Matrix Kloosterman sums

Márton Erdélyi and Árpád Tóth

Matrix Kloosterman sums

Márton Erdélyi and Árpád Tóth

We study a family of exponential sums that arises in the study of expanding horospheres on GL_n. We prove an explicit version of general purity and find optimal bounds for these sums.

1. Introduction

1A. *The subject of the paper.* We derive nontrivial bounds for certain exponential sums that are natural generalizations of the classical Kloosterman sum to the noncommutative algebra $M_n(\mathbb{F}_q)$ of $n \times n$ matrices over a finite field $\mathbb{F}_q$ with $q = p^f$ elements.

To define these sums let $\mathbb{F}_p$ be the prime field of $\mathbb{F}_q$, and $\overline{\mathbb{F}}$ be an algebraic closure of $\mathbb{F}_q$ so that for $m \geq 1$ $\mathbb{F}_{q^m} \subset \overline{\mathbb{F}}$ is the unique degree-m extension of $\mathbb{F}_q$. Let

$$\varphi_0 : \mathbb{F}_p \to \mathbb{C}^*$$

be the additive character which maps $1 \in \mathbb{F}_p$ to $\zeta = \exp(1/p) = e^{2\pi i/p}$, and fix the additive characters

$$\varphi = \varphi_0 \circ \mathrm{Tr}_{\mathbb{F}_q/\mathbb{F}_p} \quad \text{and} \quad \varphi_m = \varphi_0 \circ \mathrm{Tr}_{\mathbb{F}_{q^m}/\mathbb{F}_p}$$

of $\mathbb{F}_q$ and $\mathbb{F}_{q^m}$.

Let $M_n(\mathbb{F}_{q^m})$ be the algebra of $n \times n$ matrices over $\mathbb{F}_{q^m}$, and $\mathrm{GL}_n(\mathbb{F}_{q^m}) = M_n^*(\mathbb{F}_{q^m}) \subset M_n(\mathbb{F}_{q^m})$ be the general linear group. Let ψ (resp. ψ_m) be the additive character of $M_n(\mathbb{F}_q)$ (resp. $M_n(\mathbb{F}_{q^m})$) defined by

$$\psi = \varphi \circ \mathrm{tr}$$

(resp. $\psi_m = \varphi_m \circ \mathrm{tr}$), where $\mathrm{tr} = \mathrm{tr}_n : M_n(\mathbb{F}_{q^m}) \to \mathbb{F}_{q^m}$ is the matrix trace. For $a \in M_n(\mathbb{F}_{q^m})$ define the sum

$$K_n(a, \mathbb{F}_{q^m}) = \sum_{x \in \mathrm{GL}_n(\mathbb{F}_{q^m})} \psi_m(ax + x^{-1}), \tag{1}$$

generalizing the classical Kloosterman sum [1927]

$$K_1(\alpha, \mathbb{F}_{q^m}) = K(\alpha, \mathbb{F}_{q^m}^*) = \sum_{x \in \mathbb{F}_{q^m}^*} \varphi_m(\alpha x + x^{-1}). \tag{2}$$

When the field in question is clear, or when the arguments used do not depend on it, we will write M_n and $K_n(a)$ for $M_n(\mathbb{F}_q)$ and $K_n(a, \mathbb{F}_q)$.

The interest in these sums arose in connection with a conjecture of Marklof concerning the equidistribution of certain special points associated to expanding horospheres. This conjecture, originally motivated by Marklof's work on Frobenius numbers [2010], was proved by Einsiedler, Moses, Shah and Shapira [2016] using methods of ergodic theory. In the case of SL_2 the connection to classical Kloosterman sums was known already to Marklof (see Section 2.1 of [Einsiedler et al. 2016]) and together with Lee [2018] they proved an effective version of the conjecture for SL_3. This proof strongly hinted that nontrivial bounds of the sums in (1) could yield a proof of Marklof's conjecture with an effective power saving for higher rank situations as well. One of the main purposes of this paper is to provide such bounds; they are formulated in Theorems 1.8 and 1.10. These bounds, together with further extensions in [Erdélyi et al. 2024b], were then recently used by El-Baz, Lee and Strömbergsson for realizing the above goal in [El-Baz et al. 2022].

