3$ and an integer $1 \leq b \leq \frac{p-1}{2}$, estimate the growth rate of*

$$\#\{A \in \mathrm{M}_{d\times n}(\mathbb{Z}) \,:\, \|A\|_\infty \leq b \text{ and } \operatorname{rank}(A \bmod p) = r\} \tag{1.13}$$

as p gets large. Here $\|A\|_\infty$ denotes the maximum of the absolute values of the entries of A.

Problem 1.4 has been studied previously, for all values of n and d, by Ahmadi and Shparlinski [AS07], who obtained an asymptotic formula for (1.13) valid as $p \to \infty$ with b in a restricted range. In Section 8.2 we prove a sharp estimate on (1.13), valid for arbitrary b and p.

2. The primitive rational points on $\mathfrak{H}_q$

As in the introduction, we keep $1 \le n \le d$ fixed.

Proof of Lemma 1.1. Let q be a positive real number, and assume that $\mathfrak{S}_q$ is non-empty. This means that there is some $V \in \mathrm{M}_{d\times n}(\mathbb{R})$ such that $n_+(V)D(q) \in \Gamma\mathrm{H}$, that is

$$\begin{pmatrix} q^{-n/d}I_d & qV \\ \mathbf{0} & qI_n \end{pmatrix} = \gamma \begin{pmatrix} A & \mathbf{0} \\ U & I_n \end{pmatrix} \tag{2.1}$$

for some $\gamma \in \Gamma$, $A \in \mathrm{SL}_d(\mathbb{R})$ (if $n = d$: $A = I_n$) and $U \in \mathrm{M}_{n\times d}(\mathbb{R})$. All the entries in the last n columns of the matrix in the right hand side are integers; hence q must be an integer, and $V = q^{-1}R$ for some $R \in \mathrm{M}_{d\times n}(\mathbb{Z})$. Also, left-multiplying the relation in (2.1) by γ^{-1} and inspecting the bottom right $n \times n$ submatrix ($= I_n$), it follows that each of the standard basis vectors $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ of $\mathbb{R}^n$ is an integer linear combination of the row vectors of R and $q\boldsymbol{e}_1, \dots, q\boldsymbol{e}_n$. Hence R is q-primitive.

Conversely, assume that q is a positive integer and $R \in \mathrm{M}_{d\times n}(\mathbb{Z})$ is q-primitive. Then the homomorphism $\boldsymbol{a} \mapsto \boldsymbol{a}R \bmod q$ from $\mathbb{Z}^d$ to $\mathbb{Z}^n/q\mathbb{Z}^n$ is surjective; hence its kernel K is a subgroup of $\mathbb{Z}^d$ of index q^n. Let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_d$ be a positively oriented $\mathbb{Z}$-basis of K, where if $n = d$ we require $\boldsymbol{a}_j := q\boldsymbol{e}_j$ for $j = 1, \dots, d$ (this is ok since $K = q\mathbb{Z}^n$ if $n = d$). Let A' be the $d \times d$ matrix with row vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_d$. Then $\det(A') = q^n$, and for each $\boldsymbol{a}_j$ there is a unique $\boldsymbol{b}_j \in \mathbb{Z}^n$ such that $\boldsymbol{a}_j R + q\boldsymbol{b}_j = \mathbf{0}$. Also, since R is q-primitive, there exist $\boldsymbol{c}_1, \dots, \boldsymbol{c}_n \in \mathbb{Z}^{d+n}$ such that $\boldsymbol{c}_j \begin{pmatrix} R \\ qI_n \end{pmatrix} = \boldsymbol{e}_j$ ($j = 1, \dots, n$). Now let η be the square matrix with row vectors $(\boldsymbol{a}_1, \boldsymbol{b}_1), \dots, (\boldsymbol{a}_d, \boldsymbol{b}_d), \boldsymbol{c}_1, \dots, \boldsymbol{c}_n$; then

$$\eta \begin{pmatrix} q^{-n/d}I_d & R \\ \mathbf{0} & qI_n \end{pmatrix} = \begin{pmatrix} q^{-n/d}A' & \mathbf{0} \\ U & I_n \end{pmatrix}$$

for some $U \in \mathrm{M}_{n\times d}(\mathbb{R})$. Here $\det(q^{-n/d}A') = 1$, and if $n = d$ then $q^{-1}A' = I_n$. Hence the above matrix lies in H, and $\det(\eta) = 1$, i.e. $\eta \in \Gamma$. Therefore $\widetilde{n}_+(q^{-1}R)D(q) \in \mathfrak{S}_q$. □

It will be useful to know the cardinality of $\mathfrak{S}_q$, i.e. the cardinality of $\mathcal{R}_q$. We write $\mathbb{Z}^+$ for the set of positive integers.

Lemma 2.1. $\forall q \in \mathbb{Z}^+, \#\mathcal{R}_q = q^{dn}\prod_{p|q}\prod_{j=d+1-n}^{d}(1 - p^{-j})$.

Proof. It follows from the Chinese Remainder Theorem that the function $q \mapsto \#\mathcal{R}_q$ is multiplicative; hence it suffices to prove the lemma when q is a prime power, say $q = p^r$ ($r \ge 1$). Now, for any $R \in \mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})$ such that $R \bmod p$ is p-primitive, $R \bmod p$ has some $n\times n$ submatrix which belongs to $\mathrm{GL}_n(\mathbb{Z}/p\mathbb{Z})$; therefore the determinant of the corresponding submatrix of R itself is a unit in $\mathbb{Z}/q\mathbb{Z}$, viz., that submatrix belongs to $\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$ and R is q-primitive. Hence $R \in \mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})$ is q-primitive if and only if $R \bmod p$ is p-primitive, and so $\#\mathcal{R}_q = p^{(r-1)dn}\#\mathcal{R}_p$.

It remains to prove the lemma in the case $q = p$, a prime. Let us write $\mathbb{F}_p$ for the field $\mathbb{Z}/p\mathbb{Z}$. A matrix in $\mathrm{M}_{d\times n}(\mathbb{F}_p)$ is p-primitive if and only if it has full rank, that is, if and only if its columns are linearly independent. Note that there are exactly $p^d - 1$ full rank matrices in $\mathrm{M}_{d\times 1}(\mathbb{F}_p)$. Furthermore, for any $1 \le \ell < d$, given any matrix $A \in \mathrm{M}_{d\times \ell}(\mathbb{F}_p)$ of full rank, the column span of A has cardinality p^ℓ, and hence there are exactly $p^d - p^\ell$ ways to choose a column to the right of A to form a full rank matrix in $\mathrm{M}_{d\times(\ell+1)}(\mathbb{F}_p)$. Hence $\#\mathcal{R}_p = \prod_{\ell=0}^{n-1}(p^d - p^\ell)$, and the lemma is proved. □

In the next lemma we give a parametrization of $\mathcal{R}_q$ which will be crucial in our proof of the main theorem. If $n < d$, then we define $\Gamma^0(q)$ to be the following congruence subgroup of $\mathrm{SL}_d(\mathbb{Z})$:

$$\Gamma^0(q) = \left\{ \begin{pmatrix} A & B \\ {}^tC & D \end{pmatrix} \in \mathrm{SL}_d(\mathbb{Z}) \,:\, A \in \mathrm{M}_{d-n},\, B, C \in \mathrm{M}_{(d-n)\times n},\, D \in \mathrm{M}_n,\, B \equiv \mathbf{0} \bmod q \right\}, \tag{2.2}$$

and we fix a set $\mathcal{B}_q$ of representatives for $\Gamma^0(q)\backslash \mathrm{SL}_d(\mathbb{Z})$. When $d = n$, we set

$$\Gamma^0(q) = \mathrm{SL}_n(\mathbb{Z}) \tag{2.3}$$

and $\mathcal{B}_q := \{I_n\}$. The following lemma generalizes [EBHL22, Lemma 2.2].

Lemma 2.2. *The map*

$$\mathcal{B}_q \times \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z}) \to \mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z}) \tag{2.4}$$

given by

$$\langle \gamma, U\rangle \mapsto \gamma^{-1} \begin{pmatrix} \mathbf{0} \\ U \end{pmatrix} \qquad (\gamma \in \mathcal{B}_q,\, U \in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})), \tag{2.5}$$

is a bijection onto $\mathcal{R}_q \subset \mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})$.

(In the case $n = d$, the matrix "$\left(\begin{smallmatrix} \mathbf{0} \\ U \end{smallmatrix}\right)$" in (2.5) should be interpreted as "U".)

Proof. If $n = d$ then $\mathcal{R}_q = \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$ and $\mathcal{B}_q = \{I_n\}$ and the lemma is trivial.

From now on we assume that $1 \le n < d$. It is clear that the image of the map in (2.5) is contained in $\mathcal{R}_q$. To prove that the map is surjective, let $R \in \mathcal{R}_q$ be given. Then by the Smith Normal Form Theorem, there exist $\delta \in \mathrm{GL}_d(\mathbb{Z}/q\mathbb{Z})$ and $\eta \in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$ and a diagonal matrix $D \in \mathrm{M}_n(\mathbb{Z}/q\mathbb{Z})$ such that

$$R = \delta \begin{pmatrix} \mathbf{0} \\ D \end{pmatrix} \eta = \delta \begin{pmatrix} \mathbf{0} \\ D\eta \end{pmatrix}.$$

Note that the above identity remains true if we replace δ by $\delta\,\mathrm{diag}[u, 1, \cdots, 1]$ for any $u \in (\mathbb{Z}/q\mathbb{Z})^\times$; hence we may arrange that $\delta \in \mathrm{SL}_n(\mathbb{Z}/q\mathbb{Z})$. The reduction map from $\mathrm{SL}_n(\mathbb{Z})$ to $\mathrm{SL}_n(\mathbb{Z}/q\mathbb{Z})$ is surjective (see, e.g., the proof of [Shi94, Lemma 1.38]); hence there exists a lift $\delta' \in \mathrm{SL}_n(\mathbb{Z})$ of δ. Let γ be the unique element in $\mathcal{B}_q \cap \Gamma^0(q)\delta'^{-1}$; then $\delta' = \gamma^{-1} \begin{pmatrix} A & B \\ C & D' \end{pmatrix}$ for some $\begin{pmatrix} A & B \\ C & D' \end{pmatrix} \in \Gamma^0(q)$, and so

$$R = \gamma^{-1} \begin{pmatrix} \mathbf{0} \\ D'D\eta \end{pmatrix} \qquad \text{in } \mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z}).$$

But $R \in \mathcal{R}_q$ implies $\gamma R \in \mathcal{R}_q$; hence the rows of $D'D\eta$ generate $(\mathbb{Z}/q\mathbb{Z})^n$, that is, $D'D\eta \in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$, and we have thus proved that R lies in the image of the map in (2.5).

It remains to verify that the map is injective. Thus we assume that the two pairs $\langle \gamma, U\rangle$ and $\langle \gamma', U'\rangle$ map to the same element in $\mathcal{R}_q$. Then

$$\gamma'\gamma^{-1} \begin{pmatrix} \mathbf{0} \\ U \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ U' \end{pmatrix} \qquad \text{in } \mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z}).$$

This forces $\gamma'\gamma^{-1} \in \Gamma^0(q)$, and since $\gamma, \gamma' \in \mathcal{B}_q$ it follows that $\gamma = \gamma'$. Hence also $U = U'$, and the injectivity is proved. □

In the next lemma we give a formula which will be useful when applying Lemma 2.2 to re-express the sum in (1.7). Let us introduce, for $U \in \mathrm{M}_{n\times d}(\mathbb{R})$,

$$n_-(U) = \begin{pmatrix} I_d & \mathbf{0} \\ U & I_n \end{pmatrix}. \tag{2.6}$$

We also introduce the map

$$\widetilde{n}_- : \mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z}) \to \Gamma\backslash G \tag{2.7}$$

by setting $\widetilde{n}_-(U) := \Gamma n_-(U')$ where U' is any lift to $\mathrm{M}_{n\times d}(\mathbb{R})$ of $U \in \mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})$ (this is analogous to $\widetilde{n}_+$ in (1.3)).

Lemma 2.3. *Assume that $1 \le n \le d$. Let $\gamma \in \mathcal{B}_q$ and $U \in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$, and set*

$$R = \gamma^{-1}\begin{pmatrix}\mathbf{0}\\ U\end{pmatrix} \in \mathcal{R}_q \qquad \text{and} \qquad S = \begin{pmatrix}\mathbf{0} & U^{-1}\end{pmatrix} \in \mathrm{M}_{n\times d}(\mathbb{Z}/q\mathbb{Z}); \tag{2.8}$$

$$D_q = \begin{cases} q^{-\frac{n}{d}}\begin{pmatrix} I_{d-n} & \\ & qI_n\end{pmatrix} & \text{when } n < d,\\ I_n & \text{when } n = d.\end{cases} \tag{2.9}$$

Then

$$\widetilde{n}_+(q^{-1}R)D(q) = \widetilde{n}_-(q^{-1}S)\begin{pmatrix} D_q\gamma & \mathbf{0}\\ \mathbf{0} & I_n\end{pmatrix}. \tag{2.10}$$

(In the case $n = d$, the matrices "$\left(\begin{smallmatrix}\mathbf{0}\\ U\end{smallmatrix}\right)$" and "$(\mathbf{0}\ \ U^{-1})$" in (2.8) should be interpreted as "U" and "U^{-1}", respectively. The matrix $D_q \in \mathrm{SL}_d(\mathbb{R})$ in (2.9) should not be mixed up with the matrix $D(y)$ in $\mathrm{G} = \mathrm{SL}_{d+n}(\mathbb{R})$ defined in (1.2).)

Proof. Our task is to prove (2.10), or equivalently

$$n_-(q^{-1}S')\begin{pmatrix} D_q\gamma & \mathbf{0}\\ \mathbf{0} & I_n\end{pmatrix}\big(n_+(q^{-1}R')D(q)\big)^{-1} \in \Gamma, \tag{2.11}$$

where R' and S' are arbitrary lifts of R to $\mathrm{M}_{d\times n}(\mathbb{Z})$ and S to $\mathrm{M}_{n\times d}(\mathbb{Z})$, respectively. The matrix in (2.11) clearly has determinant one; hence it remains to prove that all its entries are integers. By a quick computation, the matrix is seen to equal

$$\begin{pmatrix} q^{\frac{n}{d}}D_q\gamma & -q^{\frac{n}{d}-1}D_q\gamma R'\\ q^{\frac{n}{d}-1}S'D_q\gamma & -q^{\frac{n}{d}-2}\,S'D_q\gamma R' + q^{-1}I_n\end{pmatrix}. \tag{2.12}$$

Here the top left block matrix is clearly in $\mathrm{M}_{d\times d}(\mathbb{Z})$, and using $\gamma R' \equiv \begin{pmatrix}\mathbf{0}\\ U\end{pmatrix} \bmod q$, the top right block matrix is seen to be in $\mathrm{M}_{d\times n}(\mathbb{Z})$; similarly the bottom left block matrix is in $\mathrm{M}_{n\times d}(\mathbb{Z})$. Finally, one verifies that $q^{\frac{n}{d}-1}S'D_q$ is in $\mathrm{M}_{n\times d}(\mathbb{Z})$ with its rightmost $n\times n$ submatrix being $\equiv U^{-1} \bmod q$; hence, since also $\gamma R' \equiv \begin{pmatrix}\mathbf{0}\\ U\end{pmatrix} \bmod q$, it follows that $q^{\frac{n}{d}-1}S'D_q\gamma R' \in I_n + q\cdot\mathrm{M}_n(\mathbb{Z})$. This implies that the bottom right block matrix in (2.12) is in $\mathrm{M}_n(\mathbb{Z})$, and the lemma is proved. □

3. Fourier analysis on the space $\Gamma\backslash\Gamma\mathrm{H}$

The material in the present section generalizes [Str15, Sec. 4]. Throughout the section we assume $1 \le n < d$.

We will parametrize the group H using the following diffeomorphism:

$$\mathrm{SL}_d(\mathbb{R}) \times \mathrm{M}_{n\times d}(\mathbb{R}) \xrightarrow{\sim} \mathrm{H}, \qquad (g, X) \mapsto \begin{pmatrix} I_d & \mathbf{0}\\ X & I_n\end{pmatrix}\begin{pmatrix} g & \mathbf{0}\\ \mathbf{0} & I_n\end{pmatrix} = \begin{pmatrix} g & \mathbf{0}\\ Xg & I_n\end{pmatrix}. \tag{3.1}$$

Note that then $\Gamma \cap \mathrm{H}$ corresponds to $\mathrm{SL}_d(\mathbb{Z}) \times \mathrm{M}_{n\times d}(\mathbb{Z})$, and the multiplication law in H is given by

$$(g, X)(g', X') = (gg', X + X'g^{-1}). \tag{3.2}$$

In particular, if F is a left $\Gamma \cap \mathrm{H}$ invariant function on H (or equivalently, a function on $\Gamma\backslash\Gamma\mathrm{H}$), then in terms of our parametrization we have $F(g, X + M) \equiv F(g, X)$ for all $M \in \mathrm{M}_{n\times d}(\mathbb{Z})$ [2], which means that for any fixed $g \in \mathrm{SL}_d(\mathbb{R})$, $X \mapsto F(g, X)$ is a function on the torus $\mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})$. We write $\widehat{F}(g; M)$ for the Fourier coefficients in the torus variable:

$$\widehat{F}(g; M) = \int_{\mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})} F(g, X)\, e^{-2\pi i \,\mathrm{tr}({}^{\mathrm{t}}MX)} dX, \tag{3.3}$$

where dX denotes the Lebesgue measure on $\mathrm{M}_{n\times d}(\mathbb{R}) \cong (\mathbb{R}^d)^n$. Thus for any $k > \frac{1}{2}nd$ and any $F \in \mathrm{C}^k(\Gamma\backslash\Gamma\mathrm{H})$, we have [Gra08, Theorem 3.2.16]

$$F(g, X) = \sum_{M\in \mathrm{M}_{n\times d}(\mathbb{Z})} \widehat{F}(g; M) e^{2\pi i \mathrm{tr}({}^{\mathrm{t}}MX)}, \tag{3.4}$$

with a uniform absolute convergence[3] over (g, X) in any compact subset of H.

Lemma 3.1. *Let $F \in \mathrm{C}(\Gamma\backslash\Gamma\mathrm{H})$. Then for any $\gamma \in \mathrm{SL}_d(\mathbb{Z})$, $g \in \mathrm{SL_d}(\mathbb{R})$ and $M \in \mathrm{M}_{n\times d}(\mathbb{Z})$,*

$$\widehat{F}(\gamma g; M) = \widehat{F}(g; M\,{}^{\mathrm{t}}\gamma^{-1}). \tag{3.5}$$

Proof. This follows from the formula (3.3) and $F(\gamma g, X) \equiv F(g, X\gamma)$, and the fact that the map $X \mapsto X\gamma$ is a diffeomorphism of the torus $\mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})$ onto itself preserving the Lebesgue measure. □

For $1 \le i \le n$ and $1 \le j \le d$, let $E_{i,j} \in \mathrm{M}_{n\times d}(\mathbb{R})$ denote the matrix with a 1 at the (i, j)th position and zeros elsewhere. We define the following differential operator:

$$(E_{i,j}F)(g, X) = \frac{\partial}{\partial t} F\big((g, X)(I_d, tE_{i,j})\big)\Big|_{t=0}. \tag{3.6}$$

Using $(g, X)(I_d, tE_{i,j}) = (g, X + tE_{i,j}g^{-1})$ and the chain rule, we get

$$(E_{i,j}F)(g, X) = \sum_{\ell=1}^{d} \overline{g}_{j,\ell} \frac{\partial F}{\partial x_{i,\ell}}(g, X), \tag{3.7}$$

where $X = (x_{i,j})_{1\le i\le n, 1\le j\le d}$ and $g^{-1} = (\overline{g}_{i,j})_{1\le i,j\le d}$.

Lemma 3.2. *Let $0 \le \kappa \le k$ with $k \in \mathbb{Z}$. Then for any $F \in \mathrm{C}^k_b(\Gamma\backslash\Gamma\mathrm{H})$, $g \in \mathrm{SL}_d(\mathbb{R})$ and $M \in \mathrm{M}_{n\times d}(\mathbb{Z})$,*

$$\big|\widehat{F}(g; M)\big| \ll_k \frac{S_{\infty,\kappa}(F)}{1 + \|M\,{}^{\mathrm{t}}g^{-1}\|_\infty^\kappa}. \tag{3.8}$$

Proof. By (3.3) and (3.7), we have for any $1 \le i \le n$ and $1 \le j \le d$,

$$\widehat{(E_{i,j}F)}(g; M) = \sum_{\ell=1}^{d} \overline{g}_{j,\ell} \int_{\mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})} \frac{\partial F}{\partial x_{i,\ell}}(g, X) e^{-2\pi i \,\mathrm{tr}({}^{\mathrm{t}}MX)} dX. \tag{3.9}$$

Hence by integration by parts,

$$\widehat{(E_{i,j}F)}(g; M) = 2\pi i \Big(\sum_{\ell=1}^{d} \overline{g}_{j,\ell} m_{i,\ell}\Big) \widehat{F}(g; M). \tag{3.10}$$

[2] We also have $F(\gamma g, X\gamma^{-1}) \equiv F(g, X)$ for all $\gamma \in \mathrm{SL}_d(\mathbb{Z})$.

[3] For any fixed ordering of $\mathrm{M}_{n\times d}(\mathbb{Z})$.

Repeated use of this formula gives

$$\widehat{(E^k_{i,j}F)}(g;M) = (2\pi i)^k \Big(\sum_{\ell=1}^{d} \overline{g}_{j,\ell} m_{i,\ell}\Big)^k \widehat{F}(g;M). \tag{3.11}$$

Recall the definition of the Sobolev norm $S_{\infty,k}$ on $\mathrm{C}^k_b(\Gamma\backslash\Gamma\mathrm{H})$; see (1.9). We may assume that the fixed basis for the Lie algebra of H which is used in this definition contains the vectors $\frac{d}{dt}\left(\begin{smallmatrix} I_d & \mathbf{0} \\ tE_{i,j} & I_n \end{smallmatrix}\right)\big|_{t=0}$ for all i,j. Then, using also (3.3), we have

$$\Big|\widehat{(E^k_{i,j}F)}(g;M)\Big| \le \big\|E^k_{i,j}F\big\|_\infty \le S_{\infty,k}(F).$$

Hence we conclude:

$$(2\pi)^k \Big|\sum_{\ell=1}^{d} \overline{g}_{j,\ell} m_{i,\ell}\Big|^k \big|\widehat{F}(g;M)\big| \le S_{\infty,k}(F).$$

Note also that, trivially,

$$\Big|\widehat{F}(g;M)\Big| \le S_{\infty,k}(F). \tag{3.12}$$

Hence

$$\Big(1 + (2\pi)^k \Big|\sum_{\ell=1}^{d} \overline{g}_{j,\ell} m_{i,\ell}\Big|^k\Big)\big|\widehat{F}(g;M)\big| \le 2S_{\infty,k}(F). \tag{3.13}$$

The above inequality holds for any $1 \le i \le n$ and $1 \le j \le d$. Note that $\sum_{\ell=1}^{d} \overline{g}_{j,\ell} m_{i,\ell}$ equals the entry of the matrix $M\,{}^{\mathrm{t}}g^{-1}$ at position i,j; hence the maximum of $\big|\sum_{\ell=1}^{d} \overline{g}_{j,\ell} m_{i,\ell}\big|$ over all i,j equals $\|M\,{}^{\mathrm{t}}g^{-1}\|_\infty$. Hence we obtain (3.8) with $\kappa = k$.

Finally, to extend to general κ, note that after possibly decreasing k we may assume that $k-1 < \kappa \le k$. If $\kappa = k$ then we are done; hence we may now assume $k-1 < \kappa < k$ (thus $\kappa > 0$ and $k \ge 1$). The bound proved above holds both for k and for $k' := k-1$; and combining these we obtain

$$\big|\widehat{F}(g;M)\big| \ll_k \left(\frac{S_{\infty,k'}(F)}{1 + \|M\,{}^{\mathrm{t}}g^{-1}\|^{k'}_\infty}\right)^{k-\kappa} \left(\frac{S_{\infty,k}(F)}{1 + \|M\,{}^{\mathrm{t}}g^{-1}\|^{k}_\infty}\right)^{\kappa-k'}.$$

This implies (3.8), by (1.11) (applied with k' in place of k) and since $(1+x^{k'})^{k-\kappa}(1+x^k)^{\kappa-k'} \ge 1+x^\kappa$ for all $x \ge 0$ (by Hölder's inequality). □

4. Effective equidistribution of Hecke points

In this section we collect the results about equidistribution of Hecke points which we will need in the proof of our main theorem. Our main reference will be [COU01]; the proofs in that paper make use of spectral theory of automorphic forms and the strong uniform bounds on matrix exponents of unitary representations obtained in [Oh02].

In this section we again assume $1 \le n < d$. Recall from Lemma 2.3 that we then have

$$D_q = q^{-\frac{n}{d}} \begin{pmatrix} I_{d-n} & \\ & qI_n \end{pmatrix} \in \mathrm{SL}_d(\mathbb{R}). \tag{4.1}$$

Lemma 4.1. *We have the disjoint coset decomposition*

$$\mathrm{SL}_d(\mathbb{Z})\, D_q\, \mathrm{SL}_d(\mathbb{Z}) = \bigsqcup_{\delta \in \Gamma^0(q)\backslash \mathrm{SL}_d(\mathbb{Z})} \mathrm{SL}_d(\mathbb{Z}) D_q \delta. \tag{4.2}$$

(Recall that $\Gamma^0(q)$ was defined in (2.2).)

Proof. Observe that

$$D_q^{-1}\mathrm{SL}_d(\mathbb{Z})D_q \cap \mathrm{SL}_d(\mathbb{Z}) = \Gamma^0(q). \tag{4.3}$$

Hence the group $\mathrm{SL}_d(\mathbb{Z})$ can be expressed as a disjoint union

$$\mathrm{SL}_d(\mathbb{Z}) = \bigcup_{\delta\in\Gamma^0(q)\backslash\mathrm{SL}_d(\mathbb{Z})} \big(D_q^{-1}\mathrm{SL}_d(\mathbb{Z})D_q \cap \mathrm{SL}_d(\mathbb{Z})\big)\delta. \tag{4.4}$$

Following [Shi94, Proposition 3.1], we get (4.2). □

We now follow the definition of Hecke operators given in [COU01]. For a complex valued function Φ on $\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})$, the Hecke operator for D_q is defined as

$$(T_{D_q}\Phi)(g) = \frac{1}{\#\big(\Gamma^0(q)\backslash\mathrm{SL}_d(\mathbb{Z})\big)} \sum_{\delta\in\Gamma^0(q)\backslash\mathrm{SL}_d(\mathbb{Z})} \Phi\left(D_q\delta g\right). \tag{4.5}$$

This makes sense since $[\mathrm{SL}_d(\mathbb{Z}) : \Gamma^0(q)] < \infty$.

The map T_{D_q} restricts to a bounded linear operator on $\mathrm{L}^2(\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R}))$. We will later also encounter the dual operator, $T^*_{D_q}$, i.e. the bounded linear operator on $\mathrm{L}^2(\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R}))$ which satisfies

$$\left\langle T_{D_q}\Phi_1, \Phi_2\right\rangle = \int_{\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})} \big[T_{D_q}\Phi_1\big](g)\overline{\Phi_2(g)}\, d\mu_0(g) = \left\langle \Phi_1, T^*_{D_q}\Phi_2\right\rangle \tag{4.6}$$

for all $\Phi_1, \Phi_2 \in \mathrm{L}^2(\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R}))$, where μ_0 is the $\mathrm{SL}_d(\mathbb{R})$-invariant probability measure on $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$. By mimicking the proof of [Shi94, Proposition 3.39] one verifies that $T^*_{D_q}$ is in fact the Hecke operator for D_q^{-1}. Using also the fact that the map $g \mapsto {}^tg^{-1}$ is an automorphism of $\mathrm{SL}_d(\mathbb{R})$ which maps D_q to D_q^{-1}, it follows from Lemma 4.1 that

$$\mathrm{SL}_d(\mathbb{Z})\, D_q^{-1}\, \mathrm{SL}_d(\mathbb{Z}) = \bigsqcup_{\delta\in\Gamma^0(q)\backslash\mathrm{SL}_d(\mathbb{Z})} \mathrm{SL}_d(\mathbb{Z})D_q^{-1}\,{}^t\delta^{-1},$$

and hence

$$\big(T^*_{D_q}\Phi\big)(g) = \frac{1}{\#\big(\Gamma^0(q)\backslash\mathrm{SL}_d(\mathbb{Z})\big)} \sum_{\delta\in\Gamma^0(q)\backslash\mathrm{SL}_d(\mathbb{Z})} \Phi\left(D_q^{-1}\,{}^t\delta^{-1}g\right) \tag{4.7}$$

for any $\Phi \in \mathrm{L}^2(\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R}))$. In fact, we will take (4.7) as a definition of $T^*_{D_q}\Phi$ for *any* function $\Phi : \mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R}) \to \mathbb{C}$.

Recall that we denote by θ the constant towards the Ramanujan conjecture for Maass wave forms on $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$.

Proposition 4.2. *Let $\kappa = \frac{d^2-1}{2}$, $\varepsilon > 0$, and $k = \lceil \kappa + \varepsilon\rceil$. Then for every $\Phi \in \mathrm{C}_b^k(\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R}))$, we have*

$$\left|(T_{D_q}\Phi)(I_d) - \int_{\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})} \Phi(g)\, d\mu_0(g)\right| \ll_\varepsilon S_{2,\kappa+\varepsilon}(\Phi) \begin{cases} q^{-\frac{1}{2}+\theta+\varepsilon} & \text{if } n = 1 \text{ and } d = 2 \\ q^{-\frac{\min\{n,d-n\}}{2}+\varepsilon} & \text{otherwise.} \end{cases} \tag{4.8}$$

Proof. It is a known result that for every $\Phi \in \mathrm{L}^2(\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R}))$,

$$\left\|T_{D_q}\Phi - \int_{\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})} \Phi(g)\, d\mu_0(g)\right\|_2 \ll_\varepsilon \|\Phi\|_2 \begin{cases} q^{-\frac{1}{2}+\theta+\varepsilon} & \text{if } n = 1 \text{ and } d = 2 \\ q^{-\frac{\min\{n,d-n\}}{2}+\varepsilon} & \text{otherwise.} \end{cases} \tag{4.9}$$

Indeed, if $d \geq 3$ then (4.9) follows by applying [COU01, Theorem 1.1 and p. 332 (Remark (3)) and Sec. 5.1] for the group $G = \mathrm{GL}_d$, and using the identification between $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$ with

$Z\,\mathrm{GL}_d(\mathbb{Z})\backslash\,\mathrm{GL}_d(\mathbb{R})$, where Z is the center of $\mathrm{GL}_d(\mathbb{R})$. In the case $d = 2$ one instead starts by noticing that (see [Shi94, Ch. 3.1–2; in particular Thm. 3.24]; alternatively follow the computation in [LM17, p. 6599(top)]):

$$T_{D_q}\Phi = \frac{1}{q\prod_{p|q}(1+p^{-1})}\sum_{a^2|q}\mu(a)\sigma_1\Big(\frac{q}{a^2}\Big)T_{q/a^2}\Phi, \tag{4.10}$$

where the sum runs over all positive integers a satisfying $a^2 \mid q$, and $\sigma_1(m) := \sum_{d|m} d$, and where T_m ($m \in \mathbb{Z}^+$) is the Hecke operator on $\mathrm{L}^2(\mathrm{SL}_2(\mathbb{Z})\backslash\,\mathrm{SL}_2(\mathbb{R}))$ defined by

$$(T_m\Phi)(g) = \frac{1}{\sigma_1(m)}\sum_{a|m}\sum_{b=0}^{\frac{m}{a}-1}\Phi\left(m^{-\frac{1}{2}}\begin{pmatrix}a & b\\ 0 & m/a\end{pmatrix}g\right) \qquad (g \in \mathrm{SL}_2(\mathbb{R})).$$

Next, by [GM03, Sec. 3] we have $\big\|T_m\Phi - \int_{\mathrm{SL}_2(\mathbb{Z})\backslash\,\mathrm{SL}_2(\mathbb{R})}\Phi\,d\mu_0\big\|_2 \ll_\varepsilon m^{-\frac{1}{2}+\theta+\varepsilon}\|\Phi\|_2$ for all $m \in \mathbb{Z}^+$. Using this bound in (4.10), the triangle inequality, and the fact that $\sum_{a^2|q}\mu(a)\sigma_1\big(\frac{q}{a^2}\big) = q\prod_{p|q}(1+p^{-1})$, we obtain (4.9) for $d = 2$. (The last step was also carried out in [LM17, pp. 6599–6600]).

Finally, after recalling the definition (1.11), the bound in (4.8) is deduced from (4.9) as in the proof of [SV05, Lemma 5]. □

5. Matrix Kloosterman sums

In this section we use the following notation:

$$e_q(x) := e^{2\pi i x/q} \qquad (q \in \mathbb{Z}^+,\ x \in \mathbb{R}).$$

For $n, q \in \mathbb{Z}^+$ and $A, B \in \mathrm{M}_n(\mathbb{Z}/q\mathbb{Z})$, we define

$$K_n(A,B;q) = \sum_{X\in\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})} e_q(\mathrm{tr}(AX + BX^{-1})). \tag{5.1}$$

5.1. **Prime moduli.** For p a prime number, we denote the field $\mathbb{Z}/p\mathbb{Z}$ by $\mathbb{F}_p$. In [ET21, Corollary 1.11], Erdélyi and Tóth have recently proved that for any prime number p and any $A, B \in \mathrm{M}_n(\mathbb{F}_p)$, not both $\mathbf{0}$,

$$\big|K_n(A,B;p)\big| \le 2p^{n^2-n+1}. \tag{5.2}$$

In fact, the main result of that paper [ET21] is that if both A and B belong to $\mathrm{GL}_n(\mathbb{F}_p)$, then the following much sharper bound holds:

$$\big|K_n(A,B;p)\big| \ll p^{(3n^2-\delta_n)/4}, \tag{5.3}$$

where $\delta_n = 0$ if n is even and $\delta_n = 1$ if n is odd [ET21, Theorem 1.8].

5.2. **General moduli.** This case is easily reduced to the case of prime power moduli, using the standard multiplicativity relation:

Lemma 5.1. *Let $q = \prod_{j=1}^r q_j$ where $q_1, \dots, q_r \in \mathbb{Z}^+$ are pairwise relatively prime, and for each j, let $c_j \in (\mathbb{Z}/q_j\mathbb{Z})^\times$ be a multiplicative inverse of $\prod_{i\neq j} q_i$ modulo q_j. Then for any $A, B \in \mathrm{M}_n(\mathbb{Z}/q\mathbb{Z})$,*

$$K_n(A,B;q) = \prod_{j=1}^r K_n(c_jA, c_jB; q_j) \tag{5.4}$$

Proof. For any integer a we have $a \equiv \sum_{j=1}^{r} (q/q_j) c_j a \bmod q$, and so $e_q(a) = \prod_{j=1}^{r} e_{q_j}(c_j a)$. In particular, for any $X \in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$,

$$e_q\big(\mathrm{tr}(AX + BX^{-1})\big) = \prod_{j=1}^{r} e_{q_j}\big(\mathrm{tr}(c_j AX + c_j BX^{-1})\big). \tag{5.5}$$

On the right-hand side, the jth factor depends only on X mod q_j, and when X runs through $\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$, the r-tuple $\langle X \bmod q_j \rangle_{j=1,\dots r}$ runs through the Cartesian product $\prod_{j=1}^{r} \mathrm{GL}_n(\mathbb{Z}/q_j\mathbb{Z})$. Hence when we sum (5.5) over $X \in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$, we obtain (5.4). □

Lemma 5.2. *Let $q \in \mathbb{Z}^+$ and $A, B \in \mathrm{M}_n(\mathbb{Z}/q\mathbb{Z})$, and let ℓ be a divisor of $\frac{q}{\prod_{p|q} p}$. Assume also that $\ell \mid B$. Then*

$$K_n(A, B; q) = \begin{cases} 0 & \text{if } \ell \nmid A \\ \ell^{n^2} K_n(\ell^{-1}A, \ell^{-1}B; \ell^{-1}q) & \text{if } \ell \mid A. \end{cases} \tag{5.6}$$

Proof. Because of the assumption on ℓ, we can fix a subset R of $\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$ containing exactly one representative for each congruence class in $\mathrm{GL}_n(\mathbb{Z}/\frac{q}{\ell}\mathbb{Z})$, and then the map $\langle Y, Z\rangle \mapsto \frac{q}{\ell} Y + Z$ is a bijection from $\mathrm{M}_n(\mathbb{Z}/\ell\mathbb{Z}) \times R$ onto $\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$. Using this parametrization in (5.1), writing $B = \ell B'$ with $B' \in \mathrm{M}_n(\mathbb{Z}/\frac{q}{\ell}\mathbb{Z})$ and noticing that $(\frac{q}{\ell} Y + Z)^{-1} \equiv Z^{-1} \bmod \frac{q}{\ell}$, we get

$$K_n(A, B; q) = \sum_{Y \in \mathrm{M}_n(\mathbb{Z}/\ell\mathbb{Z})} e_\ell\big(\mathrm{tr}(AY)\big) \sum_{Z \in R} e_q\big(\mathrm{tr}(AZ)\big) e_{q/\ell}\big(\mathrm{tr}(B'Z^{-1})\big).$$

Here the sum over Y equals ℓ^{n^2} if $\ell \mid A$, and otherwise vanishes. Hence we obtain (5.6). □

5.3. **Prime power moduli.** In the case of higher prime power moduli, we will prove a bound on $K_n(A, B; q)$ by direct and elementary computations; see Proposition 5.10 below for the final result. We remark that bounds of a similar nature, but more precise and in certain respects stronger, have independently been obtained in the recent paper [ETZ22] by Erdélyi, Tóth and Zábrádi. However we choose to include the proofs in this section in order to make our paper more self-contained and because we use a shortcut that leads to an upper bound which is sufficient for our needs. We further emphasize that, using the bounds from [ETZ22] instead, would not lead to an improvement of the exponents in our main result, Theorem 1.2.

For $q \in \mathbb{Z}^+$ and $A, B \in \mathrm{M}_n(\mathbb{Z}/q\mathbb{Z})$, we define

$$\mathcal{C}_q(A, B) = \{Y \in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z}) \,:\, AY \equiv Y^{-1}B \bmod q\}. \tag{5.7}$$

For any prime p and $C, D \in \mathrm{M}_n(\mathbb{F}_p)$, we also introduce the following matrix Gauss sum:

$$G_p(C, D) = \sum_{Z \in \mathrm{M}_n(\mathbb{F}_p)} e_p(\mathrm{tr}(CZ^2 + DZ)). \tag{5.8}$$

Lemma 5.3. *Let $q = p^\beta$ where p is a prime and $\beta \geq 2$, and set $\alpha = \lfloor \beta/2 \rfloor$. Let $A, B \in \mathrm{M}_n(\mathbb{Z}/q\mathbb{Z})$, and assume $A, B \not\equiv 0 \bmod p$. If β is even, then*

$$|K_n(A, B; q)| \leq p^{\alpha n^2} \# \mathcal{C}_{p^\alpha}(A, B). \tag{5.9}$$

If β is odd, then

$$|K_n(A, B; q)| \leq p^{\alpha n^2} \# \mathcal{C}_{p^\alpha}(A, B) \cdot \max\big\{|G_p(C, D)| \,:\, C, D \in \mathrm{M}_n(\mathbb{F}_p),\, C \neq 0\big\}. \tag{5.10}$$

Proof. Fix a subset R of $\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$ containing exactly one representative for each congruence class in $\mathrm{GL}_n(\mathbb{Z}/p^\alpha\mathbb{Z})$. Let us first assume that β is even; thus $\beta = 2\alpha$. Then the map $\langle Y, Z\rangle \mapsto Y(I + p^\alpha Z)$

is a bijection from $R \times \mathrm{M}_n(\mathbb{Z}/p^\alpha\mathbb{Z})$ onto $\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$. Using this parametrization in (5.1), and the fact that $(I + p^\alpha Z)^{-1} \equiv I - p^\alpha Z \bmod q$, we obtain

$$K_n(A,B;q) = \sum_{Y \in R} e_q(\mathrm{tr}(AY + BY^{-1})) \sum_{Z \in \mathrm{M}_n(\mathbb{Z}/p^\alpha\mathbb{Z})} e_{p^\alpha}\big(\mathrm{tr}\big((AY - Y^{-1}B)Z\big)\big).$$

Here the inner sum vanishes unless $AY - Y^{-1}B \equiv 0 \bmod p^\alpha$. Hence we obtain the bound in (5.9).

Next assume that β is odd, i.e. $\beta = 2\alpha+1$. Then we also fix a subset R' of $\mathrm{M}_n(\mathbb{Z}/p^{\alpha+1}\mathbb{Z})$ containing exactly one representative for each congruence class in $\mathrm{M}_n(\mathbb{F}_p)$. Then the map $\langle Y, Z_1, Z_2\rangle \mapsto Y(I + p^\alpha Z_1 + p^{\alpha+1}Z_2)$ is a bijection from $R \times R' \times \mathrm{M}_n(\mathbb{Z}/p^\alpha\mathbb{Z})$ onto $\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$. Using this in (5.1), together with the fact that $(I + p^\alpha Z_1 + p^{\alpha+1}Z_2)^{-1} \equiv I - p^\alpha Z_1 - p^{\alpha+1}Z_2 + p^{2\alpha}Z_1^2 \bmod q$, we obtain

$$\begin{aligned} K_n(A,B;q) = \sum_{Y \in R} e_q(\mathrm{tr}(AY + Y^{-1}B)) \sum_{Z_1 \in R'} e_{p^{\alpha+1}}\big(\mathrm{tr}\big((AY - Y^{-1}B)Z_1 + p^\alpha Y^{-1}BZ_1^2\big)\big) \\ \times \sum_{Z_2 \in \mathrm{M}_n(\mathbb{Z}/p^\alpha\mathbb{Z})} e_{p^\alpha}\big(\mathrm{tr}\big((AY - Y^{-1}B)Z_2\big)\big). \end{aligned}$$

Here the sum over Z_2 vanishes unless $AY - Y^{-1}B \equiv 0 \bmod p^\alpha$; hence we obtain

$$K_n(A,B;q) = p^{\alpha n^2} \sum_{\substack{Y \in R \\ AY - Y^{-1}B \equiv 0 \bmod p^\alpha}} e_q(\mathrm{tr}(AY + Y^{-1}B)) \cdot G_p\left(Y^{-1}B, \frac{AY - Y^{-1}B}{p^\alpha}\right),$$

and this leads to the bound in (5.10). □

In order to make the bound in Lemma 5.3 useful, we need to bound $\#\mathcal{C}_{p^\alpha}(A,B)$. We will first treat the case $\alpha = 1$, and for this we will need the following lemma.[4]

Lemma 5.4. *Let G be a finite group with the property that every irreducible linear representation of G over $\mathbb{C}$ is either realizable over $\mathbb{R}$ or has non-real character. Let $f : G \to \mathbb{Z}_{\geq 0}$ be the function that counts the number of square roots of each element in G, viz., $f(g) := \#\{x \in G \,:\, x^2 = g\}$. Then $f(g) \leq f(e)$ for all $g \in G$.*

Proof. Clearly f is a class function (i.e., invariant under conjugation), and hence $f = \sum_\chi \langle f, \chi\rangle \chi$ where the sum is taken over all irreducible characters of G [Ser77, Theorem 6]. Here

$$\langle f, \chi\rangle = \frac{1}{\#G}\sum_{g \in G} f(g)\chi(g^{-1}) = \frac{1}{\#G}\sum_{g \in G}\sum_{x \in G} \delta_{x^2=g}\,\chi(g^{-1}) = \frac{1}{\#G}\sum_{x \in G}\chi(x^{-2}) = \frac{1}{\#G}\sum_{x \in G}\chi(x^2).$$

This formula, together with the assumption of the lemma and [Ser77, Prop. 39], implies that $\langle f, \chi\rangle \in \{0, 1\}$ for all χ. Hence $f = \sum_{\chi \in S}\chi = \sum_{\chi \in S}\mathrm{Re}\,\chi$, where S is the set of those χ for which $\langle f, \chi\rangle = 1$. Hence for any $g \in G$, $f(g) = \sum_{\chi \in S}\mathrm{Re}\,\chi(g) \leq \sum_{\chi \in S}\chi(e) = f(e)$. □

Proposition 5.5. *For every prime p and every $A, B \in \mathrm{M}_n(\mathbb{F}_p)$, not both zero, we have*

$$\#\mathcal{C}_p(A,B) \ll p^{(n-1)^2+1}. \tag{5.11}$$

More precisely, if A is in $\mathrm{GL}_n(\mathbb{F}_p)$, then

$$\#\mathcal{C}_p(A,B) \ll p^{\frac{1}{2}(n^2-\delta_n)}, \tag{5.12}$$

where $\delta_n = \frac{1-(-1)^n}{2}$, while if $r = \mathrm{rank}\,A$ satisfies $1 \leq r \leq n-1$ then

$$\#\mathcal{C}_p(A,B) \ll p^{n^2-2r(n-r)}. \tag{5.13}$$

The implied constants in all three bounds are absolute.

[4] We learnt about this fact from MathOverflow, question 41784 ("Roots of permutations") [Bhb].

(A slightly more precise bound is given in [ETZ22, Theorem 1.6].)

Proof. It suffices to prove (5.12) and (5.13), since these imply (5.11). It is immediate from the definition, (5.7), that $\operatorname{rank} A \neq \operatorname{rank} B$ implies $\mathcal{C}_p(A,B) = \emptyset$; hence we may assume that $r = \operatorname{rank} A = \operatorname{rank} B$.

Let us first assume $r = n$, i.e. A and B both lie in $\mathrm{GL}_n(\mathbb{F}_p)$. Substituting $Z = AY$ in the definition of $\mathcal{C}_p(A,B)$, it follows that $\#\mathcal{C}_p(A,B)$ equals the number of elements $Z \in \mathrm{GL}_n(\mathbb{F}_p)$ with $Z^2 = AB$. Also the group $\mathrm{GL}_n(\mathbb{F}_p)$ is known to have the property that all of its linear representations are either realizable over $\mathbb{R}$ or have non-real character [Zel81, Ch. III, 12.6]. Using these facts in combination with Lemma 5.4, we conclude that

$$\#\mathcal{C}_p(A,B) \leq \#\{Z \in \mathrm{GL}_n(\mathbb{F}_p) \,:\, Z^2 = I\}. \tag{5.14}$$

The cardinality on the right-hand side of (5.14) is easy to calculate: if $Z \in \mathrm{GL}_n(\mathbb{F}_p)$ satisfies $Z^2 = I$, then all eigenvalues of Z must equal ± 1, and hence Z is conjugate over $\mathbb{F}_p$ to a matrix J in Jordan canonical form, say with Jordan blocks $J_1, \dots, J_k$ (in this order) where J_i is the $n_i \times n_i$ matrix

$$J_i = \begin{pmatrix} \varepsilon_i & 1 & & & \\ & \varepsilon_i & 1 & & 0 \\ & & \cdots & & \\ & & & \cdots & \\ & 0 & & & \varepsilon_i & 1 \\ & & & & & \varepsilon_i \end{pmatrix}$$

with $\varepsilon_i \in \{1, -1\}$. Let us first assume $p \neq 2$. Then $J^2 = I$ forces $n_i = 1$ for all i, and so we conclude that for every matrix Z belonging to the set on the right-hand side of (5.14), there is a unique $0 \leq a \leq n$ such that Z is conjugate over $\mathbb{F}_p$ to the diagonal matrix D_a having a 1's and $(n-a)$ -1's along the diagonal, in this order. Hence the right-hand side of (5.14) equals

$$\sum_{a=0}^{n} \#\{T D_a T^{-1} \,:\, T \in \mathrm{GL}_n(\mathbb{F}_p)\} = \sum_{a=0}^{n} \#(\mathrm{GL}_n(\mathbb{F}_p)/C(D_a)),$$

where $C(D_a)$ is the centralizer of D_a in $\mathrm{GL}_n(\mathbb{F}_p)$. But $C(D_a)$ consists of exactly the matrices in $\mathrm{GL}_n(\mathbb{F}_p)$ which are block diagonal with blocks of sizes $a, n-a$, and so

$$\#C(D_a) = \#\,\mathrm{GL}_a(\mathbb{F}_p) \#\,\mathrm{GL}_{n-a}(\mathbb{F}_p) = \prod_{j=0}^{a-1}(p^a - p^j) \prod_{j=0}^{n-a-1}(p^{n-a} - p^j).$$

If $a \in \{0, n\}$ this should of course be understood to say $\#C(D_a) = \#\,\mathrm{GL}_n(\mathbb{F}_p) = \prod_{j=0}^{n-1}(p^n - p^j)$. Hence the right-hand side of (5.14) equals

$$\sum_{a=0}^{n} \frac{\prod_{j=0}^{n}(p^n - p^j)}{\prod_{j=0}^{a-1}(p^a - p^j) \prod_{j=0}^{n-a-1}(p^{n-a} - p^j)}.$$

Noticing that $\prod_{j=0}^{a-1}(p^a - p^j) \asymp p^{a^2}$ uniformly over all primes p and all $a \geq 0$, the above expression is seen to be

$$\asymp \sum_{a=0}^{n} p^{n^2 - a^2 - (n-a)^2} \asymp p^{\frac{1}{2}(n^2 - \delta_n)},$$

with $\delta_n = \frac{1-(-1)^n}{2}$ and we have thus proved (5.12) in the case $r = n$, $p \neq 2$.

Next we assume $r = n$, $p = 2$. Then $J^2 = I$ forces $n_i \in \{1, 2\}$ for all i, and thus, since $-1 = 1$ in $\mathbb{F}_2$, every Jordan block appearing in J must equal $\overline{J} := \big(1\big)$ or $\overline{J}' := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Hence for every matrix Z belonging to the set in the right-hand side of (5.14), there is a unique $0 \le a \le \lfloor n/2 \rfloor$ such that Z is conjugate over $\mathbb{F}_2$ to the block diagonal matrix D'_a having a blocks $\overline{J}'$ and $n - 2a$ blocks $\overline{J}$ along the diagonal, in this order. It follows that the right-hand side of (5.14) equals

$$\sum_{0 \le a \le \lfloor n/2 \rfloor} \#\{T D'_a T^{-1} \,:\, T \in \mathrm{GL}_n(\mathbb{F}_2)\} = \sum_{0 \le a \le \lfloor n/2 \rfloor} \#(\mathrm{GL}_n(\mathbb{F}_2)/C(D'_a)). \tag{5.15}$$

Here we claim that

$$\#C(D'_a) = 2^{a(2n-3a)} \#\,\mathrm{GL}_a(\mathbb{F}_2) \#\,\mathrm{GL}_{n-2a}(\mathbb{F}_2). \tag{5.16}$$

To prove this, note that $X \in \mathrm{GL}_n(\mathbb{F}_2)$ commutes with D'_a if and only if X commutes with $D'_a - I$, and $D'_a - I$ can be conjugated, by a permutation matrix, into the matrix

$$U_a := \begin{pmatrix} \mathbf{0} & I_a & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{5.17}$$

with block sizes $a, a, n-2a$ in this order. Hence $\#C(D'_a) = \#C(U_a)$, and writing $X = \big(X_{ij}\big)_{i,j=1,2,3}$ with block sizes $a, a, n - 2a$, we find that X belongs to $C(U_a)$ if and only if $X_{11} = X_{22}$ and the three matrices X_{21}, X_{23}, X_{31} vanish. Furthermore, by considering the determinant, such a block matrix X is invertible if and only if both $X_{11} = X_{22}$ and X_{33} are invertible. Hence we obtain the formula in (5.16).

It follows from (5.15) and (5.16) that the right-hand side of (5.14) is

$$\asymp \sum_{0 \le a \le \lfloor n/2 \rfloor} 2^{n^2 - a(2n-3a) - a^2 - (n-2a)^2} = \sum_{0 \le a \le \lfloor n/2 \rfloor} 2^{\frac{1}{2}n^2 - 2(a - \frac{1}{2}n)^2} \asymp 2^{\frac{1}{2}(n^2 - \delta_n)}.$$

Hence (5.12) also holds in the case $r = n$, $p = 2$.