There is however also intrinsic interest in these sums as natural generalizations[1] of Kloosterman's sum. The relevance and widespread use in analytic number theory of $K_1(\alpha)$ (see, for example, [Heath-Brown 2000]) is immediate from the fact that it is the additive Fourier transform of the function $x \mapsto \varphi(x^{-1})$ on $\mathbb{F}_q^*$ (extended by 0),

$$\varphi(x^{-1}) = \frac{1}{q} \sum_{\alpha \in \mathbb{F}_q} K_1(-\alpha)\varphi(\alpha x),$$

and that suitable estimates are available for $K_1(\alpha)$. In fact one knows [Weil 1948a] (see also [Carlitz 1969]) that if α is not 0, then the associated zeta-function is rational,

$$Z(T) = \exp\left(-\sum_{m \geqslant 1} \frac{K_1(\alpha, \mathbb{F}_{q^m}^*)}{m} T^m\right) = \frac{1}{1 - K_1(\alpha)T + qT^2},$$

from which

$$K_1(\alpha, \mathbb{F}_{q^m}^*) = -(\lambda_1^m + \lambda_2^m)$$

---

[1] The special case when a is a scalar matrix was first considered by Hodges [1956] and reappeared again in various other contexts. See, for example, [Kim 1998; Fulman 2001; Chae and Kim 2003]. We thank Ofir Gorodetsky for bringing our attention to these earlier works.

for some $\lambda_1, \lambda_2 \in \mathbb{C}$. Weil's proof [1948b] of the Riemann hypothesis over function fields shows, using [1948a], that $|\lambda_i| = \sqrt{q}$ which gives the optimal bound

$$|K_1(\alpha, \mathbb{F}_{q^m}^*)| \leq 2q^{m/2}$$

for α not 0. (The explicit description of the connection between exponential sums of this type and the Riemann hypothesis for curves over function fields goes back to [Hasse 1935].)

There are a number of extensions of these results in the commutative setting especially the so called hyper-Kloosterman sums [Mordell 1963; SGA 4½ 1977; Luo et al. 1995; Kowalski et al. 2017], and both these and the classical Kloosterman sums are ubiquitous in the theory of exponential sums [Katz 1988]. There is also a deep connection between Kloosterman sums and modular forms [Poincaré 1911; Petersson 1932; Linnik 1963; Selberg 1965; Deshouillers and Iwaniec 1982; Goldfeld and Sarnak 1983], and the notion of Kloosterman sum is extended to GL_n [Friedberg 1987; Stevens 1987], as well as to other algebraic groups [Dabrowski 1993; Dabrowski and Reeder 1998], with many applications.

The sums $K_n(a)$ considered in this paper are more natural from a ring-theoretic point of view. If A is a finite-dimensional algebra over a finite field, then by the Wedderburn–Artin theorem the additive Fourier transform of $\psi(x^{-1})$ (extended from A^* to A by 0) leads naturally to the matrix Kloosterman sums of (1). These of course are also related to the group GL_n but at the same time they are very strongly tied to the standard representation of this group. From this ring-theoretic point of view we have again

$$\psi(x^{-1}) = \frac{1}{q^{n^2}} \sum_{a \in M_n(\mathbb{F}_q)} K_n(-a)\psi(ax)$$

in the simple ring of $n \times n$ matrices over a finite field.

The other main goal of the paper is then to generalize the classical results above from the Kloosterman sums K_1 to K_n, especially to understand the associated cohomology. The difficulty of this task stems from the fact that when $K_n(a)$ is viewed as an exponential sum on the affine variety

$$V = \{(x, \Delta) \in M_n(\mathbb{F}_q) \times \mathbb{F}_q : \det(x)\Delta = 1\},$$

the part at infinity of the projective closure of V, defined by the equation $\det x = 0$, is singular. However the sums $K_n(a)$ provide a rare example for which their cohomology and so their zeta function is explicitly expressible in terms of one-dimensional exponential sums and so the weights in the sense of Deligne [1980] can be understood in elementary terms. This realization that exponential sums on algebraic groups can be treated rather explicitly is the other main achievement of the paper. The evaluations for the matrix Kloosterman sums in concrete terms, especially the semisimple case in Theorem 1.1, is reminiscent of Herz's work on Bessel functions of matrix arguments [1955]. This link to transcendental special functions continues a long line of similar connections, for example, those of the Gauss, Jacobi, and Kloosterman sums to the gamma, beta and Bessel functions. As an important by-product the concrete nature of these evaluations lead automatically to the estimates required for the equidistribution problem mentioned above.

1B. *Statements of the results.* The statements below refer to a fixed finite field $\mathbb{F}_q$, and so we will write $K_n(a)$ for the sum $K_n(a, \mathbb{F}_q)$ in (1). We start with the following reduction. Let a_1, a_2 be matrices of size $n_1 \times n_1$ and $n_2 \times n_2$, and let $a_1 \oplus a_2$ denote the block matrix $\begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}$ of size $(n_1 + n_2) \times (n_1 + n_2)$.

Theorem 1.1. (i) *Assume that $a_1 \in M_{n_1}(\mathbb{F}_q)$, $a_2 \in M_{n_2}(\mathbb{F}_q)$ and that their characteristic polynomials $p_{a_i}(\lambda)$ are relatively prime. Then*

$$K_{n_1+n_2}(a_1 \oplus a_2) = q^{n_1 n_2} K_{n_1}(a_1) K_{n_2}(a_2).$$

(ii) *Assume that $a \in M_n(\mathbb{F}_q)$ has characteristic polynomial $p_a(\lambda) = \prod_{i=1}^n (\lambda - \alpha_i)$, with $\alpha_i \in \mathbb{F}_q$ all different. Then*

$$K_n(a) = q^{n(n-1)/2} \prod_{i=1}^n K_1(\alpha_i),$$

where $K_1(\alpha_i)$ is as in (2).