It remains to consider the case when $r = \operatorname{rank} A = \operatorname{rank} B$ satisfies $1 \le r \le n - 1$; we then wish to prove the bound (5.13). We may of course assume that $\mathcal{C}_p(A, B)$ is non-empty; thus let us fix some $Y_0 \in \mathcal{C}_p(A, B)$, and set $A_0 := AY_0 = Y_0^{-1}B \ne \mathbf{0}$. Then for any $Y \in \mathrm{GL}_n(\mathbb{F}_p)$, the condition $Y \in \mathcal{C}_p(A, B)$ is equivalent with $A_0 Y_0^{-1} Y = Y^{-1} Y_0 A_0$, viz., $Y_0^{-1} Y \in \mathcal{C}_p(A_0, A_0)$. Hence

$$\#\mathcal{C}_p(A, B) = \#\mathcal{C}_p(A_0, A_0).$$

Let $V = \{A_0 \boldsymbol{x} \,:\, \boldsymbol{x} \in \mathbb{F}_p^n\} = \{A\boldsymbol{x} \,:\, \boldsymbol{x} \in \mathbb{F}_p^n\}$; this is an r-dimensional subspace of $\mathbb{F}_p^n$. Let us note that

$$\forall Y \in \mathcal{C}_p(A_0, A_0): \qquad Y|_V \in \mathrm{GL}(V). \tag{5.18}$$

Indeed, $Y \in \mathcal{C}_p(A_0, A_0)$ implies $A_0 Y \boldsymbol{x} = Y^{-1} A_0 \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{F}_p^n$; hence $V = Y^{-1}(V)$, and so $Y|_V \in \mathrm{GL}(V)$.

Let us fix $\boldsymbol{b}_1, \dots, \boldsymbol{b}_{n-r}$ to be a basis of some complementary subspace of V in $\mathbb{F}_p^n$. Now let $Y_1 \in \mathrm{GL}(V)$ be given. Then for any $Y \in \mathcal{C}_p(A_0, A_0)$ with $Y|_V = Y_1$ (if such a Y exists at all), we have $A_0 Y(\boldsymbol{b}_j) = Y^{-1} A_0(\boldsymbol{b}_j) = Y_1^{-1} A_0(\boldsymbol{b}_j)$ for every j. This means that $Y(\boldsymbol{b}_j)$ belongs to the preimage of $Y_1^{-1} A_0(\boldsymbol{b}_j)$ under A_0. Since A_0 is a linear map on $\mathbb{F}_p^n$ of rank r, and the preimage of $Y_1^{-1} A_0(\boldsymbol{b}_j)$ under A_0 is not empty, the preimage is an affine linear subspace of $\mathbb{F}_p^n$ of dimension $n - r$. It follows that the tuple $Y(\boldsymbol{b}_1), \dots, Y(\boldsymbol{b}_{n-r})$ can be chosen in at most $p^{(n-r)^2}$ ways. Now since Y is determined by linearity from Y_1 and the elements $Y(\boldsymbol{b}_1), \dots, Y(\boldsymbol{b}_{n-r})$, we conclude that:

$$\forall Y_1 \in \mathrm{GL}(V): \qquad \#\{Y \in \mathcal{C}_p(A_0, A_0) \,:\, Y|_V = Y_1\} \le p^{(n-r)^2}. \tag{5.19}$$

It follows from (5.18), (5.19) that

$$\#\mathcal{C}_p(A,B) = \#\mathcal{C}_p(A_0,A_0) \le p^{(n-r)^2} \#\,\mathrm{GL}(V) \le p^{(n-r)^2+r^2},$$

i.e. (5.13) holds. □

We now turn to the problem of bounding $\#\mathcal{C}_{p^\alpha}(A,B)$ for $\alpha \ge 2$. We will start by proving a bound on the following quantity, which turns out to be relevant for bounding both $\#\mathcal{C}_{p^\alpha}(A,B)$ and the Gauss sum $G_p(C,D)$. For any $C \in \mathrm{M}_n(\mathbb{F}_p)$, we set

$$d(C) = \dim_{\mathbb{F}_p} \mathcal{A}(C), \qquad \text{where} \quad \mathcal{A}(C) = \{Z \in \mathrm{M}_n(\mathbb{F}_p) \,:\, CZ + ZC = 0\}. \tag{5.20}$$

Note that $\mathcal{A}(C)$ is a vector subspace of $\mathrm{M}_n(\mathbb{F}_p) \cong \mathbb{F}_p^{n^2}$, and $\#\mathcal{A}(C) = p^{d(C)}$.

Lemma 5.6. *For any $C \in \mathrm{M}_n(\mathbb{F}_p) \setminus \{0\}$, if either $p > 2$ or $C \ne I$ then*

$$d(C) \le (n-1)^2 + 1. \tag{5.21}$$

Proof. Let us fix an algebraic closure $\overline{\mathbb{F}}_p$ of $\mathbb{F}_p$. For any $C' \in \mathrm{M}_n(\overline{\mathbb{F}}_p)$ we define

$$d(C') = \dim_{\overline{\mathbb{F}}_p}\{Z \in \mathrm{M}_n(\overline{\mathbb{F}}_p) \,:\, C'Z + ZC' = 0\}.$$

Note that this formula is consistent with (5.20) if $C' \in \mathrm{M}_n(\mathbb{F}_p)$. Note also that $d(C) = d(TCT^{-1})$ for any $T \in \mathrm{GL}_n(\overline{\mathbb{F}}_p)$. We may now choose $T \in \mathrm{GL}_n(\overline{\mathbb{F}}_p)$ so that $C' := TCT^{-1}$ is in Jordan canonical form. Thus let us assume that C' has Jordan blocks $J_1, \dots, J_k$ (in this order) where J_i is the $n_i \times n_i$ matrix

$$J_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & 0 \\ & & \cdots & & \\ & & & \cdots & \\ & 0 & & & \lambda_i \;\; 1 \\ & & & & \lambda_i \end{pmatrix} \qquad (\lambda_i \in \overline{\mathbb{F}}_p). \tag{5.22}$$

Writing Z in block decomposed form as $Z = (Z_{i,j})$ with $Z_{i,j} \in \mathrm{M}_{n_i,n_j}(\overline{\mathbb{F}}_p)$ for $i,j \in \{1,\dots,k\}$, one notes that $C'Z + ZC' = 0$ holds if and only if $J_iZ_{i,j} + Z_{i,j}J_j = 0$ for all pairs i,j. Hence

$$d(C) = d(C') = \sum_{i=1}^{k}\sum_{j=1}^{k} \dim_{\overline{\mathbb{F}}_p}\{Z \in \mathrm{M}_{n_i,n_j}(\overline{\mathbb{F}}_p) \,:\, J_iZ + ZJ_j = 0\}. \tag{5.23}$$

We now claim that

$$\dim_{\overline{\mathbb{F}}_p}\{Z \in \mathrm{M}_{n_i,n_j}(\overline{\mathbb{F}}_p) \,:\, J_iZ + ZJ_j = 0\} = \begin{cases} \min(n_i,n_j) & \text{if } \lambda_i = -\lambda_j \\ 0 & \text{if } \lambda_i \ne -\lambda_j. \end{cases} \tag{5.24}$$

To prove this, note that writing $Z = (z_{a,b}) \in \mathrm{M}_{n_i,n_j}(\overline{\mathbb{F}}_p)$, the relation $J_iZ + ZJ_j = 0$ holds if and only if

$$(\lambda_i + \lambda_j)z_{a,b} + z_{a+1,b} + z_{a,b-1} = 0 \qquad (\forall a \in \{1,\dots,n_i\},\, b \in \{1,\dots,n_j\}), \tag{5.25}$$

where we understand that $z_{n_i+1,b} = 0$ for all b and $z_{a,0} = 0$ for all a. Let us first assume $\lambda_i = -\lambda_j$, so that the equation (5.25) simply reads $z_{a+1,b} + z_{a,b-1} = 0$. This implies that the matrix entries are alternating along each diagonal, viz., for any fixed $a' \in \{1,\dots,n_i\}$ and $b' \in \{1,\dots,n_j\}$ with either $a' = 1$ or $b' = 1$, we have $z_{a',b'} = -z_{a'+1,b'+1} = z_{a'+2,b'+2} = \cdots = (-1)^\ell z_{a'+\ell,b'+\ell}$ where $\ell = \min(n_i - a', n_j - b')$. But we also have $z_{a',1} = 0$ for all $a' \ge 2$ (by (5.25) applied with $a = a' - 1$ and $b = 1$) and $z_{n_i,b'} = 0$ for all $b' < n_j$ (by (5.25) applied with $a = n_i$ and $b = b' + 1$). Hence all the diagonals which start at $z_{a',1}$ with $1 < a' \le n_i$ vanish completely; and if $n_i < n_j$ then also the diagonals which start at $z_{1,b'}$ with $1 \le b' \le n_j - n_i$ vanish completely. Conversely one verifies that

any matrix having vanishing diagonals as just described, and the remaining diagonals alternating, satisfies *all* the relations in (5.25). Furthermore, there are exactly $\min(n_i, n_j)$ diagonals which are not forced to vanish. Hence (5.24) holds in the case $\lambda_i = -\lambda_j$.

Next we assume $\lambda_i \neq -\lambda_j$. Then, applying (5.25) for $b = 1$ and $a = n_i, n_i - 1, \ldots, 1$ (in this order) we get $z_{a,1} = 0$ for all a. Next, applying (5.25) for $b = 2$ and $a = n_i, n_i - 1, \ldots, 1$ gives $z_{a,2} = 0$ for all a. This may be repeated successively for $b = 3, \ldots, n_j$, finally giving $Z = 0$. Hence (5.24) holds also in the case $\lambda_i \neq -\lambda_j$.

Using (5.24) in (5.23), we obtain

$$d(C) = \sum_{i=1}^{k} \sum_{\substack{j=1 \\ (\lambda_j = -\lambda_i)}}^{k} \min(n_i, n_j). \tag{5.26}$$

This implies

$$d(C) \leq \sum_{i=1}^{k} \sum_{j=1}^{k} n_i n_j \Big(\tfrac{1}{2} + \tfrac{1}{2} \delta_{n_i = n_j = 1} \Big), \tag{5.27}$$

and using $\sum_i n_i = n$, the right-hand side of (5.27) is seen to equal $\frac{1}{2}n^2 + \frac{1}{2}m^2$, where $m := \#\{i : n_i = 1\}$. If $n_i \geq 2$ for some i then $m \leq n - 2$, and so

$$d(C) \leq \tfrac{1}{2}n^2 + \tfrac{1}{2}(n-2)^2 = (n-1)^2 + 1,$$

i.e. (5.21) holds.

Hence from now on we may assume that $n_i = 1$ for all i, viz., C is diagonalizable. Then (5.26) gives

$$d(C) = \sum_{\lambda \in S} d_\lambda d_{-\lambda},$$

where $S = \{\lambda_1, \ldots, \lambda_k\}$ is the set of eigenvalues of C and $d_\lambda = \sum_{i:\, \lambda_i = \lambda} n_i$ is the dimension of the eigenspace for λ. First assume $p \neq 2$. Then $\lambda \neq -\lambda$ for all $\lambda \in \overline{\mathbb{F}}_p \setminus \{0\}$ and thus we can choose a subset $S' \subset S$ such that $S \setminus \{0\} \subset S' \cup (-S')$ and $S' \cap (-S') = \emptyset$. Now

$$\begin{aligned} \sum_{\lambda \in S} d_\lambda d_{-\lambda} = d_0^2 + 2 \sum_{\lambda \in S'} d_\lambda d_{-\lambda} \leq d_0^2 + \frac{1}{2} \sum_{\lambda \in S'} (d_\lambda + d_{-\lambda})^2 \leq d_0^2 + \frac{1}{2} \Big(\sum_{\lambda \in S'} (d_\lambda + d_{-\lambda}) \Big)^2 \\ = d_0^2 + \tfrac{1}{2}(n - d_0)^2, \end{aligned}$$

and since $C \neq 0$ we have $0 \leq d_0 \leq n - 1$, so that

$$d_0^2 + \tfrac{1}{2}(n - d_0)^2 \leq (n-1)^2 + 1$$

(with equality if and only if $n = 2$ and $d_0 = 0$). Hence (5.21) holds.

Finally assume $p = 2$. Then $d(C) = \sum_{\lambda \in S} d_\lambda^2$. If S is a singleton set, say $S = \{\lambda\}$, then $C' = \lambda I$, and thus also $C = \lambda I$. This forces $\lambda \in \mathbb{F}_2$, and so $C \in \{0, I\}$. Hence if $C \notin \{0, I\}$ then $\#S \geq 2$, and choosing some element $\lambda' \in S$ we get

$$d(C) = d_{\lambda'}^2 + \sum_{\lambda \in S \setminus \{d_{\lambda'}\}} d_\lambda^2 \leq d_{\lambda'}^2 + (n - d_{\lambda'})^2 \leq (n-1)^2 + 1,$$

since $1 \leq d_{\lambda'} \leq n - 1$. Thus (5.21) holds. □

For $p = 2$ we will also need the following bound of similar type.

Lemma 5.7. *For any $C \in \mathrm{M}_n(\mathbb{F}_2)$,*

$$\dim_{\mathbb{F}_2}\{Z \in \mathrm{M}_n(\mathbb{F}_2) \,:\, Z + CZ + ZC = 0\} \le \frac{1}{2}n^2. \tag{5.28}$$

Proof. The proof of Lemma 5.6 carries over with some modifications. Introducing the Jordan decomposition of C exactly as in that proof, the analogue of (5.23) now says that the dimension in the left hand side of (5.28) equals

$$\sum_{i=1}^{k}\sum_{j=1}^{k} \dim_{\overline{\mathbb{F}}_2}\{Z \in \mathrm{M}_{n_i,n_j}(\overline{\mathbb{F}}_2) \,:\, Z + J_iZ + ZJ_j = 0\}. \tag{5.29}$$

Here we have, just as in (5.24),

$$\dim_{\overline{\mathbb{F}}_2}\{Z \in \mathrm{M}_{n_i,n_j}(\overline{\mathbb{F}}_2) \,:\, Z + J_iZ + ZJ_j = 0\} = \begin{cases} \min(n_i,n_j) & \text{if } 1+\lambda_i+\lambda_j = 0 \\ 0 & \text{if } 1+\lambda_i+\lambda_j \neq 0. \end{cases}$$

(Indeed, $Z + J_iZ + ZJ_j = 0$ is equivalent with (5.25) but with $\lambda_i + \lambda_j$ replaced by $1 + \lambda_i + \lambda_j$.) Hence the dimension in the left hand side of (5.28) is

$$\sum_{i=1}^{k}\sum_{\substack{j=1\\(\lambda_j=-\lambda_i-1)}}^{k} \min(n_i,n_j) \le \sum_{i=1}^{k}\sum_{\substack{j=1\\(\lambda_j=-\lambda_i-1)}}^{k} n_in_j = \sum_{\lambda\in S} d_\lambda d_{-1-\lambda},$$

where $S = \{\lambda_1,\dots,\lambda_k\}$ is the set of eigenvalues of C and $d_\lambda = \sum_{i:\,\lambda_i=\lambda} n_i$ is the generalized eigenspace dimension for λ. Now since $\lambda \neq -1-\lambda$ for all $\lambda \in \overline{\mathbb{F}}_2$, we can choose a subset $S' \subset S$ such that $S \subset S' \cup (-1-S')$ and $S' \cap (-1-S') = \emptyset$, and we then get

$$\sum_{\lambda\in S} d_\lambda d_{-1-\lambda} = 2\sum_{\lambda\in S'} d_\lambda d_{-1-\lambda} \le \frac{1}{2}\sum_{\lambda\in S'}(d_\lambda + d_{-1-\lambda})^2 \le \frac{1}{2}\Big(\sum_{\lambda\in S'}(d_\lambda + d_{-1-\lambda})\Big)^2 = \frac{1}{2}n^2.$$

□

Lemma 5.8. *Let q be a prime power. Then for any $A, B \in \mathrm{M}_n(\mathbb{Z}/q\mathbb{Z})$ with $\gcd(q, A, B) = 1$,*

$$\#\mathcal{C}_q(A,B) \ll q^{(n-1)^2+1},$$

where the implied constant is absolute.

Proof. Let us write $q = p^\alpha$ with p a prime and $\alpha \ge 1$. Let $A, B \in \mathrm{M}_n(\mathbb{Z}/q\mathbb{Z})$ and assume $\gcd(q, A, B) = 1$, viz., either $A \not\equiv 0$ or $B \not\equiv 0 \bmod p$. Note that $\#\mathcal{C}_q(A,B) = \#\mathcal{C}_q(B,A)$, since for any $Y \in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$, $AY \equiv Y^{-1}B \bmod q$ holds if and only if $(Y^{-1})^{-1}A \equiv BY^{-1} \bmod q$. Hence without loss of generality we may assume that $A \not\equiv 0 \bmod p$. In view of (5.11) in Proposition 5.5, it suffices to prove that for every $Y \in \mathcal{C}_p(A,B)$ there exist at most $p^{(\alpha-1)((n-1)^2+1)}$ lifts of Y to $\mathcal{C}_{p^\alpha}(A,B)$, that is, at most $p^{(\alpha-1)((n-1)^2+1)}$ matrices $Y' \in \mathcal{C}_{p^\alpha}(A,B)$ satisfying $[Y' \bmod p] = Y$. By induction over α, it suffices to prove that for any $\alpha \ge 1$, any $A, B \in \mathrm{M}_n(\mathbb{Z}/p^{\alpha+1}\mathbb{Z})$ with $A \not\equiv 0 \bmod p$, and any $Y \in \mathcal{C}_{p^\alpha}(A,B)$, there exist at most $p^{(n-1)^2+1}$ matrices $Y' \in \mathcal{C}_{p^{\alpha+1}}(A,B)$ satisfying $[Y' \bmod p^\alpha] = Y$. This holds trivially if there is *no* such matrix Y'; hence we may assume that there exists a matrix $Y_0' \in \mathcal{C}_{p^{\alpha+1}}(A,B)$ with $[Y_0' \bmod p^\alpha] = Y$. Now the set of matrices $Y' \in \mathrm{M}_n(\mathbb{Z}/p^{\alpha+1}\mathbb{Z})$ with $Y' \equiv Y_0' \bmod p^\alpha$ can be parametrized as $Y' = Y_0'(I + p^\alpha Z)$ with Z running through $\mathrm{M}_n(\mathbb{F}_p)$, and we then compute that

$$AY' - {Y'}^{-1}B \equiv p^\alpha(AY_0'Z + ZY_0'^{-1}B) \equiv p^\alpha(AY_0'Z + ZAY_0') \bmod p^{\alpha+1}.$$

Hence Y' lies in $\mathcal{C}_{p^{\alpha+1}}(A,B)$ if and only if $Z \in \mathcal{A}(AY_0')$. Therefore, the number of admissible lifts Y' equals $\#\mathcal{A}(AY_0') = p^{d(AY_0')}$. Note that $AY_0' \not\equiv 0 \bmod p$; hence if $p > 2$ then by Lemma 5.6 we have $d(AY_0') \le (n-1)^2 + 1$, and the proof is complete.

From now on we assume $p = 2$. In this case we decompose $\mathcal{C}_{2^\alpha}(A,B)$ as the disjoint union of the two sets

$$\mathcal{C}_{2^\alpha}^{(0)}(A,B) := \{Y \in \mathcal{C}_{2^\alpha}(A,B) \,:\, AY \not\equiv I \bmod 2\}$$

and

$$\mathcal{C}_{2^\alpha}^{(1)}(A,B) := \{Y \in \mathcal{C}_{2^\alpha}(A,B) \,:\, AY \equiv I \bmod 2\}.$$

For $\mathcal{C}_{2^\alpha}^{(0)}(A,B)$ the argument in the previous paragraph applies (since we get $AY_0' \not\equiv I \bmod 2$ as required in Lemma 5.6), and we thus obtain

$$\#\mathcal{C}_{2^\alpha}^{(0)}(A,B) \ll 2^{\alpha((n-1)^2+1)} = q^{(n-1)^2+1}.$$

We next consider $\mathcal{C}_{2^\alpha}^{(1)}(A,B)$. Note that from now on we may assume that both $A, B \in \mathrm{GL}_n(\mathbb{Z}/2^\alpha\mathbb{Z})$ since otherwise $\mathcal{C}_{2^\alpha}^{(1)}(A,B) = \emptyset$. Substituting $X = AY$ we have

$$\#\mathcal{C}_{2^\alpha}^{(1)}(A,B) = \#\{X \in \mathrm{GL}_n(\mathbb{Z}/2^\alpha\mathbb{Z}) \,:\, X^2 \equiv AB \bmod 2^\alpha \text{ and } X \equiv I \bmod 2\}.$$

Substituting next $X = I + 2U$ with $U \in \mathrm{M}_n(\mathbb{Z}/2^{\alpha-1}\mathbb{Z})$, we see that $\#\mathcal{C}_2^{(1)}(A,B) \le 1$ (with equality if and only if $AB \equiv I \bmod 2$), while for $\alpha \ge 2$ we obtain $\mathcal{C}_{2^\alpha}^{(1)}(A,B) = \emptyset$ if $AB \not\equiv I \bmod 4$, while in the case $AB \equiv I \bmod 4$ we get, after choosing $B' \in \mathrm{M}_n(\mathbb{Z}/2^{\alpha-2}\mathbb{Z})$ such that $AB \equiv I + 4B' \bmod 2^\alpha$:

$$\begin{aligned}\#\mathcal{C}_{2^\alpha}^{(1)}(A,B) &= \#\{U \in \mathrm{M}_n(\mathbb{Z}/2^{\alpha-1}\mathbb{Z}) \,:\, U + U^2 \equiv B' \bmod 2^{\alpha-2}\} \\ &= 2^{n^2}\#\{U \in \mathrm{M}_n(\mathbb{Z}/2^{\alpha-2}\mathbb{Z}) \,:\, U + U^2 \equiv B' \bmod 2^{\alpha-2}\}. \end{aligned} \tag{5.30}$$

In particular if $\alpha = 2$ then $\#\mathcal{C}_{2^\alpha}^{(1)}(A,B) = 2^{n^2} \le 2^{2((n-1)^2+1)}$, as desired. To handle the case $\alpha \ge 3$ we will prove that for any $\beta \ge 1$ and any $B' \in \mathrm{M}_n(\mathbb{Z}/2^\beta\mathbb{Z})$,

$$\#\{U \in \mathrm{M}_n(\mathbb{Z}/2^\beta\mathbb{Z}) \,:\, U + U^2 \equiv B' \bmod 2^\beta\} \le 2^{\frac{1}{2}(\beta+1)n^2}. \tag{5.31}$$

This bound is trivial for $\beta = 1$, and to prove it for $\beta \ge 2$ it suffices, by the same inductive lifting argument as in the first paragraph, to prove that for any $\beta \ge 2$ and any $U, B' \in \mathrm{M}_n(\mathbb{Z}/2^\beta\mathbb{Z})$ with $U + U^2 \equiv B' \bmod 2^\beta$, the number of $Z \in \mathrm{M}_n(\mathbb{Z}/2\mathbb{Z})$ satisfying $(U + 2^{\beta-1}Z) + (U + 2^{\beta-1}Z)^2 \equiv B' \bmod 2^\beta$ is at most $2^{\frac{1}{2}n^2}$. But the last equation is seen to be equivalent to

$$Z + UZ + ZU \equiv 0 \bmod 2,$$

and hence the claim follows from Lemma 5.7. Using (5.30) and (5.31), it follows that for all $\alpha \ge 3$,

$$\#\mathcal{C}_{2^\alpha}^{(1)}(A,B) \le 2^{\frac{1}{2}(\alpha+1)n^2} \le 8 \cdot 2^{\alpha((n-1)^2+1)} = 8q^{(n-1)^2+1},$$

where the last inequality holds since $3 + \alpha((n-1)^2+1) - \frac{1}{2}(\alpha+1)n^2 = \frac{\alpha-1}{2}(n - \frac{2\alpha}{\alpha-1})^2 + \frac{\alpha-3}{\alpha-1} \ge 0$. This completes the proof of the lemma. □

Lemma 5.9. *For any prime p and any $C, D \in \mathrm{M}_n(\mathbb{F}_p)$,*

$$|G_p(C,D)| \le p^{\frac{1}{2}(n^2+d(C))}.$$

Proof. It follows from the definition, (5.8), that

$$|G_p(C,D)|^2 = \sum_{Z,Y \in \mathrm{M}_n(\mathbb{F}_p)} e_p\big(\mathrm{tr}(C(Z^2 - Y^2) + D(Z - Y))\big).$$

Substituting $Z = X + Y$, this becomes

$$\sum_{X \in \mathrm{M}_n(\mathbb{F}_p)} \sum_{Y \in \mathrm{M}_n(\mathbb{F}_p)} e_p\big(\mathrm{tr}((XC + CX)Y) + \mathrm{tr}(CX^2 + DX)\big),$$

and here the inner sum vanishes unless $XC + CX = 0$. Hence

$$|G_p(C,D)|^2 \le \left| p^{n^2} \sum_{X \in \mathcal{A}(C)} e_p(CX^2 + DX) \right| \le p^{n^2} \#\mathcal{A}(C) = p^{n^2 + d(C)}.$$

□

Proposition 5.10. *Let q be a prime power. Then for any $A, B \in \mathrm{M}_n(\mathbb{Z}/q\mathbb{Z})$ with $\gcd(q, A, B) = 1$,*

$$|K_n(A,B;q)| \ll_n q^{n^2 - n + 1}. \tag{5.32}$$

Proof. Let us write $q = p^\beta$ with p a prime and $\beta \ge 1$. If $\beta = 1$ then (5.32) follows from (5.2); hence from now on we assume $\beta \ge 2$. It follows from $\gcd(q, A, B) = 1$ that $A \not\equiv 0$ or $B \not\equiv 0 \bmod p$. If exactly one of A or B is divisible by p then $K_n(A, B; q) = 0$ by Lemma 5.2; hence from now on we may assume that both $A, B \not\equiv 0 \bmod p$. We will now use the bound in Lemma 5.3. Thus set $\alpha = \lfloor \beta/2 \rfloor \ge 1$. By Lemma 5.8, $\#\mathcal{C}_{p^\alpha}(A,B) \ll p^{\alpha((n-1)^2+1)}$, and so

$$p^{\alpha n^2} \#\mathcal{C}_{p^\alpha}(A,B) \ll p^{2\alpha(n^2 - n + 1)}.$$

Furthermore, if $p > 2$, then by Lemma 5.6 we have $d(C) \le (n-1)^2 + 1$ for all $C \in \mathrm{M}_n(\mathbb{F}_p) \setminus \{0\}$, and hence by Lemma 5.9,

$$\max\big\{|G_p(C,D)| \,:\, C, D \in \mathrm{M}_n(\mathbb{F}_p),\ C \ne 0\big\} \le p^{\frac{1}{2}(n^2 + (n-1)^2 + 1)} = p^{n^2 - n + 1}.$$

On the other hand for $p = 2$ we have the trivial bound

$$|G_p(C,D)| \le 2^{n^2} \ll_n 2^{n^2 - n + 1}.$$

Using these bounds in Lemma 5.3, we get

$$|K_n(A,B;q)| \ll_n p^{\beta(n^2 - n + 1)} = q^{n^2 - n + 1}.$$

□

By combining Proposition 5.10 with Lemma 5.1 and Lemma 5.2, we now obtain a bound valid for general moduli.

Theorem 5.11. *Let $\varepsilon > 0$, $q \ge 2$ and $A, B \in \mathrm{M}_n(\mathbb{Z}/q\mathbb{Z})$. If $\gcd(q, A, B) = 1$ then*

$$|K_n(A,B;q)| \ll_{n,\varepsilon} q^{n^2 - n + 1 + \varepsilon}. \tag{5.33}$$

If $\ell = \gcd(q, A)$ then

$$\big|K_n(A,B;q)\big| \ll_{n,\varepsilon} q^{n^2} (q/\ell)^{-n+1+\varepsilon}. \tag{5.34}$$

(See also [ETZ22, Theorem 1.8] for somewhat stronger and more precise bounds.)

Proof. If $\gcd(q, A, B) = 1$ then it follows from Lemma 5.1 and Proposition 5.10 that

$$|K_n(A,B;q)| \ll_{n,\varepsilon} q^{n^2 - n + 1 + \varepsilon}.$$

Next we assume instead $\ell = \gcd(q, A)$. Write $q = \prod_{i=1}^s p_i^{\alpha_i}$ and $\ell = \prod_{i=1}^s p_i^{\gamma_i}$ (thus $0 \le \gamma_i \le \alpha_i$ and $0 < \alpha_i$ for all i). Picking $c_i \in (\mathbb{Z}/p^{\alpha_i}\mathbb{Z})^\times$ as in Lemma 5.1 we have

$$K_n(A,B;q) = \prod_{i=1}^s K_n(c_i A, c_i B; p^{\alpha_i}).$$

For each i, if $\gamma_i < \alpha_i$ then $p^{\alpha_i} \nmid A$, and so by Lemma 5.2 we have

$$K_n(c_iA, c_iB; p^{\alpha_i}) = p^{\gamma_i n^2} K_n(p^{-\gamma_i}c_iA, p^{-\gamma_i}c_iB; p^{\alpha_i-\gamma_i})$$

if $p^{\gamma_i} \mid B$, and otherwise $K_n(c_iA, c_iB; p^{\alpha_i}) = 0$. Hence if $\gamma_i < \alpha_i$ then by Proposition 5.10,

$$|K_n(c_iA, c_iB; p^{\alpha_i})| \le Cp^{\gamma_i n^2} p^{(\alpha_i-\gamma_i)(n^2-n+1)} = Cp^{\alpha_i n^2} p^{(\alpha_i-\gamma_i)(-n+1)},$$

where $C = C(n) \ge 1$ is the implied constant in (5.32). In the remaining case, when $\gamma_i = \alpha_i$, we use the *trivial* bound $|K_n(c_iA, c_iB; p^{\alpha_i})| \le p^{\alpha_i n^2}$. Multiplying over all i, we obtain:

$$|K_n(A, B; q)| \le q^{n^2} \prod_{\substack{i=1 \\ (\gamma_i<\alpha_i)}}^{s} \left(Cp^{(\alpha_i-\gamma_i)(-n+1)}\right) \ll_{n,\varepsilon} q^{n^2}(q/\ell)^{-n+1+\varepsilon}.$$

Hence we have proved (5.34). □

Finally we deal with the Ramanujan sum case.

Proposition 5.12. *Let p be a prime and $m \in \mathbb{Z}^+$. For $A \in \mathrm{M}_n(\mathbb{Z}/p^m\mathbb{Z})$, when $p^{m-1} \nmid A$,*

$$K_n(\mathbf{0}, A; p^m) = 0. \tag{5.35}$$

Assume that $p^{m-1} \mid A$ and let $r \in \{0, 1, \dots, n\}$ be the rank of the matrix $p^{-(m-1)}A$ in $\mathrm{M}_n(\mathbb{F}_p)$. Then

$$K_n(\mathbf{0}, A; p^m) = p^{(m-1)n^2}(-1)^r p^{-\frac{r(r+1)}{2}+rn} \prod_{i=0}^{n-r-1} (p^{n-r} - p^i). \tag{5.36}$$

Proof. Applying Lemma 5.2 with $\ell = p^{m-1}$, the first claim, (5.35), follows immediately, and the second claim, (5.36), is reduced to the case $m = 1$. Now (5.36) follows from [ET21, Thm. 1.9], since $\left|\mathrm{GL}_{n-r}(\mathbb{F}_p)\right| = \prod_{i=0}^{n-r-1}(p^{n-r} - p^i)$. □

Corollary 5.13. *Let q be a positive integer. We have*

$$|K_n(\mathbf{0}, A; q)| \le \begin{cases} q^{n^2}\left(\frac{q}{\gcd(q,A)}\right)^{-n} & \text{when } \prod_{p\mid q} p^{\mathrm{ord}_p(q)-1} \mid A, \\ 0 & \text{otherwise.} \end{cases} \tag{5.37}$$

Proof. Let us write $m_p = \mathrm{ord}_p(q)$. If $\prod_{p\mid q} p^{m_p-1} \nmid A$ then $K_n(\mathbf{0}, A; q) = 0$, by the first part of Proposition 5.12. From now on we assume $\prod_{p\mid q} p^{m_p-1} \mid A$. Letting r_p be the rank of the matrix $p^{-(m_p-1)}A$ in $\mathrm{M}_n(\mathbb{F}_p)$, we have by the second part of Proposition 5.12:

$$|K_n(\mathbf{0}, A; q)| \le \prod_{p\mid q} p^{(m_p-1)n^2} p^{-\frac{r_p(r_p+1)}{2}+r_pn} p^{(n-r_p)^2} \tag{5.38}$$

$$= \prod_{p\mid q} p^{m_pn^2+\frac{r_p(r_p-1)}{2}-r_pn} = q^{n^2} \prod_{p\mid q} p^{r_p\left(\frac{r_p-1}{2}-n\right)}. \tag{5.39}$$

One verifies that $r_p\left(\frac{r_p-1}{2} - n\right) \le -n$ whenever $r_p \ge 1$; furthermore the product of all primes $p \mid q$ satisfying $r_p \ge 1$ equals $\frac{q}{\gcd(q,A)}$; hence the inequality in (5.37) follows. □

6. Geometry of numbers

Let us fix integers $1 \le n < d$. For any real numbers $\kappa > nd$ and $a, b > 0$, and any $g \in \mathrm{SL}_d(\mathbb{R})$, we define

$$\Phi_{a,b}^{(\kappa)}(g) = \sum_{\substack{X \in \mathrm{M}_{n\times d}(\mathbb{Z}) \\ X \neq \mathbf{0}}} \frac{1}{a + b\|Xg\|_\infty^\kappa}, \tag{6.1}$$

Note that the condition $\kappa > nd$ ensures that the series on the right-hand side converges (see also Lemma 6.2 below). Furthermore, $\Phi_{a,b}^{(\kappa)}$ is (left) $\mathrm{SL}_d(\mathbb{Z})$-invariant; indeed, for any $\gamma \in \mathrm{SL}_d(\mathbb{Z})$ we have

$$\Phi_{a,b}^{(\kappa)}(\gamma g) = \sum_{\substack{X \in \mathrm{M}_{n\times d}(\mathbb{Z}) \\ X \neq \mathbf{0}}} \frac{1}{a + b\|X\gamma g\|_\infty^\kappa} = \Phi_{a,b}^{(\kappa)}(g). \tag{6.2}$$

Our goal in the present section is to prove a bound on the integral of $\Phi_{a,b}^{(\kappa)}$ over $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$; see Proposition 6.4 below. This bound will play an important role in our proof of the main theorem in Section 7. We start by proving, in Lemma 6.2 below, a pointwise bound on $\Phi_{a,b}^{(\kappa)}$.

For a lattice L in $\mathbb{R}^d$ we write $\lambda_i = \lambda_i(L)$ $(i = 1, \dots, d)$ for its successive minima with respect to the unit ball, i.e.,

$$\lambda_i(L) := \min\big\{\lambda \in \mathbb{R}_{\ge 0} \,:\, L \text{ contains } i \text{ linearly independent vectors of length} \le \lambda\big\}. \tag{6.3}$$

Thus $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_d$. Let $\mathcal{B}_R^d \subset \mathbb{R}^d$ be the ball of radius R with centre at the origin in $\mathbb{R}^d$.

Lemma 6.1. *For every lattice L in $\mathbb{R}^d$ and every $R > 0$,*

$$\#(L \cap \mathcal{B}_R^d) \asymp_d \prod_{i=1}^{d} \Big(1 + \frac{R}{\lambda_i(L)}\Big).$$

Proof. See [GS91, Proposition 6].[5] □

Lemma 6.2. *For any $\kappa > nd$, $a, b > 0$ and $g \in \mathrm{SL}_d(\mathbb{R})$, writing $\lambda_i := \lambda_i(\mathbb{Z}^d g)$ for $i = 1, \dots, d$, we have*

$$\Phi_{a,b}^{(\kappa)}(g) \ll \begin{cases} b^{-1}\lambda_1^{-\kappa} & \text{if } \lambda_1 \ge (a/b)^{1/\kappa} \\ a^{-1}\displaystyle\prod_{i=1}^{d}\Big(1 + \frac{(a/b)^{1/\kappa}}{\lambda_i}\Big)^n & \text{if } \lambda_1 \le (a/b)^{1/\kappa}, \end{cases} \tag{6.4}$$

where the implied constant depends only on n, d and κ.

Proof. As we will see, the lemma follows from the definition (6.1) and Lemma 6.1 by a simple computation using dyadic decomposition. Note that for any $r > 0$, both sides in (6.4) are scaled by a factor r^{-1} when replacing $\langle a, b\rangle$ by $\langle ra, rb\rangle$; hence we may without loss of generality assume $b = 1$. Now set $c := \lambda_1/\sqrt{d}$ and

$$N_m := \#\big\{X \in \mathrm{M}_{n\times d}(\mathbb{Z}) \,:\, c\,2^{m-1} \le \|Xg\|_\infty < c\,2^m\big\} \qquad (m \in \mathbb{Z}).$$

[5] As noted in the erratum of [GS91], in the statement of [GS91, Proposition 6], "$\lambda_1 \cdots \lambda_k/M(K)$" should read "$\lambda_1 \cdots \lambda_k M(K)$", and in the last line of the proof, "$\sim V(K_0)$" should be corrected to "$\sim V(K_0)^{-1}$".

Then $N_m = 0$ for all $m \le 0$, since every non-zero vector $\boldsymbol{v} \in \mathbb{Z}^d g$ satisfies $\|\boldsymbol{v}\|_\infty \ge \|\boldsymbol{v}\|/\sqrt{d} \ge c$, for $\|\boldsymbol{v}\| = \sqrt{{}^t\boldsymbol{v}\boldsymbol{v}}$. Hence

$$\Phi_{a,1}^{(\kappa)}(g) = \sum_{\substack{X \in \mathrm{M}_{d\times n}(\mathbb{Z}) \\ X \neq \mathbf{0}}} \frac{1}{a + \|Xg\|_\infty^\kappa} \ll_\kappa \sum_{m=1}^\infty \frac{N_m}{a + (c\,2^m)^\kappa} \ll_{\kappa,d} \sum_{m=1}^\infty \frac{N_m}{a + (\lambda_1\,2^m)^\kappa}.$$

Here

$$N_m \le \Big(\#(\mathbb{Z}^d g \cap \mathcal{B}^d_{c2^m})\Big)^n \ll_d \widetilde{N}_m := \prod_{i=1}^d \Big(1 + \frac{\lambda_1 2^m}{\lambda_i}\Big)^n, \tag{6.5}$$

by Lemma 6.1 (and since $c \ll_d \lambda_1$), and also

$$\frac{1}{a + (\lambda_1\,2^m)^\kappa} < A_m := \min\big(a^{-1}, (\lambda_1\,2^m)^{-\kappa}\big), \tag{6.6}$$

and so

$$\Phi_{a,1}^{(\kappa)}(g) \ll_{\kappa,d} \sum_{m=1}^\infty \widetilde{N}_m A_m. \tag{6.7}$$

Here we note that the sequence $\widetilde{N}_1, \widetilde{N}_2, \dots$ is increasing and satisfies $(\frac{5}{3})^n \widetilde{N}_m \le \widetilde{N}_{m+1} \le 2^{nd}\widetilde{N}_m$ for all $m \ge 1$ (where the first inequality comes from behavior of the factor corresponding to $i = 1$ in (6.5)). Letting m_0 be the unique real number satisfying $\lambda_1 2^{m_0} = a^{1/\kappa}$, it follows that the sequence $\widetilde{N}_m A_m$ is geometrically increasing with a ratio $\ge (\frac{5}{3})^n$ for $m \le m_0$ and geometrically decreasing with the ratio $2^{nd-\kappa}$ for $m \ge m_0$. Hence if $m_0 \le 0$ (viz., if $\lambda_1 \ge a^{1/\kappa}$) then $\Phi_{a,1}^{(\kappa)}(g) \ll_{\kappa,d,n} \widetilde{N}_1 A_1 \ll_{d,n} \lambda_1^{-\kappa}$, while if $m_0 \ge 0$ then

$$\begin{aligned}\Phi_{a,1}^{(\kappa)}(g) &\ll_{\kappa,d,n} \widetilde{N}_{\lfloor m_0\rfloor} A_{\lfloor m_0\rfloor} + \widetilde{N}_{\lceil m_0\rceil} A_{\lceil m_0\rceil} \le a^{-1}\big(\widetilde{N}_{\lfloor m_0\rfloor} + \widetilde{N}_{\lceil m_0\rceil}\big) \ll_{n,d} a^{-1}\widetilde{N}_{m_0} \\ &= a^{-1}\prod_{i=1}^d \Big(1 + \frac{a^{1/\kappa}}{\lambda_i}\Big)^n. \end{aligned} \tag{6.8}$$

□

We will also make use of Rogers' formula, [Rog55, Theorem 4], which can be stated as follows (see [SS22, Theorem 1.5 and Sec. 2]). Recall from Section 4 that μ_0 denotes the invariant probability measure on $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$.

Theorem 6.3. *For any $1 \le n < d$, and for any Borel measurable function $\rho : \mathrm{M}_{n\times d}(\mathbb{R}) \to \mathbb{R}_{\ge 0}$ we have*

$$\int_{\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})} \sum_{\substack{X \in \mathrm{M}_{n\times d}(\mathbb{Z}) \\ X \neq \mathbf{0}}} \rho(Xg)\,d\mu_0(g) = \sum_{m=1}^n \sum_{B \in A_{n,m}} \int_{\mathrm{M}_{m\times d}(\mathbb{R})} \rho(BX)\,dX, \tag{6.9}$$

where for each $m \in \{1, \dots, n\}$, $A_{n,m}$ is a subset of $\mathrm{M}_{n\times m}(\mathbb{Z})$ such that the map $B \mapsto B\,\mathbb{R}^m$ is a bijection from $A_{n,m}$ onto the family of rational m-dimensional subspaces of $\mathbb{R}^n$, [6] *and $(B\,\mathbb{R}^m) \cap \mathbb{Z}^n = B\,\mathbb{Z}^m$ for each $B \in A_{n,m}$; furthermore, dX is the standard md-dimensional Lebesgue measure on $\mathrm{M}_{m\times d}(\mathbb{R})$.*

[6] Recall that a linear subspace $V \subset \mathbb{R}^n$ is said to be rational if $V = \mathrm{Span}_{\mathbb{R}}(V \cap \mathbb{Z}^n)$.

In the above theorem, note that (6.9) should be understood as an identity between extended real numbers, i.e. either both sides of the equality sign are finite and equal, or else both sides are $+\infty$.

Finally we are now ready to prove the main result of the present section.

Proposition 6.4. *For any $\kappa > nd$ and $a, b > 0$ we have*

$$\int_{\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})} \Phi_{a,b}^{(\kappa)}(g)\, d\mu_0(g) \ll a^{-1}\Big(\frac{a}{b}\Big)^{\frac{d}{\kappa}}\Big(1+\frac{a}{b}\Big)^{(n-1)\frac{d}{\kappa}}, \tag{6.10}$$

where the implied constant depends only on n, d and κ.

In particular the proposition implies that the integral on the left-hand side of (6.10) is finite.

Proof. We apply Theorem 6.3 with the following choice of $\rho = \rho_{a,b} : \mathrm{M}_{n\times d}(\mathbb{R}) \to \mathbb{R}_{\geq 0}$:

$$\rho_{a,b}(Y) := \frac{1}{a + b\|Y\|_\infty^\kappa}.$$

With this choice, (6.9) says that

$$\int_{\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})} \Phi_{a,b}^{(\kappa)}(g)\, d\mu_0(g) = \sum_{m=1}^{n} J_{a,b}(m), \tag{6.11}$$

where

$$J_{a,b}(m) := \sum_{B\in A_{n,m}} \int_{\mathrm{M}_{m\times d}(\mathbb{R})} \rho_{a,b}(BX)\, dX.$$

Note that for any $c > 0$, by substituting $X = c^{1/\kappa} X_{\mathrm{new}}$ we have

$$J_{a,b}(m) = c^{md/\kappa} J_{a,bc}(m). \tag{6.12}$$

Furthermore, we have the trivial scaling property

$$J_{a,b}(m) = c J_{ac,bc}(m). \tag{6.13}$$

Combining these we get:

$$J_{a,b}(m) = \Big(\frac{a}{b}\Big)^{md/\kappa} J_{a,a}(m) = a^{-1}\Big(\frac{a}{b}\Big)^{md/\kappa} J_{1,1}(m). \tag{6.14}$$

Let us now also note that, by Lemma 6.1 and Lemma 6.2,

$$\Phi_{1,1}^{(\kappa)}(g) \ll_{n,d,\kappa} \#(\mathbb{Z}^d g \cap \mathcal{B}_1^d)^n, \qquad \forall g \in \mathrm{SL}_d(\mathbb{R}).$$

Furthermore, by Schmidt [Sch58, Theorem 2] we have

$$\int_{\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})} \#(\mathbb{Z}^d g \cap \mathcal{B}_1^d)^n\, d\mu_0(g) < \infty.$$

Hence $\int_{\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})} \Phi_{1,1}^{(\kappa)}(g)\, d\mu_0(g) < \infty$, and thus $J_{1,1}(m) < \infty$ for each $m \in \{1, \dots, n\}$. Hence

$$\int_{\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})} \Phi_{a,b}^{(\kappa)}(g)\, d\mu_0(g) = \sum_{m=1}^{n} J_{a,b}(m) = \sum_{m=1}^{n} a^{-1}\Big(\frac{a}{b}\Big)^{md/\kappa} J_{1,1}(m) \ll_{n,d,\kappa} \sum_{m=1}^{n} a^{-1}\Big(\frac{a}{b}\Big)^{m\frac{d}{\kappa}}$$
$$\ll_n a^{-1} \max\Big(\Big(\frac{a}{b}\Big)^{\frac{d}{\kappa}}, \Big(\frac{a}{b}\Big)^{n\frac{d}{\kappa}}\Big),$$

viz., the bound in (6.10) holds. □

7. Proof of the main theorem

In this section we give the proof of Theorem 1.2. The proof is split into the two cases $n = d$ and $n < d$. The first of these is treated in Section 7.1: as we will see, the proof in this case is a fairly easy consequence of the bounds on the matrix Kloosterman sums proved in Section 5. The proof in the case $n < d$ is carried out in Sections 7.2–7.5; the proof depends crucially on the bounds in Section 5 in this case as well, but we additionally need to invoke Hecke equidistribution and methods from geometry of numbers.

We stress that throughout the present section, the implied constant in any "$\ll$" may depend on d (thus may also depend on n), without this being explicitly indicated in the notation.

7.1. **The case $n = d$.** In this case we have $\kappa = 2n^2$ and $k = 2n^2+1$ in the statement of Theorem 1.2. Furthermore, $\mathcal{R}_q = \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$ and $\mathcal{B}_q = \{I_n\}$ (see Section 2); hence by Lemma 2.3, for any $f \in \mathrm{C}_b^k(\mathrm{M}_n(\mathbb{R}/\mathbb{Z}) \times \Gamma\backslash\Gamma\mathrm{H})$ and $q \in \mathbb{Z}^+$, we have

$$\mathcal{A}_q(f) = \frac{1}{\#\mathcal{R}_q} \sum_{R \in \mathcal{R}_q} f\left(q^{-1}R, \widetilde{n}_+(q^{-1}R)D(q)\right) = \frac{1}{\#\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})} \sum_{R \in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})} f\left(q^{-1}R, \widetilde{n}_-(q^{-1}R^{-1})\right). \tag{7.1}$$

Note also that $\mathrm{H} \cong \mathrm{M}_n(\mathbb{R})$ and $\Gamma\backslash\Gamma\mathrm{H} \cong \mathrm{M}_n(\mathbb{R}/\mathbb{Z})$ in the present case; hence we have the following Fourier expansion, for $X_1, X_2 \in \mathrm{M}_n(\mathbb{R}/\mathbb{Z})$:

$$f(X_1, \widetilde{n}_-(X_2)) = \sum_{N,M \in \mathrm{M}_n(\mathbb{Z})} \widehat{f}(N, M) e^{2\pi i \mathrm{tr}({}^t\!NX_1)} e^{2\pi i \mathrm{tr}({}^t\!MX_2)}, \tag{7.2}$$

where

$$\widehat{f}(N, M) = \int_{\mathrm{M}_n(\mathbb{R}/\mathbb{Z})} \int_{\mathrm{M}_n(\mathbb{R}/\mathbb{Z})} f(T_1, \widetilde{n}_-(T_2)) e^{-2\pi i \mathrm{tr}({}^t\!NT_1)} e^{-2\pi i \mathrm{tr}({}^t\!MT_2)}\, dT_1\, dT_2. \tag{7.3}$$

The sum in (7.2) is absolutely convergent, uniformly with respect to X_1, X_2, since $f \in \mathrm{C}_b^k$ with $k = 2n^2 + 1 > n^2$ [Gra08, Theorem 3.2.16].

By applying integration by parts in a similar way as in Lemma 3.2, we have, for any $0 \le \lambda \le k$ and $N, M \in \mathrm{M}_{n\times n}(\mathbb{Z})$,

$$\left|\widehat{f}(N, M)\right| \ll \min\left(\frac{S_{\infty,\lambda}(f)}{1 + \|N\|_\infty^\lambda}, \frac{S_{\infty,\lambda}(f)}{1 + \|M\|_\infty^\lambda}\right) \ll \frac{S_{\infty,\lambda}(f)}{1 + \|N\|_\infty^\lambda + \|M\|_\infty^\lambda}. \tag{7.4}$$

Substituting (7.2) into (7.1), and then using the definition of the matrix Kloosterman sum $K_n(A, B; q)$ in (5.1) and the basic identity $K_n({}^t\!A, {}^t\!B; q) = K_n(A, B; q)$, we obtain

$$\mathcal{A}_q(f) = \widehat{f}(\mathbf{0}, \mathbf{0}) + \frac{1}{\#\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})} \sum_{\substack{N,M \in \mathrm{M}_n(\mathbb{Z}) \\ N \neq \mathbf{0} \text{ or } M \neq \mathbf{0}}} \widehat{f}(N, M) \sum_{R \in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})} e^{2\pi i \frac{\mathrm{tr}({}^t\!NR) + \mathrm{tr}({}^t\!MR^{-1})}{q}} = \int_{\mathrm{M}_n(\mathbb{R}/\mathbb{Z})} \int_{\mathrm{M}_n(\mathbb{R}/\mathbb{Z})} f\left(T_1, \widetilde{n}_-(T_2)\right) dT_1\, dT_2 + E(q), \tag{7.5}$$

where

$$E(q) := \frac{1}{\#\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})} \sum_{\substack{N,M \in \mathrm{M}_n(\mathbb{Z}) \\ N \neq \mathbf{0} \text{ or } M \neq \mathbf{0}}} \widehat{f}(N, M) K_n(N, M; q).$$

Here, in order to bound the Kloosterman sum for $M = \mathbf{0}$ we apply (5.37) in Corollary 5.13, while for $M \neq \mathbf{0}$ we use (5.34) in Theorem 5.11 if $n \geq 2$, and the classical Weil bound if $n = 1$. Using also

$$\#\,\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z}) = q^{n^2}\prod_{p\mid q}\prod_{j=1}^{n}(1-p^{-j}) > q^{n^2}\prod_{j=2}^{n}\zeta(j)^{-1}\prod_{p\mid q}(1-p^{-1}) \gg q^{n^2}(\log q)^{-1} \tag{7.6}$$

(where the first equality holds by Lemma 2.1, and the last relation holds by Mertens' third theorem), and (7.4), we obtain:

$$\big|E(q)\big| \ll_\varepsilon \frac{S_{\infty,\lambda}(f)}{q^{n^2-\varepsilon}}\Bigg(\sum_{\substack{N\in \mathrm{M}_n(\mathbb{Z})\\ N\neq\mathbf{0}}}\frac{q^{n^2-n}\gcd(q,N)^n}{\|N\|_\infty^\lambda} + \sum_{\substack{N,M\in \mathrm{M}_n(\mathbb{Z})\\ M\neq\mathbf{0}}}\frac{q^{n^2-\vartheta+\varepsilon}\gcd(q,M)^\vartheta}{1+\|N\|_\infty^\lambda+\|M\|_\infty^\lambda}\Bigg), \tag{7.7}$$

where ϑ is as in the statement of Theorem 1.2, viz., $\vartheta = n-1$ if $n > 1$ and $\vartheta = \frac{1}{2}$ if $n = 1$.

In the first sum, we substitute $\ell = \gcd(q, N)$ and $N = \ell\, N_{\mathrm{new}}$; this gives

$$\sum_{\substack{N\in \mathrm{M}_n(\mathbb{Z})\\ N\neq\mathbf{0}}}\frac{\gcd(q,N)^n}{\|N\|_\infty^\lambda} = \sum_{\ell\mid q}\ell^{n-\lambda}\sum_{\substack{N\in \mathrm{M}_n(\mathbb{Z})\\ N\neq\mathbf{0}}}\|N\|_\infty^{-\lambda}.$$

Using here the fact that

$$|\{N\in \mathrm{M}_n(\mathbb{Z}) : \|N\|_\infty = m\}| \leq 2n^2(2m+1)^{n^2-1} \qquad (\forall m\in\mathbb{Z}_{\geq 0}), \tag{7.8}$$

we obtain

$$\sum_{\substack{N\in \mathrm{M}_n(\mathbb{Z})\\ N\neq\mathbf{0}}}\frac{\gcd(q,N)^n}{\|N\|_\infty^\lambda} \ll \sum_{\ell\mid q}\ell^{n-\lambda}\sum_{m=1}^{\infty} m^{n^2-1-\lambda}. \tag{7.9}$$

The last sum converges if and only if $\lambda > n^2$, and when this holds the total expression is bounded above by a constant which only depends on λ.

Similarly, regarding the second sum in (7.7) we have:

$$\begin{aligned}
\sum_{\substack{N,M\in \mathrm{M}_n(\mathbb{Z})\\ M\neq\mathbf{0}}}\frac{\gcd(q,M)^\vartheta}{1+\|N\|_\infty^\lambda+\|M\|_\infty^\lambda} &\ll \sum_{\ell\mid q}\ell^\vartheta\sum_{\substack{N,M\in \mathrm{M}_n(\mathbb{Z})\\ M\neq\mathbf{0}}}\frac{1}{1+\|N\|_\infty^\lambda+\ell^\lambda\|M\|_\infty^\lambda}\\
&\ll \sum_{\ell\mid q}\ell^\vartheta\sum_{u=0}^{\infty}\sum_{m=1}^{\infty}\frac{(u+1)^{n^2-1}m^{n^2-1}}{1+u^\lambda+\ell^\lambda m^\lambda}\\
&\ll \sum_{\ell\mid q}\ell^\vartheta\Bigg(\sum_{m=1}^{\infty}\ell^{-\lambda}m^{n^2-1-\lambda}\sum_{u=0}^{\ell m}(u+1)^{n^2-1} + \sum_{u=\ell+1}^{\infty}u^{n^2-1-\lambda}\sum_{1\leq m<u/\ell}m^{n^2-1}\Bigg)\\
&\ll \sum_{\ell\mid q}\ell^\vartheta\Bigg(\sum_{m=1}^{\infty}\ell^{-\lambda}m^{n^2-1-\lambda}\,(\ell m)^{n^2} + \sum_{u=\ell+1}^{\infty}u^{n^2-1-\lambda}\,(u/\ell)^{n^2}\Bigg).
\end{aligned}$$

Both the sums in the last expression are convergent if and only if $\lambda > 2n^2$, and *if* $\lambda > 2n^2$ then we obtain

$$\sum_{\substack{N,M\in \mathrm{M}_n(\mathbb{Z})\\ M\neq\mathbf{0}}}\frac{\gcd(q,M)^\vartheta}{1+\|N\|_\infty^\lambda+\|M\|_\infty^\lambda} \ll_\lambda \sum_{\ell\mid q}\ell^{\vartheta+n^2-\lambda} \ll_\varepsilon q^\varepsilon. \tag{7.10}$$

(If $n \geq 2$ then we even have $\sum_{\ell|q} \ell^{\vartheta+n^2-\lambda} \ll 1$.)

In conclusion, using (7.9) and (7.10) in (7.7), it follows that for any $\lambda > 2n^2$,

$$\big|E(q)\big| \ll_{\lambda,\varepsilon} S_{\infty,\lambda}(f)\left(q^{-n+\varepsilon} + q^{-\vartheta+3\varepsilon}\right) \ll S_{\infty,\lambda}(f)\, q^{-\vartheta+3\varepsilon}.$$

Using this in (7.5), setting $\varepsilon = \frac{1}{3}\varepsilon_{\text{new}}$ and then choosing $\lambda = 2n + \varepsilon_{\text{new}}$, we obtain the relation (1.12), i.e. we have proved Theorem 1.2 in the case $n = d$. □

7.2. **The case $n < d$.** To start the proof in this case, let $\kappa, \vartheta, \kappa', \vartheta', k, \varepsilon, f$ and q be given as in the statement of Theorem 1.2. By Lemma 2.2 and Lemma 2.3,

$$\begin{aligned}(7.11)\quad \mathcal{A}_q(f) &= \frac{1}{\#\mathcal{R}_q} \sum_{\gamma\in\mathcal{B}_q} \sum_{U\in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})} f\left(q^{-1}\gamma^{-1}\begin{pmatrix}\mathbf{0}\\ U\end{pmatrix}, \widetilde{n}_-\left(q^{-1}\begin{pmatrix}\mathbf{0} & U^{-1}\end{pmatrix}\right)\begin{pmatrix} D_q\gamma & \\ & I_n\end{pmatrix}\right)\\ &= \frac{1}{\#\mathcal{R}_q} \sum_{\gamma\in\mathcal{B}_q} \sum_{U\in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})} \sum_{N\in \mathrm{M}_{d\times n}(\mathbb{Z})} \widehat{f}\left(N; \widetilde{n}_-\left(q^{-1}\begin{pmatrix}\mathbf{0} & U^{-1}\end{pmatrix}\right)\begin{pmatrix} D_q\gamma & \\ & I_n\end{pmatrix}\right) e^{2\pi i \frac{\mathrm{tr}\left({}^{\mathrm{t}}N\gamma^{-1}\begin{pmatrix}\mathbf{0}\\ U\end{pmatrix}\right)}{q}},\end{aligned}$$

where, for $N \in \mathrm{M}_{d\times n}(\mathbb{Z})$ and $h \in \mathrm{H}$,

$$(7.12)\qquad \widehat{f}(N;h) = \int_{\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})} f(T,h) e^{-2\pi i\, \mathrm{tr}({}^{\mathrm{t}}NT)}\, dT.$$

Recall from the statement of Theorem 1.2 that $f \in \mathrm{C}_b^k(\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z}) \times \Gamma\backslash\Gamma\mathrm{H})$. By applying integration by parts in a similar way as in Lemma 3.2, we have, for any $0 \leq \lambda \leq k$, $N \in \mathrm{M}_{d\times n}(\mathbb{Z})$ and $h \in \mathrm{H}$,

$$(7.13)\qquad \big|\widehat{f}(N;h)\big| \ll \frac{S_{\infty,\lambda}(f)}{1 + \|N\|_\infty^\lambda}.$$

Recall our parametrization of H in (3.1); note that this can be expressed as $h = n_-(X)\begin{pmatrix} g & \\ & I_n\end{pmatrix}$. In line with this we set, for $g \in \mathrm{SL}_d(\mathbb{R})$, $X \in \mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})$ and $N \in \mathrm{M}_{d\times n}(\mathbb{Z})$:

$$(7.14)\qquad F(g,X;N) = \widehat{f}\left(N, \widetilde{n}_-(X)\begin{pmatrix} g & \\ & I_n\end{pmatrix}\right).$$

By (3.4),

$$(7.15)\qquad F(g,X;N) = \sum_{M\in \mathrm{M}_{n\times d}(\mathbb{Z})} \widehat{F}(g;M,N) e^{2\pi i \mathrm{tr}({}^{\mathrm{t}}MX)}$$

where

$$\begin{aligned}(7.16)\quad \widehat{F}(g;M,N) &= \int_{\mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})} F(g,T;N) e^{-2\pi i \mathrm{tr}({}^{\mathrm{t}}MT)}\, dT\\ &= \int_{\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})} \int_{\mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})} f\left(T_1, \widetilde{n}_-(T_2)\begin{pmatrix} g & \\ & I_n\end{pmatrix}\right) e^{-2\pi i \mathrm{tr}({}^{\mathrm{t}}NT_1)} e^{-2\pi i \mathrm{tr}({}^{\mathrm{t}}MT_2)}\, dT_2\, dT_1.\end{aligned}$$

Hence we have

$$\mathcal{A}_q(f) = \frac{1}{\#\mathcal{R}_q} \sum_{\substack{N\in \mathrm{M}_{d\times n}(\mathbb{Z})\\ M\in \mathrm{M}_{n\times d}(\mathbb{Z})}} \sum_{\gamma\in\mathcal{B}_q} \sum_{U\in \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})} \widehat{F}\left(D_q\gamma; M, N\right) e^{2\pi i \frac{\mathrm{tr}\left({}^{\mathrm{t}}M\left(\mathbf{0}\, U^{-1}\right)\right) + \mathrm{tr}\left({}^{\mathrm{t}}N\gamma^{-1}\begin{pmatrix}\mathbf{0}\\ U\end{pmatrix}\right)}{q}}.$$

Recalling now the definition of the matrix Kloosterman sum, (5.1), and using $\begin{pmatrix}\mathbf{0}\\ U\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ I_n\end{pmatrix} U$ and

$$\mathrm{tr}\big({}^{\mathrm{t}}M\big(\mathbf{0}\,U^{-1}\big)\big) = \mathrm{tr}\big({}^{\mathrm{t}}MU^{-1}\big(\mathbf{0}\,I_n\big)\big) = \mathrm{tr}\big(\big(\mathbf{0}\,I_n\big)\,{}^{\mathrm{t}}MU^{-1}\big), \tag{7.17}$$

we obtain

$$\mathcal{A}_q(f) = \frac{1}{\#\mathcal{R}_q} \sum_{\substack{N\in \mathrm{M}_{d\times n}(\mathbb{Z})\\ M\in \mathrm{M}_{n\times d}(\mathbb{Z})}} \sum_{\gamma\in\mathcal{B}_q} \widehat{F}\big(D_q\gamma; M, N\big) K_n\Big(\big(\mathbf{0}\,I_n\big)\,{}^{\mathrm{t}}M,\ {}^{\mathrm{t}}N\gamma^{-1}\begin{pmatrix}\mathbf{0}\\ I_n\end{pmatrix}; q\Big). \tag{7.18}$$

We split this sum into three parts by separating out the two cases $N = M = \mathbf{0}$ and $[N \neq \mathbf{0}$, $M = \mathbf{0}]$:

$$\mathcal{A}_q(f) = E_{0,q}(f) + E_{1,q}(f) + E_{2,q}(f), \tag{7.19}$$

where

$$E_{0,q}(f) = \frac{\#\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})}{\#\mathcal{R}_q} \sum_{\gamma\in\mathcal{B}_q} \widehat{F}\left(D_q\gamma; \mathbf{0}, \mathbf{0}\right), \tag{7.20}$$

$$E_{1,q}(f) = \frac{1}{\#\mathcal{R}_q} \sum_{\substack{N\in \mathrm{M}_{d\times n}(\mathbb{Z})\\ N\neq\mathbf{0}}} \sum_{\gamma\in\mathcal{B}_q} \widehat{F}\left(D_q\gamma; \mathbf{0}, N\right) K_n\left(\mathbf{0},\ {}^{\mathrm{t}}N\gamma^{-1}\begin{pmatrix}\mathbf{0}\\ I_n\end{pmatrix}; q\right) \tag{7.21}$$

and

$$E_{2,q}(f) = \frac{1}{\#\mathcal{R}_q} \sum_{\substack{N\in \mathrm{M}_{d\times n}(\mathbb{Z})\\ M\in \mathrm{M}_{n\times d}(\mathbb{Z})\setminus\{\mathbf{0}\}}} \sum_{\gamma\in\mathcal{B}_q} \widehat{F}\left(D_q\gamma; M, N\right) K_n\Big(\big(\mathbf{0}\,I_n\big)\,{}^{\mathrm{t}}M,\ {}^{\mathrm{t}}N\gamma^{-1}\begin{pmatrix}\mathbf{0}\\ I_n\end{pmatrix}; q\Big). \tag{7.22}$$

7.3. **The main term: $E_{0,q}(f)$.** We apply the equidistribution of Hecke points to the sum $E_{0,q}(f)$ in (7.20). Note that by Lemma 3.1, $\widehat{F}(g;\mathbf{0},\mathbf{0})$ is a left $\mathrm{SL}_d(\mathbb{Z})$-invariant function of $g \in \mathrm{SL}_d(\mathbb{R})$. Using (4.5) and $\#\mathcal{R}_q = \#\mathcal{B}_q \cdot \#\,\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$ (which holds by Lemma 2.2), we have

$$E_{0,q}(f) = \frac{1}{\#\mathcal{B}_q} \sum_{\gamma\in\mathcal{B}_q} \widehat{F}\left(D_q\gamma; \mathbf{0}, \mathbf{0}\right) = \big(T_{D_q}\widehat{F}(\,\cdot\,;\mathbf{0},\mathbf{0})\big)(I_d). \tag{7.23}$$

Recall that we are keeping $1 \le n < d$, and that $\kappa', \vartheta', k, \varepsilon, f$ and q are given as in the statement of Theorem 1.2; in particular we have $f \in \mathrm{C}_b^k((\mathbb{R}/\mathbb{Z})^{dn} \times \Gamma\backslash\Gamma\mathrm{H})$ where k is an integer with $k > \kappa' + \varepsilon$. It follows that $\widehat{F}(\,\cdot\,;\mathbf{0},\mathbf{0}) \in \mathrm{C}_b^k(\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R}))$, and now by (7.23) and Proposition 4.2 we have

$$\left|E_{0,q}(f) - \int_{\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})} \widehat{F}(g;\mathbf{0},\mathbf{0})\,d\mu_0(g)\right| \ll_\varepsilon S_{2,\kappa'+\varepsilon}(\widehat{F}(\,\cdot\,;\mathbf{0},\mathbf{0}))\,q^{-\vartheta'+\varepsilon}. \tag{7.24}$$

Here

$$\begin{aligned}\int_{\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})} &\widehat{F}(g;\mathbf{0},\mathbf{0})\,d\mu_0(g)\\ &= \int_{\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})} \int_{\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})} \int_{\mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})} f\left(T_1, \widetilde{n}_-(T_2)\begin{pmatrix} g & \\ & I_n\end{pmatrix}\right) dT_2\ dT_1\,d\mu_0(g)\\ &= \int_{\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})} \int_{\Gamma\backslash\Gamma\mathrm{H}} f(T,g)\,d\mu_{\mathrm{H}}(g)\,dT.\end{aligned} \tag{7.25}$$

Finally, in order to compare $S_{2,\kappa'+\varepsilon}(\widehat{F}(\,\cdot\,;\mathbf{0},\mathbf{0}))$ with $S_{2,\kappa'+\varepsilon}(f)$, let $\mathcal{B}$ and $\mathcal{B}'$ be the fixed linear bases for the Lie algebra of $\mathrm{SL}_d(\mathbb{R})$ and the Lie algebra of $\mathrm{M}_{d\times n}(\mathbb{R})\times\mathrm{H}$ which are used in the definitions of the Sobolev norms; we may then assume that $\mathcal{B} \subset \mathcal{B}'$ when the Lie algebra of $\mathrm{SL}_d(\mathbb{R})$ is embedded

in the Lie algebra of $\mathrm{M}_{d\times n}(\mathbb{R})\times \mathrm{H}$ via the differential of the homomorphism $g\mapsto \Big(\mathbf{0},\left(\begin{smallmatrix} g & \mathbf{0}\\ \mathbf{0} & I_n\end{smallmatrix}\right)\Big)$. Then for any monomial $\mathcal{D}$ in $\mathcal{B}$ of order $\le k$, and every $g\in \mathrm{SL}_d(\mathbb{R})$, we have

$$\big[\mathcal{D}\widehat{F}\big](g;\mathbf{0},\mathbf{0}) = \int_{\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})}\int_{\mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})} [\mathcal{D}f]\left(T_1,\widetilde{n}_-(T_2)\begin{pmatrix} g & \\ & I_n\end{pmatrix}\right)\,dT_2\,dT_1,$$

and hence

$$\big|\big[\mathcal{D}\widehat{F}\big](g;\mathbf{0},\mathbf{0})\big|^2 \le \int_{\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})}\int_{\mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z})} \left|[\mathcal{D}f]\left(T_1,\widetilde{n}_-(T_2)\begin{pmatrix} g & \\ & I_n\end{pmatrix}\right)\right|^2\,dT_2\,dT_1.$$

Integrating the last inequality over g, it follows that

$$\begin{aligned}\big\|\big[\mathcal{D}\widehat{F}\big](\,\cdot\,;\mathbf{0},\mathbf{0})\big\|_{\mathrm{L}^2(\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R}))} &= \sqrt{\int_{\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})}\big|\big[\mathcal{D}\widehat{F}\big](g;\mathbf{0},\mathbf{0})\big|^2\,d\mu_0(g)}\\ &\le \|\mathcal{D}f\|_{\mathrm{L}^2((\mathbb{R}/\mathbb{Z})^{dn}\times\Gamma\backslash\Gamma\mathrm{H})}.\end{aligned}$$

For any integer $0\le k_1\le k$, by summing the above inequality over all monomials in $\mathcal{B}$ of order $\le k_1$, it follows that $S_{2,k_1}\big(\widehat{F}(\,\cdot\,;\mathbf{0},\mathbf{0})\big)\le S_{2,k_1}(f)$. Hence also $S_{2,\lambda}\big(\widehat{F}(\,\cdot\,;\mathbf{0},\mathbf{0})\big)\le S_{2,\lambda}(f)$ for any real number $0\le\lambda\le k$. Using this fact together with (7.25) in (7.24), we conclude:

$$\left|E_{0,q}(f) - \int_{\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z})}\int_{\Gamma\backslash\Gamma\mathrm{H}} f(T,g)\,d\mu_{\mathrm{H}}(g)\,dT\right| \ll_\varepsilon S_{2,\kappa'+\varepsilon}(f)\,q^{-\vartheta'+\varepsilon}. \tag{7.26}$$

7.4. **Error term 1: $E_{1,q}(f)$.** It follows from (7.16) and (7.13) that

$$\big|\widehat{F}(g;M,N)\big| \ll \frac{S_{\infty,\lambda}(f)}{1+\|N\|_\infty^\lambda}$$

for all $0\le\lambda\le k$, $g\in\mathrm{SL}_d(\mathbb{R})$ and $N\in\mathrm{M}_{d\times n}(\mathbb{Z})$, $M\in\mathrm{M}_{n\times d}(\mathbb{Z})$. Using this bound together with Corollary 5.13 in (7.21), we obtain:

$$\begin{aligned}|E_{1,q}(f)| &\ll S_{\infty,\lambda}(f)\frac{q^{n^2}}{\#\mathcal{R}_q}\sum_{\substack{N\in\mathrm{M}_{d\times n}(\mathbb{Z})\\ N\neq\mathbf{0}}}\|N\|_\infty^{-\lambda}\sum_{\gamma\in\mathcal{B}_q}\left(\frac{q}{\gcd\big(q,\,{}^{\mathrm{t}}N\gamma^{-1}\left(\begin{smallmatrix}\mathbf{0}\\ I_n\end{smallmatrix}\right)\big)}\right)^{-n}\\ &\le S_{\infty,\lambda}(f)\frac{q^{n^2-n}}{\#\mathcal{R}_q}\sum_{\substack{N\in\mathrm{M}_{d\times n}(\mathbb{Z})\\ N\neq\mathbf{0}}}\|N\|_\infty^{-\lambda}\sum_{\ell\mid q}\ell^n\mathcal{A}_\ell(N),\end{aligned} \tag{7.27}$$

where

$$\mathcal{A}_\ell(N) = \#\big\{\gamma\in\mathcal{B}_q\ :\ \ell\mid {}^{\mathrm{t}}N\gamma^{-1}\left(\begin{smallmatrix}\mathbf{0}\\ I_n\end{smallmatrix}\right)\big\}.$$

For any $N\in\mathrm{M}_{d\times n}(\mathbb{Z})$, $\ell\mid q$, $\gamma\in\mathcal{B}_q$ and $U\in\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$, by multiplying by U from the right, it follows that the relation $\ell\mid {}^{\mathrm{t}}N\gamma^{-1}\left(\begin{smallmatrix}\mathbf{0}\\ I_n\end{smallmatrix}\right)$ is equivalent with $\ell\mid {}^{\mathrm{t}}N\gamma^{-1}\left(\begin{smallmatrix}\mathbf{0}\\ U\end{smallmatrix}\right)$. Hence by Lemma 2.2,

$$\begin{aligned}\mathcal{A}_\ell(N) &\le \frac{\#\big\{X\in\mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})\ :\ {}^{\mathrm{t}}NX\equiv\mathbf{0}\bmod\ell\big\}}{\#\,\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})}\\ &= \frac{\#\big\{X\in\mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})\ :\ {}^{\mathrm{t}}N'X\equiv\mathbf{0}\bmod\ell'\big\}}{\#\,\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})},\end{aligned} \tag{7.28}$$

where we write $N':=\gcd(\ell,N)^{-1}N$ and $\ell':=\gcd(\ell,N)^{-1}\ell$. To bound the last expression, note that $X\in\mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})$ satisfies the relation ${}^{\mathrm{t}}N'X\equiv\mathbf{0}\bmod\ell'$ if and only if ${}^{\mathrm{t}}N'X\equiv\mathbf{0}\bmod\gcd(p^r,\ell')$ holds for every prime power p^r dividing q (with $r\ge1$). But by construction we have $\gcd(\ell',N')=1$; hence if $p\mid\ell'$ then ${}^{\mathrm{t}}N'$ has at least one row, say $\boldsymbol{n}=\boldsymbol{n}(p)\in\mathbb{Z}^d$, which is not divisible by

p, which means that there are exactly $p^{rdn}/\gcd(p^r,\ell')^n$ matrices $X \in \mathrm{M}_{d\times n}(\mathbb{Z}/p^r\mathbb{Z})$ satisfying $\boldsymbol{n}X \equiv \mathbf{0} \bmod \gcd(p^r,\ell')$. Hence

$$\#\big\{X \in \mathrm{M}_{d\times n}(\mathbb{Z}/p^r\mathbb{Z}) \,:\, {}^{\mathrm{t}}N'X \equiv \mathbf{0} \bmod \gcd(p^r,\ell')\big\} \le \frac{p^{rdn}}{\gcd(p^r,\ell')^n}.$$

Using this bound for each prime power p^r dividing q, and multiplying, it follows that

$$\#\big\{X \in \mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z}) \,:\, {}^{\mathrm{t}}N'X \equiv \mathbf{0} \bmod \ell'\big\} \le \frac{q^{dn}}{{\ell'}^n} = \frac{q^{dn}}{\ell^n}\gcd(\ell,N)^n. \tag{7.29}$$

Recalling also (7.6), we conclude:

$$\mathcal{A}_\ell(N) \ll_{n,\varepsilon} q^{(d-n)n+\varepsilon}\ell^{-n}\gcd(\ell,N)^n. \tag{7.30}$$

Let us also note that, by Lemma 2.1 and since $d > n$,

$$\#\mathcal{R}_q = q^{dn}\prod_{p\mid q}\prod_{j=d+1-n}^{d}(1-p^{-j}) > q^{dn}\prod_{j=d+1-n}^{d}\zeta(j)^{-1} \gg q^{dn}. \tag{7.31}$$

Using the bounds (7.28), (7.30) and (7.31) in (7.27), we obtain:

$$|E_{1,q}(f)| \ll_\varepsilon S_{\infty,\lambda}(f)\,q^{-n+\varepsilon}\sum_{\substack{N\in\mathrm{M}_{d\times n}(\mathbb{Z})\\ N\neq\mathbf{0}}}\|N\|_\infty^{-\lambda}\sum_{\ell\mid q}\gcd(\ell,N)^n.$$

Recall that this holds for any real number λ in the interval $0 \le \lambda \le k$, with k as in Theorem 1.2.

Now note that for each positive integer ℓ we have[7]

$$\sum_{\substack{N\in\mathrm{M}_{d\times n}(\mathbb{Z})\\ N\neq\mathbf{0}}}\|N\|_\infty^{-\lambda}\gcd(\ell,N)^n \le \sum_{\delta\mid\ell}\sum_{\substack{N\in\mathrm{M}_{d\times n}(\delta\mathbb{Z})\\ N\neq\mathbf{0}}}\|N\|_\infty^{-\lambda}\delta^n = \sum_{\delta\mid\ell}\delta^{n-\lambda}\sum_{\substack{N'\in\mathrm{M}_{d\times n}(\mathbb{Z})\\ N'\neq\mathbf{0}}}\|N'\|_\infty^{-\lambda}, \tag{7.32}$$

where we substituted $N = \delta N'$. However, using the fact that

$$\#\big\{N' \in \mathrm{M}_{d\times n}(\mathbb{Z}) \,:\, \|N'\|_\infty = m\big\} \le 2dn(2m+1)^{dn-1} \qquad (\forall m \in \mathbb{Z}_{\ge 0}), \tag{7.33}$$

one verifies that the sum $\sum_{N'\neq\mathbf{0}}\|N'\|_\infty^{-\lambda}$ is finite whenever $\lambda > dn$; and in this case the expression in (7.32) is $\ll_\lambda \sum_{\delta\mid\ell}\delta^{n-\lambda} \ll_\lambda 1$, since $n-\lambda < n-dn \le -1$. We may here choose $\lambda = dn+1$ (this is permissible since $dn+1 \le 2dn < k$), and conclude:

$$|E_{1,q}(f)| \ll_\varepsilon S_{\infty,dn+1}(f)\,q^{-n+\varepsilon}\sum_{\ell\mid q}1 \ll_\varepsilon S_{\infty,dn+1}(f)\,q^{-n+2\varepsilon}. \tag{7.34}$$

7.5. **Error term 2: $E_{2,q}(f)$.** Recalling (7.22), for any $N \in \mathrm{M}_{d\times n}(\mathbb{Z})$ we let

$$E_{2,q}(f;N) = \frac{1}{\#\mathcal{R}_q}\sum_{\substack{M\in\mathrm{M}_{n\times d}(\mathbb{Z})\\ M\neq\mathbf{0}}}\sum_{\gamma\in\mathcal{B}_q}\widehat{F}\,(D_q\gamma;M,N)\,K_n\bigg(\begin{pmatrix}\mathbf{0}\ I_n\end{pmatrix}{}^{\mathrm{t}}M,\ {}^{\mathrm{t}}N\gamma^{-1}\begin{pmatrix}\mathbf{0}\\ I_n\end{pmatrix};q\bigg), \tag{7.35}$$

so that

$$E_{2,q}(f) = \sum_{N\in\mathrm{M}_{d\times n}(\mathbb{Z})}E_{2,q}(f;N). \tag{7.36}$$

[7]Here we work with nonnegative sums taking values in $\mathbb{R}_{\ge 0}\cup\{+\infty\}$; note that a priori we may have $\sum_{\substack{N\in\mathrm{M}_{d\times n}(\mathbb{Z})\\ N\neq\mathbf{0}}}\|N\|_\infty^{-\lambda}\gcd(\ell,N)^n = +\infty$; however our computation shows that $\sum_{\substack{N\in\mathrm{M}_{d\times n}(\mathbb{Z})\\ N\neq\mathbf{0}}}\|N\|_\infty^{-\lambda}\gcd(\ell,N)^n < \infty$ whenever $\lambda > dn$.

By Lemma 3.2, using also (7.12) and (7.14), we have

$$\left|\widehat{F}\left(D_q\gamma; M, N\right)\right| \ll \frac{S_{\infty,\lambda}(f)}{1+\|M\,{}^{\mathrm{t}}(D_q\gamma)^{-1}\|_\infty^\lambda}, \tag{7.37}$$

for any real number λ in the interval $0 \le \lambda \le k$. Also, by (7.13) and (7.16),

$$\left|\widehat{F}\left(D_q\gamma; M, N\right)\right| \ll \frac{S_{\infty,\lambda}(f)}{1+\|N\|_\infty^\lambda}.$$

Hence for all $M \in \mathrm{M}_{n\times d}(\mathbb{Z})$ and $N \in \mathrm{M}_{d\times n}(\mathbb{Z})$, we have

$$\left|\widehat{F}\left(D_q\gamma; M, N\right)\right| \ll \frac{S_{\infty,\lambda}(f)}{1+\|M\,{}^{\mathrm{t}}(D_q\gamma)^{-1}\|_\infty^\lambda+\|N\|_\infty^\lambda}.$$

Using this bound in (7.35), it follows that for every $N \in \mathrm{M}_{d\times n}(\mathbb{Z})$,

$$\begin{aligned}|E_{2,q}(f;N)| &\ll \frac{S_{\infty,\lambda}(f)}{\#\mathcal{R}_q}\sum_{\substack{M\in\mathrm{M}_{n\times d}(\mathbb{Z})\\ M\neq\mathbf{0}}}\sum_{\gamma\in\mathcal{B}_q}\frac{\left|K_n\Big(\big(\mathbf{0}\ I_n\big)\,{}^{\mathrm{t}}M,\ {}^{\mathrm{t}}N\gamma^{-1}\left(\begin{smallmatrix}\mathbf{0}\\ I_n\end{smallmatrix}\right);q\Big)\right|}{1+\|M\,{}^{\mathrm{t}}(D_q\gamma)^{-1}\|_\infty^\lambda+\|N\|_\infty^\lambda}\\ &\ll_\varepsilon \frac{S_{\infty,\lambda}(f)}{\#\mathcal{R}_q}\sum_{\substack{M\in\mathrm{M}_{n\times d}(\mathbb{Z})\\ M\neq\mathbf{0}}}\sum_{\gamma\in\mathcal{B}_q}\frac{q^{n^2-\vartheta+\varepsilon}\gcd\big(q,\big(\mathbf{0}\ I_n\big)\,{}^{\mathrm{t}}M\big)^\vartheta}{1+\|M\,{}^{\mathrm{t}}(D_q\gamma)^{-1}\|_\infty^\lambda+\|N\|_\infty^\lambda},\end{aligned} \tag{7.38}$$

where we recall that $\vartheta = n-1$ if $n \ge 2$, $\vartheta = \frac{1}{2}$ if $n = 1$; in the last step we used (5.34) in Theorem 5.11 if $n \ge 2$, and the classical Weil bound if $n = 1$. Writing here $M = \big(M_0\ \ M_1\big)$ with $M_0 \in \mathrm{M}_{n\times(d-n)}(\mathbb{Z})$ and $M_1 \in \mathrm{M}_n(\mathbb{Z})$, and setting $\ell := \gcd\big(q, \big(\mathbf{0}\ \ I_n\big)\,{}^{\mathrm{t}}M\big) = \gcd(q, M_1)$ and $M_1' := \ell^{-1}M_1$, it follows that

$$\begin{aligned}|E_{2,q}(f;N)| \ll_\varepsilon S_{\infty,\lambda}(f)\frac{q^{n^2-\vartheta+\varepsilon}}{\#\mathcal{R}_q}\sum_{\ell|q}\ell^\vartheta \sum_{M_0\in\mathrm{M}_{n\times(d-n)}(\mathbb{Z})}\ \sum_{\substack{M_1'\in\mathrm{M}_n(\mathbb{Z})\\ M_0=\mathbf{0}\Rightarrow M_1'\neq\mathbf{0}}}\\ \sum_{\gamma\in\mathcal{B}_q}\Big(1+\big\|\big(M_0\ \ \ell M_1'\big)D_q^{-1}\,{}^{\mathrm{t}}\gamma^{-1}\big\|_\infty^\lambda+\|N\|_\infty^\lambda\Big)^{-1}.\end{aligned}$$

Setting now $X := \big(M_0\ \ M_1'\big)$ we have, using (4.1),

$$\big(M_0\quad \ell M_1'\big)\,D_q^{-1} = \ell^{\frac{n}{d}}XD_{q/\ell}^{-1},$$

and thus the last bound can be expressed as follows:

$$\begin{aligned}&\big|E_{2,q}(f;N)\big|\\ &\ll_\varepsilon S_{\infty,\lambda}(f)\frac{q^{n^2-\vartheta+\varepsilon}}{\#\mathcal{R}_q}\sum_{\ell|q}\ell^\vartheta\sum_{\substack{X\in\mathrm{M}_{n\times d}(\mathbb{Z})\\ X\neq\mathbf{0}}}\sum_{\gamma\in\mathcal{B}_q}\Big(1+\|N\|_\infty^\lambda+\ell^{\frac{n}{d}\lambda}\big\|XD_{q/\ell}^{-1}\,{}^{\mathrm{t}}\gamma^{-1}\big\|_\infty^\lambda\Big)^{-1}.\end{aligned} \tag{7.39}$$

Assuming from now on that $\lambda > nd$, and using the majorant function $\Phi_{a,b}^{(\kappa)}$ introduced in (6.1), the last bound can be expressed:

$$\big|E_{2,q}(f;N)\big| \ll_\varepsilon S_{\infty,\lambda}(f)\frac{q^{n^2-\vartheta+\varepsilon}}{\#\mathcal{R}_q}\sum_{\ell|q}\ell^\vartheta\sum_{\gamma\in\mathcal{B}_q}\Phi^{(\lambda)}_{1+\|N\|_\infty^\lambda,\ell^{\lambda n/d}}\big(D_{q/\ell}^{-1}\,{}^{\mathrm{t}}\gamma^{-1}\big). \tag{7.40}$$

We will need the following simple lemma.

Lemma 7.1. *For any function $\Phi : \mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R}) \to \mathbb{C}$, $u \mid q$ and $g \in \mathrm{SL}_d(\mathbb{R})$,*

$$\sum_{\gamma\in\mathcal{B}_q} \Phi\big(D_u^{-1}\ {}^t\gamma^{-1}g\big) = \#\mathcal{B}_q \cdot \big(T^*_{D_u}\Phi\big)(g). \tag{7.41}$$

Proof. We have

$$\sum_{\gamma\in\mathcal{B}_q} \Phi\big(D_u^{-1}\ {}^t\gamma^{-1}g\big) = \sum_{\gamma_1\in\Gamma^0(q)\backslash\Gamma^0(u)} \sum_{\gamma_2\in\mathcal{B}_u} \Phi\big(D_u^{-1}\ {}^t(\gamma_1\gamma_2)^{-1}\,g\big).$$

But $\gamma_1 \in \Gamma^0(u)$ implies $D_u\,\gamma_1 D_u^{-1} \in \mathrm{SL}_d(\mathbb{Z})$; hence $D_u^{-1}\ {}^t\gamma_1^{-1}D_u \in \mathrm{SL}_d(\mathbb{Z})$ and $\Phi\big(D_u^{-1}\ {}^t(\gamma_1\gamma_2)^{-1}\,g\big) = \Phi\big(D_u^{-1}\ {}^t\gamma_2^{-1}\,g\big)$, and so we get

$$\begin{aligned}\sum_{\gamma\in\mathcal{B}_q} \Phi\big(D_u^{-1}\ {}^t\gamma^{-1}g\big) &= \#\big(\Gamma^0(q)\backslash\Gamma^0(u)\big) \sum_{\gamma_2\in\mathcal{B}_u} \Phi\big(D_u^{-1}\ {}^t\gamma_2^{-1}g\big)\\ &= \#\big(\Gamma^0(q)\backslash\Gamma^0(u)\big)\cdot\#\mathcal{B}_u\cdot(T^*_{D_u}\Phi)(g) = \#\mathcal{B}_q\cdot(T^*_{D_u}\Phi)(g),\end{aligned}$$

where the second equality holds by (4.7). □

Using Lemma 7.1 and $\#\mathcal{R}_q = \#\mathcal{B}_q \cdot \#\,\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$, the bound in (7.40) can be rewritten as follows:

$$\big|E_{2,q}(f;N)\big| \ll_\varepsilon S_{\infty,\lambda}(f)\frac{q^{n^2-\vartheta+\varepsilon}}{\#\,\mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})}\sum_{\ell\mid q}\ell^\vartheta\cdot\Big[T^*_{D_{q/\ell}}\Phi^{(\lambda)}_{1+\|N\|_\infty^\lambda,\ell^{\lambda n/d}}\Big](I_d). \tag{7.42}$$

We will bound $\big(T^*_{D_{q/\ell}}\Phi^{(\lambda)}_{a,b}\big)(I_d)$ from above by an integral over $\mathrm{SL}_d(\mathbb{Z})\backslash\,\mathrm{SL}_d(\mathbb{R})$. Fix a fundamental domain $\mathcal{F}_d$ for $\mathrm{SL}_d(\mathbb{Z})\backslash\,\mathrm{SL}_d(\mathbb{R})$ containing I_d in its interior, and then fix an open neighbourhood $\Omega \subset \mathcal{F}_d$ of I_d so small that for every $w \in \Omega$ and every $\boldsymbol{v} \in \mathbb{R}^d$,

$$\frac{1}{2}\|\boldsymbol{v}\|_\infty \le \|\boldsymbol{v}w\|_\infty \le 2\|\boldsymbol{v}\|_\infty. \tag{7.43}$$

This implies that for every $w \in \Omega$ and every $A \in \mathrm{M}_{n\times d}(\mathbb{R})$,

$$\frac{1}{2}\|A\|_\infty \le \|Aw\|_\infty \le 2\|A\|_\infty. \tag{7.44}$$

Hence for any $a, b > 0$, $g \in \mathrm{SL}_d(\mathbb{R})$, $w \in \Omega$ and $X \in \mathrm{M}_{n\times d}(\mathbb{R})$, we have

$$2^{-\lambda}\big(a + b\|Xg\|_\infty^\lambda\big) \le a + b\|Xgw\|_\infty^\lambda \le 2^\lambda\big(a + b\|Xg\|_\infty^\lambda\big),$$

and thus, recalling (6.1), we conclude that

$$2^{-\lambda}\Phi^{(\lambda)}_{a,b}(g) \le \Phi^{(\lambda)}_{a,b}(gw) \le 2^\lambda\Phi^{(\lambda)}_{a,b}(g). \tag{7.45}$$

Recalling now that, by (4.7),

$$\big[T^*_{D_{q/\ell}}\Phi\big](g) = \frac{1}{\#\mathcal{B}_{q/\ell}}\sum_{\gamma\in\mathcal{B}_{q/\ell}}\Phi(D_{q/\ell}^{-1}\ {}^t\gamma^{-1}g),$$

and applying the left inequality in (7.45) with $g = D_{q/\ell}^{-1}\ {}^t\gamma^{-1}$ for each $\gamma \in \mathcal{B}_{q/\ell}$, we conclude that

$$2^{-\lambda}\big[T^*_{D_{q/\ell}}\Phi^{(\lambda)}_{a,b}\big](I_d) \le \big[T^*_{D_{q/\ell}}\Phi^{(\lambda)}_{a,b}\big](w), \qquad \forall w \in \Omega.$$

Hence

$$\big[T^*_{D_{q/\ell}}\Phi^{(\lambda)}_{a,b}\big](I_d) \le \frac{2^\lambda}{\int_\Omega d\mu_0(w)}\int_\Omega\big[T^*_{D_{q/\ell}}\Phi^{(\lambda)}_{a,b}\big](w)\,d\mu_0(w) \le \frac{2^\lambda}{\int_\Omega d\mu_0(w)}\int_{\mathcal{F}_d}\big[T^*_{D_{q/\ell}}\Phi^{(\lambda)}_{a,b}\big](g)\,d\mu_0(g), \tag{7.46}$$

where the last inequality holds since $\Phi_{a,b}^{(\lambda)}(g) > 0$ everywhere. It should be noted that in (7.46) we are again working with nonnegative sums and integrals taking values in $\mathbb{R}_{\geq 0} \cup \{+\infty\}$; a priori one or both of the integrals in (7.46) may equal $+\infty$, however we will see below that this is not the case.

Next, using (4.6) we have $\langle T_{D_q}^* \Phi, 1\rangle = \langle \Phi, T_{D_q} 1\rangle = \langle \Phi, 1\rangle$ for all $\Phi \in \mathrm{L}^2(\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R}))$, i.e.,

$$\int_{\mathcal{F}_d} \big[T_{D_q}^* \Phi\big](g)\, d\mu_0(g) = \int_{\mathcal{F}_d} \Phi(g)\, d\mu_0(g). \tag{7.47}$$

It follows that (7.47) also holds as a relation in $\mathbb{R}_{\geq 0} \cup \{+\infty\}$, for any left $\mathrm{SL}_d(\mathbb{Z})$-invariant Borel measurable function $\Phi : \mathrm{SL}_d(\mathbb{R}) \to \mathbb{R}_{\geq 0}$. Using this fact in (7.46) we conclude:

$$\big[T_{D_{q/\ell}}^* \Phi_{a,b}^{(\lambda)}\big](I_d) \leq \frac{2^\lambda}{\int_\Omega d\mu_0(w)} \int_{\mathcal{F}_d} \Phi_{a,b}^{(\lambda)}(g)\, d\mu_0(g), \tag{7.48}$$

and by Proposition 6.4 (and since we are assuming $\lambda > nd$) this implies

$$\big[T_{D_{q/\ell}}^* \Phi_{a,b}^{(\lambda)}\big](I_d) \ll_\lambda a^{-1} \Big(\frac{a}{b}\Big)^{\frac{d}{\lambda}} \Big(1 + \frac{a}{b}\Big)^{(n-1)\frac{d}{\lambda}}. \tag{7.49}$$

Using (7.6) and (7.49) in (7.42), we obtain:

$$|E_{2,q}(f; N)| \ll_{\lambda,\varepsilon} S_{\infty,\lambda}(f) q^{-\vartheta+2\varepsilon} \sum_{\ell|q} \ell^\vartheta \cdot \big(1 + \|N\|_\infty^\lambda\big)^{\frac{d}{\lambda}-1} \ell^{-n} \bigg(1 + \frac{1 + \|N\|_\infty^\lambda}{\ell^{\lambda n/d}}\bigg)^{(n-1)\frac{d}{\lambda}}. \tag{7.50}$$

Hence, using also (7.36) and (7.33) (with $m = a-1$), we have

$$|E_{2,q}(f)| \ll_{\lambda,\varepsilon} S_{\infty,\lambda}(f) q^{-\vartheta+2\varepsilon} \sum_{\ell|q} \ell^{\vartheta-n} \cdot \sum_{a=1}^\infty a^{dn-1} \cdot a^{d-\lambda} \bigg(1 + \frac{a}{\ell^{n/d}}\bigg)^{(n-1)d}.$$

Here we must require $\lambda > 2dn$ in order for the sum over a to converge. *Assuming* $\lambda > 2dn$, the sum over a is bounded independently of ℓ, since

$$\sum_{a=1}^\infty a^{dn-1} \cdot a^{d-\lambda} \bigg(1 + \frac{a}{\ell^{n/d}}\bigg)^{(n-1)d} \leq \sum_{a=1}^\infty (a+1)^{2dn-1-\lambda} \ll_\lambda 1.$$

Hence, using also $\vartheta < n$, we obtain:

$$|E_{2,q}(f)| \ll_{\lambda,\varepsilon} S_{\infty,\lambda}(f) q^{-\vartheta+3\varepsilon}. \tag{7.51}$$

Setting here $\varepsilon = \frac{1}{3}\varepsilon_{\text{new}}$ and then choosing $\lambda = \kappa + \varepsilon_{\text{new}}$ (recall that $\kappa = 2dn$), the bound becomes $S_{\infty,\kappa+\varepsilon}(f) q^{-\vartheta+\varepsilon}$. Note that this bound subsumes the one in (7.34), since $\vartheta < n$ and $dn + 1 < \kappa$. Hence, recalling that $\mathcal{A}_q(f) = E_{0,q}(f) + E_{1,q}(f) + E_{2,q}(f)$ (see (7.19)) and using (7.26), (7.34) and (7.51), we obtain (1.12), i.e. we have proved Theorem 1.2 in the case $n < d$. □

8. An application and a by-product

In this section, we illustrate how both our result and its proof can be used to prove statements about solutions of Diophantine equations over a finite field $\mathbb{F}_p$, $p \geq 3$ prime, in small boxes (see also [Shp15] for a comprehensive survey).

Indeed, Section 8.1 is an application of our main theorem, Theorem 1.2, to estimating the probability that a randomly chosen (according to a rather general probability measure) system of affine congruences has a given number of small solutions.

In Section 8.2, which is rather an application of the technique introduced in the proof of the main theorem, we give a sharp upper bound – and prove the corresponding lower bound in a much more elementary way – for the number of $\mathbb{F}_p$-points in small boxes on the variety of the set of (rectangular) matrices with a given rank.

8.1. Application: small solutions of linear congruences.

Let p be an odd prime, and consider the affine variety $V \subset \mathbb{A}^d$ defined by the system of equations

$$f_1(x_1, \dots, x_d) = \cdots = f_n(x_1, \dots, x_d) = 0, \tag{8.1}$$

where $f_1, \dots, f_n$ are polynomials in $\mathbb{F}_p[X_1, \dots, X_d]$. An important question is what can be said about existence of $\mathbb{F}_p$-points of V, or about the number of $\mathbb{F}_p$-points of V, inside a small "box" or more general small domain in $\mathbb{F}_p^d$; see, e.g., the recent survey [Shp15]. In particular, a much studied problem is how small *integer* solutions the system (8.1) has. For a random choice of polynomials $f_1, \dots, f_n$, one expects the size of the smallest integer solution to typically be of size comparable to $p^{n/d}$.

In [SV05], this question was studied for a random system of linear congruences. It was proved in [SV05] that if the variety V is taken uniformly random among all linear, or all affine linear, subspaces of $\mathbb{F}_p^d$ of codimension n, then for any given nice subset Ω of $\mathbb{R}^d$, as $p \to \infty$, there exists an explicit limit distribution for the number of integer points which lie in $p^{n/d}\Omega$ (viz., are "small") and which project to points in V.

In Theorem 8.1 below we prove a variant of these results, where instead the linear polynomials $f_1, \dots, f_n$ in (8.1) are taken random with respect to a given probability measure of a fairly general type: We take the constant terms of $f_1, \dots, f_n$ to be arbitrary *fixed* integers $b_1, \dots, b_n$, while the tuple of degree one coefficients is chosen uniformly random among all points R in $\mathbb{F}_p^{dn}$ such that $p^{-1}R$ belongs to a given (nice) subset U of the torus $(\mathbb{R}/\mathbb{Z})^{dn}$ and the equations are linearly independent. In other words, we ask about the number of integer solutions $\boldsymbol{x} \in \mathbb{Z}^d$ of size $\ll p^{n/d}$ to the congruence equation $\boldsymbol{x}R \equiv \boldsymbol{b} \bmod p$, for fixed $\boldsymbol{b} \in \mathbb{Z}^n$ and R chosen uniformly random in the set $\{R \in \mathcal{R}_p \,:\, p^{-1}R \in U\}$. We will prove that, for any given nice subset Ω of $\mathbb{R}^d$, there exists an explicit limit distribution for the number of such solutions $\boldsymbol{x}$ in $\mathbb{Z}^d \cap p^{n/d}\Omega$, as $p \to \infty$. In fact, our proof allows the modulus p to run through all *integers*, and we will state the theorem in this form, writing q in place of p.

We say that a subset Ω of Euclidean space $\mathbb{R}^m$ or of the torus $(\mathbb{R}/\mathbb{Z})^m$ ($m \geq 1$) is *smooth* if $\mathrm{vol}(\partial_\varepsilon\Omega) \ll \varepsilon$ as $\varepsilon \to 0$, where $\partial_\varepsilon\Omega$ is the ε-neighborhood of the boundary of Ω. In the following we will view $\mathrm{M}_{d\times n}(\mathbb{Z}/q\mathbb{Z})$ as a subset of $(\mathbb{R}/q\mathbb{Z})^{dn}$; this means that for any subset $U \subset (\mathbb{R}/\mathbb{Z})^{dn}$, we can write $\mathcal{R}_q \cap qU$ for the set of all $R \in \mathcal{R}_q$ satisfying $q^{-1}R \in U$.

Theorem 8.1. *Let $d > n \geq 1$; let U be a smooth subset of $(\mathbb{R}/\mathbb{Z})^{dn}$ of positive volume; let Ω be a smooth and bounded subset of $\mathbb{R}^d$; let $\boldsymbol{b} \in \mathbb{Z}^n$, and let $\varepsilon > 0$. If $\boldsymbol{b} = \boldsymbol{0}$ then we assume that Ω contains a neighborhood of the origin. Then for any $r \in \mathbb{Z}_{\geq 0}$ there exists a constant $c(\Omega, \boldsymbol{b}, r) \geq 0$ such that, for any positive integer q, the number of $R \in \mathcal{R}_q \cap qU$ such that the congruence equation*

$\boldsymbol{x}R \equiv \boldsymbol{b} \bmod q$ has exactly r solutions $\boldsymbol{x}$ in $\mathbb{Z}^d \cap q^{n/d}\Omega$ is

$$\#\big(\mathcal{R}_q \cap qU\big) \cdot \big(c(\Omega, \boldsymbol{b}, r) + O_{U,\Omega,\boldsymbol{b},r,\varepsilon}(q^{-\alpha+\varepsilon})\big), \tag{8.2}$$

where $\alpha = \alpha(d,n) = \min\Big(\dfrac{n-1}{1+2dn}, \dfrac{n}{d^2}, \dfrac{d-n}{d^2}\Big)$ *if* $n \geq 2$, *and* $\alpha(d,1) = \min\Big(\dfrac{1}{2+4d}, \dfrac{1}{d^2}\Big)$.

In order to state the explicit formula for the limit probabilities $c(\Omega, \boldsymbol{b}, r)$ in Theorem 8.1, let $\mathrm{ASL}_d(\mathbb{R})$ be the affine special linear group of order d, that is, $\mathrm{ASL}_d(\mathbb{R}) = \mathrm{SL}_d(\mathbb{R}) \ltimes \mathbb{R}^d$ with multiplication law

$$(g, \boldsymbol{v})(g', \boldsymbol{v}') = (gg', \boldsymbol{v}g' + \boldsymbol{v}') \qquad \big(g, g' \in \mathrm{SL}_d(\mathbb{R}),\ \boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^d\big).$$

(Note that $\mathrm{ASL}_d(\mathbb{R})$ is isomorphic with our group H in the special case $n = 1$.) The group $\mathrm{ASL}_d(\mathbb{R})$ acts on $\mathbb{R}^d$ from the right through $\boldsymbol{x}(g, \boldsymbol{v}) := \boldsymbol{x}g + \boldsymbol{v}$ ($\boldsymbol{x} \in \mathbb{R}^d$). We identify the homogeneous space $\mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R})$ with the space of *grids* (=translates of lattices) of covolume one in $\mathbb{R}^d$, through $\mathrm{ASL}_d(\mathbb{Z})g \leftrightarrow \mathbb{Z}^d g$ ($g \in \mathrm{ASL}_d(\mathbb{R})$), and we denote by μ the invariant probability measure on $\mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R})$. We take $\mathrm{SL}_d(\mathbb{R})$ to be embedded in $\mathrm{ASL}_d(\mathbb{R})$ through $g \mapsto (g, \boldsymbol{0})$; thus $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$ becomes identified in the standard way with the space of lattices of covolume one in $\mathbb{R}^d$. Recall that μ_0 denotes the invariant probability measure on $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$. Now we have:

$$c(\Omega, \boldsymbol{b}, r) = \begin{cases} \mu_0(\{g \in \mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R}) \ :\ \#(\mathbb{Z}^d g \cap \Omega) = r\}) & \text{if } \boldsymbol{b} = \boldsymbol{0}; \\ \mu(\{g \in \mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R}) \ :\ \#(\mathbb{Z}^d g \cap \Omega) = r\}) & \text{if } \boldsymbol{b} \neq \boldsymbol{0}. \end{cases} \tag{8.3}$$

In particular note that for $\boldsymbol{b} \neq \boldsymbol{0}$, $c(\Omega, \boldsymbol{b}, r)$ is independent of $\boldsymbol{b}$!

Remark 8.2. The formulas for the limit probabilities in (8.3) are the same as those in [SV05]. In fact, in the case $\boldsymbol{b} = \boldsymbol{0}$, by specializing to $U = (\mathbb{R}/\mathbb{Z})^{dn}$ and restricting q to run through primes, Theorem 8.1 gives back [SV05, Theorem 2] but with a weaker error term. Indeed, as R runs through $\mathcal{R}_p$, the set $\{\boldsymbol{x} \in \mathbb{F}_p^d : \boldsymbol{x}R = \boldsymbol{0}\}$ runs through all the linear subspaces of $\mathbb{F}_p^d$ of codimension n, visiting each such subspace exactly $\prod_{j=1}^{n-1}(p^n - p^j)$ times. Similarly, the limit result of [SV05, Theorem 3] (without an error term) follows *formally* by applying Theorem 8.1 with $q = p$ prime, $U = (\mathbb{R}/\mathbb{Z})^{dn}$, and averaging over all $\boldsymbol{b}$ in $\mathbb{F}_p^n$; this is of course not a rigorous deduction, since $\boldsymbol{b}$ is required to be a fixed *integer* vector in Theorem 8.1, and the error term in (8.2) is allowed to depend on $\boldsymbol{b}$ in an uncontrolled way.

Remark 8.3. As we will see, the proof of Theorem 8.1 can easily be extended to give the following more general statement: Let d, n, U be as in Theorem 8.1; let $k \in \mathbb{Z}^+$, and let $\Omega_1, \ldots, \Omega_k$ be smooth and bounded subsets of $\mathbb{R}^d$. Let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k \in \mathbb{Z}^n$; for each j such that $\boldsymbol{b}_j = \boldsymbol{0}$, we assume that Ω_j contains a neighbourhood of the origin. Let $r_1, \ldots, r_k \in \mathbb{Z}_{\geq 0}$. Then for any $q \in \mathbb{Z}^+$, the number of $R \in \mathcal{R}_q \cap qU$ such that for each $j = 1, \ldots, k$, the equation $\boldsymbol{x}R \equiv \boldsymbol{b}_j \bmod q$ has exactly r_j solutions $\boldsymbol{x} \in \mathbb{Z}^d \cap q^{n/d}\Omega_j$, is

$$\#\big(\mathcal{R}_q \cap qU\big) \cdot \big(c + O(q^{-\alpha+\varepsilon})\big), \tag{8.4}$$

where $\alpha = \alpha(d,n)$ is as before, $c \in \mathbb{R}_{\geq 0}$ is a constant which depends on $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$, $\Omega_1, \ldots, \Omega_k$ and $r_1, \ldots, r_k$ (see (8.17) below), and where the implied constant in the "big O" may depend on U, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k, \Omega_1, \ldots, \Omega_k, r_1, \ldots, r_k, \varepsilon$.

Remark 8.4. In the case $d = n$ we have $\mathcal{R}_q = \mathrm{GL}_n(\mathbb{Z}/q\mathbb{Z})$ for every $q \in \mathbb{Z}^+$, and hence for any $\boldsymbol{b} \in \mathbb{Z}^n$, the equation $\boldsymbol{x}R \equiv \boldsymbol{b} \bmod q$ has a unique solution $\boldsymbol{x} = \boldsymbol{b}R^{-1}$ in $(\mathbb{Z}/q\mathbb{Z})^n$. We now have the following result analogous to Theorem 8.1: For any smooth subsets $U \subset (\mathbb{R}/\mathbb{Z})^{n^2}$ and $\Omega \subset (\mathbb{R}/\mathbb{Z})^n$ of positive volume, and any $\boldsymbol{b} \in \mathbb{Z}^n \setminus \{\boldsymbol{0}\}$ and $q \in \mathbb{Z}^+$, the number of $R \in \mathcal{R}_q \cap qU$ such that $\boldsymbol{x} = \boldsymbol{b}R^{-1}$ lies in $q\Omega$ equals

$$\#\big(\mathcal{R}_q \cap qU\big)\,\mathrm{vol}(\Omega)(1 + O_{U,\Omega,\boldsymbol{b},\varepsilon}(q^{-\beta+\varepsilon})),$$

where $\beta = \beta(n) = \frac{n-1}{1+2n^2}$ if $n \geq 2$, and $\beta(1) = \frac{1}{6}$. We give the proof at the end of the present section. In analogy with Remark 8.3, the above result may also be generalized into an asymptotic formula for the number of $R \in \mathcal{R}_q \cap qU$ such that $\boldsymbol{x} = \boldsymbol{b}_j R^{-1}$ lies in $q\Omega_j$ for each $j = 1, \ldots, k$.

We now start preparing for the proof of Theorem 8.1. For each $\boldsymbol{b} \in \mathbb{R}^n$ we denote by $J_{\boldsymbol{b}}$ the following Lie group homomorphism:

$$J_{\boldsymbol{b}} : \mathrm{H} \to \mathrm{ASL}_d(\mathbb{R}), \qquad J_{\boldsymbol{b}} \begin{pmatrix} Z & \boldsymbol{0} \\ V & I_n \end{pmatrix} = (Z, \boldsymbol{b}V). \tag{8.5}$$

If $\boldsymbol{b} \in \mathbb{Z}^n$ then $J_{\boldsymbol{b}}(\Gamma \cap \mathrm{H}) \subset \mathrm{ASL}_d(\mathbb{Z})$, and hence $J_{\boldsymbol{b}}$ induces a smooth map

$$\widetilde{J}_{\boldsymbol{b}} : \Gamma\backslash\Gamma\mathrm{H} \to \mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R}).$$

Lemma 8.5. *For any $q \in \mathbb{Z}^+$, $R \in \mathcal{R}_q$, $\boldsymbol{b} \in \mathbb{Z}^n$, and any subset $\Omega \subset \mathbb{R}^d$, the number of solutions $\boldsymbol{x} \in \mathbb{Z}^d \cap q^{n/d}\Omega$ to the congruence equation $\boldsymbol{x}R \equiv \boldsymbol{b} \bmod q$ equals $\#\big(\mathbb{Z}^d \widetilde{J}_{\boldsymbol{b}}\big(\widetilde{n}_+(q^{-1}R)D(q)\big) \cap \Omega\big)$.*

(Here for any $\alpha \in \mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R})$ we write "$\mathbb{Z}^d\alpha$" for the corresponding grid; thus $\mathbb{Z}^d\alpha := \mathbb{Z}^d g$ for any $g \in \mathrm{ASL}_d(\mathbb{R})$ such that $\alpha = \mathrm{ASL}_d(\mathbb{Z})g$.)

Proof. Let R' be a lift of R to $\mathrm{M}_{d\times n}(\mathbb{Z})$. We know from Lemma 1.1 that $\widetilde{n}_+(q^{-1}R)D(q) \in \Gamma\backslash\Gamma\mathrm{H}$; hence there exists some $\gamma \in \Gamma$ such that $\gamma\, n_+(q^{-1}R')D(q) \in \mathrm{H}$. Writing $\gamma = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ (with $A \in \mathrm{M}_d(\mathbb{Z})$, $B \in \mathrm{M}_{d\times n}(\mathbb{Z})$, $C \in \mathrm{M}_{n\times d}(\mathbb{Z})$, $D \in \mathrm{M}_n(\mathbb{Z})$), we then have

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} q^{-\frac{n}{d}} I_d & R' \\ \boldsymbol{0} & qI_n \end{pmatrix} = \begin{pmatrix} q^{-\frac{n}{d}} A & \boldsymbol{0} \\ q^{-\frac{n}{d}} C & I_n \end{pmatrix}. \tag{8.6}$$

In particular we have $AR' + qB = \boldsymbol{0}$ in $\mathrm{M}_{d\times n}(\mathbb{Z})$, and this implies that the lattice $\mathbb{Z}^d A$ is contained in the kernel of the homomorphism $\boldsymbol{x} \mapsto [\boldsymbol{x}R' \bmod q]$ from $\mathbb{Z}^d$ to $\mathbb{Z}^n/q\mathbb{Z}^n$. This homomorphism is surjective since $R \in \mathcal{R}_q$; hence the kernel is a subgroup of index q^n in $\mathbb{Z}^d$; furthermore, (8.6) implies $\det A = q^n$, so that also $\mathbb{Z}^d A$ has index q^n in $\mathbb{Z}^d$. Hence $\mathbb{Z}^d A$ in fact *equals* the kernel:

$$\mathbb{Z}^d A = \{\boldsymbol{x} \in \mathbb{Z}^d \,:\, \boldsymbol{x}R' \equiv \boldsymbol{0} \bmod q\}. \tag{8.7}$$

Also from (8.6) we have $CR' + qD = I_n$; hence $\boldsymbol{b}CR' \equiv \boldsymbol{b} \bmod q$. This fact combined with (8.7) implies

$$\{\boldsymbol{x} \in \mathbb{Z}^d \,:\, \boldsymbol{x}R \equiv \boldsymbol{b} \bmod q\} = \mathbb{Z}^d A + \boldsymbol{b}C.$$

Note also that (8.6) implies $\mathbb{Z}^d \widetilde{J}_{\boldsymbol{b}}\big(\widetilde{n}_+(q^{-1}R)D(q)\big) = q^{-\frac{n}{d}}(\mathbb{Z}^d A + \boldsymbol{b}C)$. The lemma follows from the last two facts. □

Lemma 8.6. *For any $\boldsymbol{b} \in \mathbb{Z}^n \setminus \{\boldsymbol{0}\}$, $\mu_{\mathrm{H}} \circ \widetilde{J}_{\boldsymbol{b}}^{-1} = \mu$. On the other hand, for $\boldsymbol{b} = \boldsymbol{0}$ we have $\widetilde{J}_{\boldsymbol{0}}(\Gamma\backslash\Gamma\mathrm{H}) = \mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$ and $\mu_{\mathrm{H}} \circ \widetilde{J}_{\boldsymbol{0}}^{-1} = \mu_0$.*

Proof. As before, let $\mathcal{F}_d \subset \mathrm{SL}_d(\mathbb{R})$ be a fundamental domain for $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$; also let $\widetilde{\mu}_0$ be the Haar measure on $\mathrm{SL}_d(\mathbb{R})$ which induces the measure μ_0 on $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$. Then the measure μ_{H} can be explicitly described as follows: For any Borel set $B \subset \Gamma\backslash\Gamma\mathrm{H}$,

$$\mu_{\mathrm{H}}(B) = \int_{\mathcal{F}_d} \mathrm{vol}\left(\left\{ X \in \mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z}) \,:\, \widetilde{n}_-(X) \begin{pmatrix} g & \boldsymbol{0} \\ \boldsymbol{0} & I_n \end{pmatrix} \in B \right\} \right) d\widetilde{\mu}_0(g), \tag{8.8}$$

where vol is the Lebesgue measure on the torus $\mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z}) \cong (\mathbb{R}/\mathbb{Z})^{dn}$. Similarly, if we introduce the map $\iota : (\mathbb{R}/\mathbb{Z})^d \to \mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R})$ by setting $\iota(\boldsymbol{x}) := \mathrm{ASL}_d(\mathbb{Z})(I_d, \boldsymbol{x}')$ where $\boldsymbol{x}'$ is any lift to $\mathbb{R}^d$ of $\boldsymbol{x} \in (\mathbb{R}/\mathbb{Z})^d$, then for any Borel set $A \subset \mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R})$,

$$\mu(A) = \int_{\mathcal{F}_d} \mathrm{vol}\big(\{\boldsymbol{x} \in (\mathbb{R}/\mathbb{Z})^d \,:\, \iota(\boldsymbol{x})(g, \boldsymbol{0}) \in A\}\big)\, d\widetilde{\mu}_0(g), \tag{8.9}$$

where now vol also denotes the Lebesgue measure on $(\mathbb{R}/\mathbb{Z})^d$. Assuming $\boldsymbol{b} \in \mathbb{Z}^d \setminus \{\mathbf{0}\}$, our task is to prove that

$$\mu_{\mathrm{H}}(\widetilde{J}_{\boldsymbol{b}}^{-1}(A)) = \mu(A) \tag{8.10}$$

holds for any Borel set $A \subset \mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R})$. This follows using the formulas (8.8) and (8.9), together with the fact that

$$\widetilde{n}_-(X)\begin{pmatrix} g & \mathbf{0} \\ \mathbf{0} & I_n \end{pmatrix} \in \widetilde{J}_{\boldsymbol{b}}^{-1}(A) \Leftrightarrow \iota(\boldsymbol{b}X)\cdot(g,\mathbf{0}) \in A \qquad (\forall X \in \mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z}),\ g \in \mathrm{SL}_d(\mathbb{R})),$$

and the fact that for any Borel set $A' \subset (\mathbb{R}/\mathbb{Z})^d$,

$$\mathrm{vol}\big(\{X \in \mathrm{M}_{n\times d}(\mathbb{R}/\mathbb{Z}) \,:\, \boldsymbol{b}X \in A'\}\big) = \mathrm{vol}(A'). \tag{8.11}$$

In the remaining case, $\boldsymbol{b} = \mathbf{0}$, the statements of the lemma are immediate from (8.5) and (8.8). □

As in [SV05], we introduce a notion of smoothness for subsets of arbitrary homogeneous spaces, as follows: Let $X = \Lambda\backslash L$ where L is a Lie group and Λ a lattice in L, and let $\widetilde{\mu}$ be the L-invariant probability measure on X. (We will apply the following to the three cases $X = (\Gamma \cap \mathrm{H})\backslash \mathrm{H}$, $X = \mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R})$ and $X = \mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$.) We fix a left invariant Riemannian metric d on L. This metric descends to a Riemannian metric on $X = \Lambda\backslash L$, which we also denote by d, and using this metric, for any subset $\Omega \subset X$ and any $\varepsilon > 0$, we define the ε-neighborhood of the boundary of Ω,

$$\partial_\varepsilon \Omega := \big\{p \in X \,:\, \big[\exists q \in \partial\Omega \text{ s.t. } \mathsf{d}(p,q) < \varepsilon\big]\big\}.$$

Now the set Ω is said to be *smooth* if $\widetilde{\mu}(\partial_\varepsilon\Omega) \ll \varepsilon$ as $\varepsilon \to 0$.

Next, for any subset $\Omega \subset \mathbb{R}^d$ and any $r \in \mathbb{Z}_{\geq 0}$, we let $\widetilde{\Omega}_r$ be the subset of $\mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R})$ corresponding to those grids of covolume one in $\mathbb{R}^d$ which intersect Ω in exactly r points, viz.,

$$\widetilde{\Omega}_r = \big\{\mathrm{ASL}_d(\mathbb{Z})g \,:\, g \in \mathrm{ASL}_d(\mathbb{R}),\ \#(\mathbb{Z}^d g \cap \Omega) = r\big\}. \tag{8.12}$$

Lemma 8.7. *For any smooth subset $\Omega \subset \mathbb{R}^d$, any $r \in \mathbb{Z}_{\geq 0}$ and any $\boldsymbol{b} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$, $\widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)$ is a smooth subset of $\Gamma\backslash\Gamma\mathrm{H}$.*

Proof. As in [SV05, Lemma 10], one proves that $\widetilde{\Omega}_r$ is a smooth subset of $\mathrm{ASL}_d(\mathbb{Z})\backslash \mathrm{ASL}_d(\mathbb{R})$. Next, it is an immediate verification from (8.5) that $\|dJ_{\boldsymbol{b}}(\boldsymbol{v})\| \leq C\|\boldsymbol{v}\|$ holds for any point $h \in \mathrm{H}$ and any tangent vector $\boldsymbol{v} \in T_h\mathrm{H}$, where the two norms are the Riemannian norms on $T_{J_{\boldsymbol{b}}(h)}(\mathrm{ASL}_d(\mathbb{R}))$, and on $T_h\mathrm{H}$, respectively, and where C is a positive constant which is independent of h and $\boldsymbol{v}$. It follows that

$$\mathsf{d}(J_{\boldsymbol{b}}(h_1), J_{\boldsymbol{b}}(h_2)) \leq C\,\mathsf{d}(h_1,h_2), \qquad \forall h_1, h_2 \in \mathrm{H},$$

and this, in turn, implies that

$$\mathsf{d}(\widetilde{J}_{\boldsymbol{b}}(p_1), \widetilde{J}_{\boldsymbol{b}}(p_2)) \leq C\,\mathsf{d}(p_1,p_2), \qquad \forall p_1, p_2 \in (\Gamma\cap\mathrm{H})\backslash\mathrm{H}.$$

Now let $\varepsilon > 0$ be given. Then for any point $p \in \partial_\varepsilon\big(\widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)\big)$ there exist points $p_1, p_2 \in \Gamma\mathrm{H}\backslash\mathrm{H}$ satisfying $p_1 \in \widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)$, $p_2 \notin \widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)$, $\mathsf{d}(p,p_1) < \varepsilon$ and $\mathsf{d}(p,p_2) < \varepsilon$. It follows that $\widetilde{J}_{\boldsymbol{b}}(p_1) \in \widetilde{\Omega}_r$, $\widetilde{J}_{\boldsymbol{b}}(p_2) \notin \widetilde{\Omega}_r$, $\mathsf{d}(\widetilde{J}_{\boldsymbol{b}}(p_1), \widetilde{J}_{\boldsymbol{b}}(p)) < C\,\varepsilon$ and $\mathsf{d}(\widetilde{J}_{\boldsymbol{b}}(p_2), \widetilde{J}_{\boldsymbol{b}}(p)) < C\,\varepsilon$; and hence $\widetilde{J}_{\boldsymbol{b}}(p) \in \partial_{C\,\varepsilon}\widetilde{\Omega}_r$. We have thus proved:

$$\partial_\varepsilon\big(\widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)\big) \subset \widetilde{J}_{\boldsymbol{b}}^{-1}\big(\partial_{C\,\varepsilon}\widetilde{\Omega}_r\big). \tag{8.13}$$

Hence, using also Lemma 8.6 and the fact that $\widetilde{\Omega}_r$ is smooth,

$$\mu_{\mathrm{H}}\big(\partial_\varepsilon\big(\widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)\big)\big) \leq \mu_{\mathrm{H}}\big(\widetilde{J}_{\boldsymbol{b}}^{-1}\big(\partial_{C\,\varepsilon}\widetilde{\Omega}_r\big)\big) = \mu\big(\partial_{C\,\varepsilon}\widetilde{\Omega}_r\big) \ll \varepsilon.$$

Hence $\widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)$ is smooth. □

Remark 8.8. One verifies that the constant C in the proof of the previous lemma can be taken to be $C_1\|\boldsymbol{b}\|$, where $\|\boldsymbol{b}\|$ is the Euclidean norm of $\boldsymbol{b}$ and where C_1 is a constant which only depends on the Riemannian metrics on H and $\mathrm{ASL}_d(\mathbb{R})$.

The following is the analogue of Lemma 8.7 in the case $\boldsymbol{b} = \mathbf{0}$.

Lemma 8.9. *For any smooth subset $\Omega \subset \mathbb{R}^d$ which contains a neighbourhood of the origin, and for any $r \in \mathbb{Z}_{\geq 0}$, $\widetilde{J}_{\mathbf{0}}^{-1}(\widetilde{\Omega}_r)$ is a smooth subset of $\Gamma\backslash\Gamma\mathrm{H}$.*

Proof. Since $\widetilde{J}_{\mathbf{0}}$ maps $\Gamma\backslash\Gamma\mathrm{H}$ into $\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})$, we have $\widetilde{J}_{\mathbf{0}}^{-1}(\widetilde{\Omega}_r) = \widetilde{J}_{\mathbf{0}}^{-1}(\widetilde{\Omega}'_r)$ where

$$\widetilde{\Omega}'_r = \big\{\mathrm{SL}_d(\mathbb{Z})g \,:\, g \in \mathrm{SL}_d(\mathbb{R}),\ \#(\mathbb{Z}^d g \cap \Omega) = r\big\}.$$

This set $\widetilde{\Omega}'_r$ is a smooth subset of $\mathrm{SL}_d(\mathbb{Z})\backslash\mathrm{SL}_d(\mathbb{R})$, by [SV05, Lemma 4]. Now the proof of Lemma 8.7 carries over to the present case. □

Proof of Theorem 8.1. Let $U, \Omega, \boldsymbol{b}, r$ be given as in the statement of the theorem. Let $\chi_1 : \mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z}) \to \{0,1\}$ be the characteristic function of U; let $\chi_2 : \Gamma\backslash\Gamma\mathrm{H} \to \{0,1\}$ be the characteristic function of $\widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)$, and let $\chi : \mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z}) \times \Gamma\backslash\Gamma\mathrm{H} \to \{0,1\}$ be the characteristic function of $U \times \widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)$. Then by Lemma 8.5 and (8.12), for any $q \in \mathbb{Z}^+$, the number of $R \in \mathcal{R}_q \cap qU$ such that the equation $\boldsymbol{x}R \equiv \boldsymbol{b} \bmod q$ has exactly r solutions $\boldsymbol{x} \in \mathbb{Z}^d \cap q^{n/d}\Omega$ is

$$\sum_{R\in\mathcal{R}_q} \chi\big(q^{-1}R, \widetilde{n}_+(q^{-1}R)D(q)\big). \tag{8.14}$$

Since U is smooth by assumption, and $\widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)$ is smooth by Lemma 8.7 or Lemma 8.9, it follows from the proof of [SV05, Lemma 1] that for every $0 < \delta \leq \frac{1}{2}$ there exist functions $f_{1,\pm} \in \mathrm{C}^\infty(\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z}))$ and $f_{2,\pm} \in \mathrm{C}^\infty(\Gamma\backslash\Gamma\mathrm{H})$ satisfying

$$0 \leq f_{j,-} \leq \chi_j \leq f_{j,+} \leq 1 \qquad \text{and} \qquad S_{\infty,k}(f_{j,\pm}) \ll_k \delta^{-k}$$

for $j = 1, 2$ and any real $k \geq 0$, and also

$$\mathrm{vol}\Big(\overline{\{A \,:\, f_{1,\pm}(A) \neq \chi_1(A)\}}\Big) \ll \delta \qquad \text{and} \qquad \mu_{\mathrm{H}}\Big(\overline{\{\Gamma h \,:\, f_{2,\pm}(\Gamma h) \neq \chi_2(\Gamma h)\}}\Big) \ll \delta.$$

Define the two functions $f_\pm \in \mathrm{C}_b^\infty(\mathrm{M}_{d\times n}(\mathbb{R}/\mathbb{Z}) \times \Gamma\backslash\Gamma\mathrm{H})$ through $f_\pm(T,p) = f_{1,\pm}(T)f_{2,\pm}(p)$. Then

$$0 \leq f_- \leq \chi \leq f_+ \leq 1; \qquad S_{\infty,k}(f_\pm) \ll_k \delta^{-k}; \qquad S_{2,k'}(f_\pm) \ll_k \delta^{\frac{1}{2}-k'}$$

for any real $k \geq 0$ and $k' \geq 1$. Recalling the definition (1.7), it follows that the sum in (8.14) is bounded from above by $\#\mathcal{R}_q \cdot \mathcal{A}_q(f_+)$, and by Theorem 1.2 this equals

$$\begin{aligned} &\#\mathcal{R}_q\bigg(\int_{\mathrm{M}_{d\times n}(\mathbb{Z}\backslash\mathbb{R})} f_{1,+}\,dT \int_{\Gamma\backslash\Gamma\mathrm{H}} f_{2,+}\,d\mu_{\mathrm{H}} + O\Big(S_{\infty,\kappa+\varepsilon}(f_+)q^{-\vartheta+\varepsilon} + S_{2,\kappa'+\varepsilon}(f_+)q^{-\vartheta'+\varepsilon}\Big)\bigg) \\ &= \#\mathcal{R}_q\bigg(\mathrm{vol}(U)\mu_{\mathrm{H}}(\widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)) + O(\delta) + O\Big(\delta^{-\kappa-\varepsilon}q^{-\vartheta+\varepsilon} + \delta^{\frac{1}{2}-\kappa'-\varepsilon}q^{-\vartheta'+\varepsilon}\Big)\bigg). \end{aligned} \tag{8.15}$$

Similarly, the sum in (8.14) is bounded from below by $\#\mathcal{R}_q \cdot \mathcal{A}_q(f_-)$, which is again estimated by the right-hand side of (8.15). It follows that also the sum in (8.14) itself is estimated by the right-hand side of (8.15). Note here that by Lemma 8.6, $\mu_{\mathrm{H}}(\widetilde{J}_{\boldsymbol{b}}^{-1}(\widetilde{\Omega}_r)) = c(\Omega, \boldsymbol{b}, r)$, the constant defined in (8.3). We now optimize by choosing $\delta = q^{-\alpha}$ with $\alpha = \alpha(d,n) := \min\big(\frac{\vartheta}{1+\kappa}, \frac{\vartheta'}{\frac{1}{2}+\kappa'}\big)$. Using also $\mathrm{vol}(U) > 0$, it follows that

$$\sum_{R\in\mathcal{R}_q} \chi\big(q^{-1}R, \widetilde{n}_+(q^{-1}R)D(q)\big) = \#\mathcal{R}_q \cdot \mathrm{vol}(U)\big(c(\Omega,\boldsymbol{b},r) + O\big(q^{-\alpha+\varepsilon'}\big)\big), \tag{8.16}$$

where ε' depends on ε, with $\varepsilon' \to 0$ as $\varepsilon \to 0$. From now on we write ε in place of ε'. Let us note that the exponent $\alpha = \alpha(d,n)$ here satisfies the formula stated in Theorem 8.1; this is immediate from the formulas for $\kappa, \vartheta, \kappa', \vartheta'$ in Theorem 1.2, where in the special case $d = 2$, $n = 1$ we make use of the fact that $\theta \le \frac{7}{64} < \frac{3}{10}$.

It remains to prove that we can replace the factor $\#\mathcal{R}_q \cdot \mathrm{vol}(U)$ in (8.16) by $\#(\mathcal{R}_q \cap qU)$. To this end, note first that by repeating the above argument but with $\chi_2 \equiv f_{2,\pm} \equiv 1$, we obtain

$$\#(\mathcal{R}_q \cap qU) = \#\mathcal{R}_q \cdot \mathrm{vol}(U)\big(1 + O\big(q^{-\alpha+\varepsilon}\big)\big),$$

that is, there exists a constant $C = C(U,\varepsilon) > 0$ such that

$$\#\mathcal{R}_q \cdot \mathrm{vol}(U)\big(1 - Cq^{-\alpha+\varepsilon}\big) \le \#(\mathcal{R}_q \cap qU) \le \#\mathcal{R}_q \cdot \mathrm{vol}(U)\big(1 + Cq^{-\alpha+\varepsilon}\big).$$

We have $\big(1 + Cq^{-\alpha+\varepsilon}\big)^{-1} \ge 1 - O(q^{-\alpha+\varepsilon})$ and so $\#\mathcal{R}_q \cdot \mathrm{vol}(U) \ge \#(\mathcal{R}_q \cap qU)\big(1 - O(q^{-\alpha+\varepsilon})\big)$; and if $Cq^{-\alpha+\varepsilon} \le \frac{1}{2}$ then also $\big(1 - Cq^{-\alpha+\varepsilon}\big)^{-1} \le 1 + O(q^{-\alpha+\varepsilon})$, allowing us to conclude $\#\mathcal{R}_q \cdot \mathrm{vol}(U) = \#(\mathcal{R}_q \cap qU) \cdot \big(1 + O(q^{-\alpha+\varepsilon})\big)$. Using the last estimate in (8.16) gives (8.2), and even when $Cq^{-\alpha+\varepsilon} > \frac{1}{2}$ we conclude that (8.2) is a valid bound from below on the quantity in (8.14). Note also that the statement around (8.2) holds trivially if $\#(\mathcal{R}_q \cap qU) = 0$; hence from now on we may assume $\#(\mathcal{R}_q \cap qU) \ge 1$. Now to complete the proof, note that $Cq^{-\alpha+\varepsilon} > \frac{1}{2}$ implies $q \ll 1$, hence $\#\mathcal{R}_q \ll 1$, and so $\#\mathcal{R}_q \cdot \mathrm{vol}(U) \le 1 + O(q^{-\alpha+\varepsilon}) \le \#(\mathcal{R}_q \cap qU) \cdot (1 + O(q^{-\alpha+\varepsilon}))$, provided that we take the implied constant sufficiently large. Using the last inequality in (8.16) gives the desired upper bound. □

Proof of the statement in Remark 8.3. The proof of Theorem 8.1 carries over, with essentially the only difference being that $\chi_2 : \Gamma\backslash\Gamma\mathrm{H} \to \{0,1\}$ is now taken to be the characteristic function of the intersection $\cap_{j=1}^k \widetilde{J}_{\boldsymbol{b}_j}^{-1}(\widetilde{\Omega}_{j,r_j})$, where $\widetilde{\Omega}_{j,r_j} = \big\{\mathrm{ASL}_d(\mathbb{Z})g \ : \ g \in \mathrm{ASL}_d(\mathbb{R}),\ \#(\mathbb{Z}^d g \cap \Omega_j) = r_j\big\}$. Now by a property valid in arbitrary metric spaces, $\partial\big(\cap_{j=1}^k \widetilde{J}_{\boldsymbol{b}_j}^{-1}(\widetilde{\Omega}_{j,r_j})\big) \subset \cup_{j=1}^k \partial\big(\widetilde{J}_{\boldsymbol{b}_j}^{-1}(\widetilde{\Omega}_{j,r_j})\big)$, and hence $\partial_\varepsilon\big(\cap_{j=1}^k \widetilde{J}_{\boldsymbol{b}_j}^{-1}(\widetilde{\Omega}_{j,r_j})\big) \subset \cup_{j=1}^k \partial_\varepsilon\big(\widetilde{J}_{\boldsymbol{b}_j}^{-1}(\widetilde{\Omega}_{j,r_j})\big)$ for every ε. Hence, using the fact that each set $\widetilde{J}_{\boldsymbol{b}_j}^{-1}(\widetilde{\Omega}_{j,r_j})$ is smooth, it follows that also the intersection $\cap_{j=1}^k \widetilde{J}_{\boldsymbol{b}_j}^{-1}(\widetilde{\Omega}_{j,r_j})$ is smooth. The rest of the proof is essentially the same as before, and we obtain (8.4) with

$$c = \mu_H\Big(\bigcap_{j=1}^k \widetilde{J}_{\boldsymbol{b}_j}^{-1}(\widetilde{\Omega}_{j,r_j})\Big). \tag{8.17}$$

□

Proof of the statement in Remark 8.4. Recall that when $d = n$, the map $\widetilde{n}_-$ (see (2.7)) gives an identification between $\mathrm{M}_n(\mathbb{R}/\mathbb{Z})$ and $\Gamma\backslash\Gamma\mathrm{H}$. Let $m_{\boldsymbol{b}}$ be the 'multiplication' map from $\Gamma\backslash\Gamma\mathrm{H}$ to $(\mathbb{R}/\mathbb{Z})^n$ given by $m_{\boldsymbol{b}}(\widetilde{n}_-(A)) = \boldsymbol{b}A$ for all $A \in \mathrm{M}_n(\mathbb{R}/\mathbb{Z})$. Let $\chi_1 : \mathrm{M}_n(\mathbb{R}/\mathbb{Z}) \to \{0,1\}$ be the characteristic function of U; let $\chi_2 : \Gamma\backslash\Gamma\mathrm{H} \to \{0,1\}$ be the characteristic function of $m_{\boldsymbol{b}}^{-1}(\Omega)$, and let $\chi : \mathrm{M}_n(\mathbb{R}/\mathbb{Z}) \times \Gamma\backslash\Gamma\mathrm{H} \to \{0,1\}$ be the characteristic function of $U \times m_{\boldsymbol{b}}^{-1}(\Omega)$. Recall that $\widetilde{n}_+(q^{-1}R)D(q) = \widetilde{n}_-(q^{-1}R^{-1})$ for every $R \in \mathcal{R}_q$, by Lemma 2.3. Hence the number of $R \in \mathcal{R}_q \cap qU$ such that $\boldsymbol{x} = \boldsymbol{b}R^{-1}$ lies in $q\Omega$ is now again given by the sum in (8.14). One verifies that $\mu_{\mathrm{H}} \circ m_{\boldsymbol{b}}^{-1} = \mathrm{vol}$, the Lebesgue measure on $(\mathbb{R}/\mathbb{Z})^n$, and by an argument as in Lemma 8.7, $m_{\boldsymbol{b}}^{-1}(\Omega)$ is a smooth subset of $\Gamma\backslash\Gamma\mathrm{H}$. Now the proof of Theorem 8.1 carries over to the present case. □

8.2. **By-product: counting matrices.** The next theorem gives an optimal bound on the following quantity, for any given $1 \le r < n < d$, any prime p and integer $1 \le b \le \frac{p-1}{2}$:

$$N_{p,b} := \#\{X \in \mathrm{M}_{d\times n}(\mathbb{Z}) : \|X\|_\infty \le b \text{ and } \mathrm{rank}(X \bmod p) = r\}. \tag{8.18}$$

The proof of this bound is a by-product of the proof of our main result, Theorem 1.2; in particular it uses an interpretation in terms of Hecke operators, and Rogers' formula (Theorem 6.3).

Theorem 8.10. *Let $1 \leq r < n < d$. For every prime p and integer $1 \leq b \leq \frac{p-1}{2}$,*

$$N_{p,b} \asymp_d \max\left(b^{dr}, b^{dn} p^{-(d-r)(n-r)}\right). \tag{8.19}$$

Remark 8.11. The same counting problem was considered by Ahmadi and Shparlinski in [AS07, Theorem 9]. They were however interested in obtaining asymptotics, which they did through results ultimately relying on Deligne-type methods for estimating the number of $\mathbb{F}_p$-points on varieties. Their large p asymptotics are non-trivial in the range where b is large, specifically – with our notation – whenever $b \geq p^{\gamma_{r,n,d}+\varepsilon}$ for some positive ε, where

$$\gamma_{r,n,d} = \max\left(\frac{1}{2} + \frac{(n-r)(d-r)}{2nd}, 1 - \frac{1}{2(n-r)(d-r)+2}\right).$$

It should be noted that their result is valid for arbitrary $n, d \geq 1$.

Our method yields an upper bound of the correct order of magnitude for arbitrary p and b; however we are not able to handle the case of square matrices ($n = d$). This stems from the application of Rogers' formula in our approach; Theorem 6.3 is only valid under the assumption $n < d$.

Proof. The main work will be spent on proving that (8.19) gives a valid bound from above on $N_{p,b}$. To start, let $\pi_p \colon \mathbb{Z}^d \twoheadrightarrow \mathbb{F}_p^d$ be the canonical projection; denote by $\mathrm{Gr}_{r,d}(\mathbb{F}_p)$ the space of r-dimensional linear subspaces of $\mathbb{F}_p^d$, and by X_d the space $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$. Observe that for a linear subspace V of $\mathbb{F}_p^d$ with dimension r, $\pi_p^{-1}(V)$ is a sublattice of $\mathbb{Z}^d$ whose covolume is p^{d-r}, hence $p^{\frac{r-d}{d}}\pi_p^{-1}(V)$ is a unimodular lattice in $\mathbb{R}^d$. We may thus introduce the map

$$\phi_p \colon \mathrm{Gr}_{r,d}(\mathbb{F}_p) \to X_d, \qquad \phi_p(V) = p^{\frac{r}{d}-1}\pi_p^{-1}(V).$$

It follows that an upper bound for $N_{p,b}$ is given by

$$\begin{aligned}
&\sum_{V \in \mathrm{Gr}_{r,d}(\mathbb{F}_p)} \#\{(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \in (\pi_p^{-1}(V) \cap [-b,b]^d)^n : \dim \mathrm{Span}_{\mathbb{R}}(\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}) \geq r\}) \\
&= \sum_{V \in \mathrm{Gr}_{r,d}(\mathbb{F}_p)} \#\{(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \in (\phi_p(V) \cap b\,\tilde{C})^n : \dim \mathrm{Span}_{\mathbb{R}}(\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}) \geq r\}),
\end{aligned}$$

where $\tilde{C}$ is the cube $\tilde{C} = p^{\frac{r}{d}-1}[-1,1]^d \subset \mathbb{R}^d$.

At this point we recall the connection between the Grassmannian over $\mathbb{F}_p$ and lattices, namely that there is a bijection between $\mathrm{Gr}_{r,d}(\mathbb{F}_p)$ and the lattices $p\mathbb{Z}^d \subset L \subset \mathbb{Z}^d$ of index p^{d-r}. Furthermore, the family of such lattices can be used to define a Hecke operator: Set

$$D'_p = p^{\frac{r}{d}-1}\begin{pmatrix} I_r & \\ & pI_{d-r} \end{pmatrix} \in \mathrm{SL}_d(\mathbb{R}),$$

and introduce, as in Section 4, the Hecke operator $T_{D'_p}$, acting on functions on X_d. It then follows from [Shi94, Lemma 3.13] that

$$(T_{D'_p}\Phi)(L) = \frac{1}{\#\,\mathrm{Gr}_{r,d}(\mathbb{F}_p)} \sum_{\substack{pL \subset L' \subset L \\ [L:L'] = p^{d-r}}} \Phi\left(p^{\frac{r}{d}-1}L'\right),$$

for any $\Phi : X_d \to \mathbb{C}$.

Hence:

$$N_{p,b} \leq \#\,\mathrm{Gr}_{r,d}(\mathbb{F}_p)\big[T_{D'_p}(F_p)\big](\mathbb{Z}^d),$$

where $F_p\colon X_d \to \mathbb{Z}_{\geq 0}$ is defined, for $L \in X_d$, by

$$F_p(L) = \#\{(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \in (L \cap b\,\tilde{C})^n : \dim \operatorname{Span}_{\mathbb{R}}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \geq r\}.$$

We now proceed as in the proof of our main theorem, specifically the part after (7.42), whose role is now played by the above inequality.

As in that proof, let Ω be an open neighbourhood of the identity matrix I_d in $\mathrm{SL}_d(\mathbb{R})$ such that (7.43) holds for all $w \in \Omega$ and $\boldsymbol{v} \in \mathbb{R}^d$; it then follows that

$$\forall w \in \Omega : \qquad N_{p,b} \leq \#\operatorname{Gr}_{r,d}(\mathbb{F}_p)\big[T_{D'_p}(\tilde{F}_p)\big](\mathbb{Z}^d w),$$

where

$$\begin{aligned}\tilde{F}_p\colon X_d &\to \mathbb{Z}_{\geq 0}\\ L &\mapsto \#\{(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \in (L \cap 2b\,\tilde{C})^n : \dim \operatorname{Span}_{\mathbb{R}}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\} \geq r\}.\end{aligned}$$

Hence

$$N_{p,b} \ll_d \#\operatorname{Gr}_{r,d}(\mathbb{F}_p) \int_{X_d} [T_{D'_p}(\tilde{F}_p)](L)\, d\mu_0(L) = \#\operatorname{Gr}_{r,d}(\mathbb{F}_p) \int_{X_d} \tilde{F}_p\, d\mu_0. \tag{8.20}$$

(Recall that μ_0 is the $\mathrm{SL}_d(\mathbb{R})$-invariant probability measure on X_d.) The last upper bound can be rewritten as

$$N_{p,b} \ll_d \#\operatorname{Gr}_{r,d}(\mathbb{F}_p) \int_{X_d} \sum_{(\boldsymbol{v}_1,\dots,\boldsymbol{v}_n) \in L^n} \chi_{p,b}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)\, d\mu_0(L), \tag{8.21}$$

where $\chi_{p,b}\colon (\mathbb{R}^d)^n \to \{0,1\}$ is the characteristic function of the set $\{(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \in (2b\,\tilde{C})^n : \dim \operatorname{Span}_{\mathbb{R}}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\} \geq r\}$. We now rewrite the integrand in such a way that we can apply Theorem 6.3: writing $L = \mathbb{Z}^d g$ for some $g \in \mathrm{SL}_d(\mathbb{R})$, it is

$$\sum_{(m_1,\dots,m_n) \in (\mathbb{Z}^d)^n} \chi_{p,b}(m_1 g, \dots, m_n g) = \sum_{X \in \mathrm{M}_{n\times d}(\mathbb{Z})} \chi_{p,b}(Xg) = \sum_{\substack{X \in \mathrm{M}_{n\times d}(\mathbb{Z})\\ X \neq \mathbf{0}}} \chi_{p,b}(Xg).$$

We are now in a position to use Theorem 6.3 and deduce that the integral in (8.21) is equal to

$$\sum_{m=1}^{n} \sum_{B \in A_{n,m}} \int_{\mathrm{M}_{m\times d}(\mathbb{R})} \chi_{p,b}(BX)\, dX,$$

where we recall that the matrix B has rank m, and so BX has rank $\leq m$. By the definition of $\chi_{p,b}$, it follows that the integrand vanishes whenever $m \leq r-1$, so the sum is equal to

$$\sum_{m=r}^{n} \sum_{B \in A_{n,m}} \int_{\mathrm{M}_{m\times d}(\mathbb{R})} \chi_{p,b}(BX)\, dX.$$

If we now define $\chi\colon \mathrm{M}_{n\times d}(\mathbb{R}) \to \{0,1\}$ to be the characteristic function of the set of matrices $M \in \mathrm{M}_{n\times d}(\mathbb{R})$ such that $\|M\|_\infty \leq 1$ and $\operatorname{rank} M \geq r$, we have, using $2b\,\tilde{C} = 2bp^{\frac{r}{d}-1}[-1,1]^d$:

$$\begin{aligned}\sum_{B \in A_{n,m}} \int_{\mathrm{M}_{m\times d}(\mathbb{R})} \chi_{p,b}(BX)\, dX &= \sum_{B \in A_{n,m}} \int_{\mathrm{M}_{m\times d}(\mathbb{R})} \chi((2b)^{-1}p^{1-\frac{r}{d}}BX)\, dX\\ &= (2bp^{\frac{r}{d}-1})^{md} \sum_{B \in A_{n,m}} \int_{\mathrm{M}_{m\times d}(\mathbb{R})} \chi(BX)\, dX \ll_d b^{md}p^{-m(d-r)}.\end{aligned}$$

In the last step we used the fact that $\sum_{B\in A_{n,m}} \int_{\mathrm{M}_{m\times d}(\mathbb{R})} \chi(BX)\,dX \le \int_{X_d} \sum_{X\in \mathrm{M}_{n\times d}(\mathbb{Z})} \chi(Xg)\,d\mu_0(g) < \infty$, by Theorem 6.3 and [Sch58, Theorem 2]. Plugging the last bound back into (8.21), we finally obtain

$$N_{p,b} \ll_d \#\,\mathrm{Gr}_{r,d}(\mathbb{F}_p) \sum_{m=r}^{n} (b^d p^{-(d-r)})^m \ll_d p^{r(d-r)} \max\big((b^d p^{-(d-r)})^r, (b^d p^{-(d-r)})^n\big)$$
$$= \max\big(b^{dr}, b^{dn} p^{-(d-r)(n-r)}\big),$$

i.e. we have proved that (8.19) gives a valid upper bound on $N_{p,b}$.

To finish, we prove that the same expression is also a lower bound on $N_{p,b}$; it should be noted that this proof is completely elementary. As a first step we note that

$$\#\{Y \in \mathrm{M}_r(\mathbb{Z}) \,:\, \|Y\|_\infty \le b \text{ and } \det Y \not\equiv 0 \bmod p\} > b^{r^2}. \tag{8.22}$$

This is proved by induction: First, by immediate inspection (using $b \ge 1$), we have

$$\#\{y \in \mathbb{Z} \cap [-b,b] \,:\, y \not\equiv a \bmod p\} > b \qquad (\forall a \in \mathbb{Z}). \tag{8.23}$$

This fact, applied with $a = 0$, means that (8.22) holds for $r = 1$. Next, for $r \ge 2$, write $Y = (y_{ij}) \in \mathrm{M}_r(\mathbb{Z})$, and let Y' be the top left $(r-1)\times(r-1)$ submatrix of Y. Then by expanding $\det Y$ along the bottom row, we have $\det Y = y_{r,r} \cdot \det Y' + h$, where h is an integer which is independent of $y_{r,r}$. Hence for any fixed choice of Y' with $\|Y'\|_\infty \le b$ and $\det Y' \not\equiv 0 \bmod p$, and any fixed choice of the entries $y_{r,i}$ and $y_{i,r}$ $(i = 1,\dots,r-1)$, there is some $a \in \mathbb{Z}$ such that $\det Y \equiv 0 \bmod p$ holds if and only if $y_{r,r} \equiv a \bmod p$; and so by (8.23) there are more than b choices of $y_{r,r} \in \mathbb{Z}\cap[-b,b]$ which make $\det Y \not\equiv 0 \bmod p$. Since the number of choices of Y' as above is $\ge b^{(r-1)^2}$ (by induction), and each entry $y_{r,i}$ and $y_{i,r}$ $(i = 1,\dots,r-1)$ can be chosen in more than b ways, it follows that (8.22) holds.

Note that any matrix $X \in \mathrm{M}_{d\times n}(\mathbb{Z})$ with $\|X\|_\infty \le b$ whose top left $r\times r$ submatrix has determinant $\not\equiv 0 \bmod p$ and whose last $n-r$ columns vanish identically, belongs to the set in (8.18). Hence (8.22) immediately implies that

$$N_{p,b} \ge b^{r^2} \cdot b^{r(d-r)} = b^{dr}. \tag{8.24}$$

Next we will prove that we also have $N_{p,b} \gg b^{dn} p^{-(d-r)(n-r)}$. Let

$$B' := (\mathbb{Z} \cap [-\tfrac{1}{2}b, \tfrac{1}{2}b])^d \quad \text{and} \quad B := (\mathbb{Z}\cap[b,b])^d,$$

and note that $\#(\mathbb{Z}\cap[-\frac{1}{2}b,\frac{1}{2}b]) > \frac{1}{2}b$ and hence $\#B' > (\frac{1}{2}b)^d$. We claim that for every vector subspace $V \subset \mathbb{F}_p^d$ of dimension r,

$$\#(B \cap \pi_p^{-1}(V)) > (\tfrac{1}{2}b)^d p^{r-d}. \tag{8.25}$$

To prove this, set $H := \max\{\#\alpha^{-1}(\boldsymbol{w}) \,:\, \boldsymbol{w} \in \mathbb{F}_p^d/V\}$, where α is the projection map from B' to $\mathbb{F}_p^d/V$. Then $(\frac{1}{2}b)^d < \#B' \le \#(\mathbb{F}_p^d/V)\cdot H = p^{d-r}H$, and so $H > (\frac{1}{2}b)^d p^{r-d}$. But the definition of H implies that there exist H distinct vectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_H$ in B' satisfying $\pi_p(\boldsymbol{v}_1) \equiv \cdots \equiv \pi_p(\boldsymbol{v}_H) \bmod V$. It follows that $\boldsymbol{v}_1 - \boldsymbol{v}_i$ for $i = 1,2,\dots,H$ are H distinct vectors lying in $B \cap \pi_p^{-1}(V)$, and hence $\#(B\cap\pi_p^{-1}(V)) \ge H > (\frac{1}{2}b)^d p^{r-d}$, i.e. (8.25) is proved.

Now let us construct matrices X belonging to the set in (8.18) as follows: First choose the left $d\times r$ submatrix X' of X to have all entries in $\mathbb{Z}\cap[-b,b]$ and full rank mod p. By the argument giving (8.24), this choice can be made in $\ge b^{dr}$ ways. Let $V \subset \mathbb{F}_p^d$ be the span of the columns of X' reduced mod p. Finally, pick each remaining column of X as an arbitrary vector in $B \cap \pi_p^{-1}(V)$.

By (8.25), these columns can be chosen in more than $\big((\frac{1}{2}b)^d p^{r-d}\big)^{n-r}$ ways, and our construction guarantees that X belongs to the set in (8.18). Hence

$$N_{p,b} > b^{dr} \cdot \big((\tfrac{1}{2}b)^d p^{r-d}\big)^{n-r} \tag{8.26}$$

Together, (8.24) and (8.26) imply the desired lower bound, $N_{p,b} \gg_d \max\big(b^{dr}, b^{dn}p^{-(d-r)(n-r)}\big)$ (with the implied constant being $2^{-d(n-r)}$). □

References

[AS07] Omran Ahmadi and Igor E. Shparlinski, *Distribution of matrices with restricted entries over finite fields*, Indag. Math. (N.S.) **18** (2007), no. 3, 327–337.

[Bhb] Alex B. (https://mathoverflow.net/users/35416/alex b), *Roots of permutations*, MathOverflow, URL:https://mathoverflow.net/q/41788 (version: 2019-03-12).

[COU01] Laurent Clozel, Hee Oh, and Emmanuel Ullmo, *Hecke operators and equidistribution of Hecke points*, Invent. Math. **144** (2001), no. 2, 327–351. MR 1827734

[EBHL22] Daniel El-Baz, Bingrong Huang, and Min Lee, *Effective equidistribution of primitive rational points on expanding horospheres*, J. Eur. Math. Soc. (2022), DOI 10.4171/JEMS/1238.

[EMSS16] Manfred Einsiedler, Shahar Mozes, Nimish Shah, and Uri Shapira, *Equidistribution of primitive rational points on expanding horospheres*, Compositio Mathematica **152** (2016), no. 4, 667–692.

[ET21] Márton Erdélyi and Árpád Tóth, *Matrix Kloosterman sums*, arXiv:2109.00762, 2021.

[ETZ22] Márton Erdélyi, Árpád Tóth, and Gergely Zábrádi, *Matrix Kloosterman sums modulo prime powers*, ArXiv (2022), 1–17.

[GM03] Daniel Goldstein and Andrew Mayer, *On the equidistribution of Hecke points*, Forum Math. **15** (2003), no. 2, 165–189.

[Gra08] Loukas Grafakos, *Classical Fourier analysis*, second ed., Graduate Texts in Mathematics, vol. 249, Springer, New York, 2008.

[GS91] Henri Gillet and Christophe Soulé, *On the number of lattice points in convex symmetric bodies and their duals*, Israel J. Math. **74** (1991), no. 2-3, 347–357. MR 1135244; erratum, ibid. **171** (2009), 443–444

[Kim03] Henry H. Kim, *Functoriality for the exterior square of* GL_4 *and the symmetric fourth of* GL_2, J. Amer. Math. Soc. **16** (2003), no. 1, 139–183, With appendix 1 by Dinakar Ramakrishnan and appendix 2 by Kim and Peter Sarnak. MR 1937203

[Li15] Han Li, *Effective limit distribution of the Frobenius numbers*, Compositio Mathematica **151** (2015), no. 5, 898–916.

[LM17] Min Lee and Jens Marklof, *Effective equidistribution of rational points on expanding horospheres*, International Mathematics Research Notices (2017), rnx081.

[Mar10a] Jens Marklof, *The asymptotic distribution of Frobenius numbers*, Inventiones mathematicae **181** (2010), no. 1, 179–207.

[Mar10b] ———, *Horospheres, Farey fractions and Frobenius numbers*, Oberwolfach Reports **29** (2010), 28–32.

[MS13] Jens Marklof and Andreas Strömbergsson, *Diameters of random circulant graphs*, Combinatorica **33** (2013), no. 4, 429–466.

[Oh02] Hee Oh, *Uniform pointwise bounds for matrix coefficients of unitary representations and applications to kazhdan constants*, Duke Math. J. **113** (2002), no. 1, 133–192.

[Rat91] Marina Ratner, *On Raghunathan's measure conjecture*, Ann. of Math. (2) **134** (1991), no. 3, 545–607. MR 1135878

[Rog55] C. Ambrose Rogers, *Mean values over the space of lattices*, Acta Math. **94** (1955), 249–287.

[Sch58] Wolfgang Schmidt, *On the convergence of mean values over lattices*, Canad. J. Math. **10** (1958), 103–110.

[Ser77] Jean-Pierre Serre, *Linear representations of finite groups*, Springer-Verlag, New York, 1977, Translated from the second French edition by Leonard L. Scott, Graduate Texts in Mathematics, Vol. 42.

[Sha98] Nimish A. Shah, *Invariant measures and orbit closures on homogeneous spaces for actions of subgroups generated by unipotent elements*, Lie groups and ergodic theory (Mumbai, 1996), Tata Inst. Fund. Res. Stud. Math., vol. 14, Tata Inst. Fund. Res., Bombay, 1998, pp. 229–271. MR 1699367

[Shi94] Goro Shimura, *Introduction to the arithmetic theory of automorphic functions*, Publications of the Mathematical Society of Japan, vol. 11, Princeton University Press, Princeton, NJ, 1994, Reprint of the 1971 original, Kanô Memorial Lectures, 1. MR 1291394

[Shp15] Igor E. Shparlinski, *Points on varieties over finite fields in small boxes*, SCHOLAR—a scientific celebration highlighting open lines of arithmetic research, Contemp. Math., vol. 655, Amer. Math. Soc., Providence, RI, 2015, pp. 209–233. MR 3453122

[SS22] Andreas Strömbergsson and Anders Södergren, *On a mean value formula for multiple sums over a lattice and its dual*, 2022, preprint, arXiv:2211.05454 [math.NT].

[Str15] Andreas Strömbergsson, *An effective Ratner equidistribution result for* $\mathrm{SL}(2,\mathbb{R}) \ltimes \mathbb{R}^2$, Duke Math. J. **164** (2015), no. 5, 843–902.

[SV05] Andreas Strömbergsson and Akshay Venkatesh, *Small solutions to linear congruences and Hecke equidistribution*, Acta Arithmetica **118** (2005), no. 1, 41–78 (eng).

[Ven10] Akshay Venkatesh, *Sparse equidistribution problems, period bounds and subconvexity*, Ann. of Math. (2) **172** (2010), no. 2, 989–1094.

[Zel81] Andrey V. Zelevinsky, *Representations of finite classical groups*, Lecture Notes in Mathematics, vol. 869, Springer-Verlag, Berlin-New York, 1981, A Hopf algebra approach. MR 643482

Institute of Analysis and Number Theory, TU Graz, Steyrergasse 30, 8010 Graz, Austria
Email address: daniel.elbaz.88@gmail.com

School of Mathematics, University of Bristol, Bristol BS8 1TW, U.K.
Email address: min.lee@bristol.ac.uk

Department of Mathematics, Uppsala University, Box 480, SE-75106, Uppsala, Sweden
Email address: astrombe@math.uu.se