Now assume that all roots of the characteristic polynomial of a are in $\mathbb{F}_q$. By the theorem above we may assume that a has a unique eigenvalue α. Our first result in this direction is for nilpotent matrices.

Theorem 1.2. *Assume that $a \in M_n(\mathbb{F}_q)$ is nilpotent. Then*

$$K_n(a) = K_n(0) = (-1)^n q^{n(n-1)/2}. \tag{3}$$

In general we have:

Theorem 1.3. *Assume that $a \in \mathrm{GL}_n(\mathbb{F}_q)$ has a unique eigenvalue $\alpha \neq 0$. Denote by λ the partition of n consisting of the sizes of the blocks in the Jordan normal form of a. There exists a polynomial $P_\lambda(A, G, K) \in \mathbb{Z}[A, G, K]$, that depends only on the block partition λ, such that*

$$K_n(a) = P_\lambda(q, q-1, K_1(\alpha)).$$

Remark 1. While irrelevant for estimates, the separation of q and $q - 1$ in the above polynomial is natural from the cohomological point of view, as these correspond to sums over the additive group $\mathbb{A}^1$ and the multiplicative group $\mathbb{G}_m$.

The proof of Theorem 1.3 is constructive and allows one to express the Kloosterman sums $K_n(a)$ recursively as a polynomial in q, $q - 1$ and $K_1(\alpha_i)$, where α_i runs through the eigenvalues of a. For example, if I_n denotes the identity matrix of size $n \times n$ then we have:

Theorem 1.4. *Assume that $a = \alpha I_n$, $\alpha \neq 0$. Then*

$$K_n(\alpha I_n) = q^{n-1} K_1(\alpha) K_{n-1}(\alpha I_{n-1}) + q^{2n-2}(q^{n-1} - 1) K_{n-2}(\alpha I_{n-2}). \tag{4}$$

The recursion formulas for a general partition are somewhat complicated to state but easy to describe algorithmically. See Section 5B, which also contains further examples. These formulas are based on a parabolic Bruhat decomposition (Section 2B). Using the finer decomposition via a Borel subgroup, one can derive closed form expressions. For example, we have:

Theorem 1.5. *If $\alpha \in \mathbb{F}_q^*$ then*

$$K_n(\alpha I_n) = \sum_{\substack{w \in S_n \\ w^2 = I}} q^{n(n-1)/2 + N_w} (q-1)^{e_w} K_1(\alpha)^{f_w},$$

where S_n is the symmetric group and where for an involution $w \in S_n$, f_w (resp. e_w) is the number of fixed points (resp. involution pairs) of w (i.e., $n = 2e + f$) and

$$N_w = \big|\{(i, j) \mid 1 \le i < j \le n, \ w(j) < w(i) \le j\}\big|.$$

One can similarly express $K_n(a)$ for $a = \alpha I + \sum_{i=1}^{n-1} e_{i,i+1}$, when $\alpha \ne 0$ and where $e_{i,j}$ is the elementary matrix with 1 in the (i, j)-th position and 0 everywhere else; see (58) in Section 5E.

The use of the Borel subgroup Bruhat decomposition is also very suitable for deriving estimates for these generalized Kloosterman sums. As a first step we have the following.

Theorem 1.6. *If a has a unique eigenvalue α then*

$$|K_n(a)| \le |K_n(\alpha I)| \le 4q^{(3n^2 - \delta(n))/4},$$

where $\delta(n) = 0$ if n is even and $\delta(n) = 1$ if n is odd.

Thus if the characteristic polynomial of a splits over $\mathbb{F}_q$, then the estimates do not require much input from étale cohomology. However to bound the sum $K_n(a)$ in general it seems unavoidable to use cohomological methods. The main input from étale cohomology is Lemma 3.16 from which we can derive the following.

Theorem 1.7. *Let $a \in \mathrm{GL}_n(\mathbb{F}_q)$ be a regular semisimple element (i.e., the characteristic polynomial p_a has no multiple roots over $\overline{\mathbb{F}}$). Then the exponential sum $K_n(a)$ is cohomologically pure — that is, all the cohomology groups are trivial but the middle one and all the weights are n^2.*

In particular for these regular semisimple elements, we have "square-root cancellation"

$$|K_n(a)| \le 2^n q^{n^2/2}.$$

Remark 2. The conditions on the multiplicities of the roots of p_a can be formulated as polynomial equations with integral coefficients in the variables $a_{i,j}$; thus Theorem 1.7 is a concrete illustration of the theorems on "generic purity" of Katz and Laumon [1985] and Fouvry and Katz [2001, Theorem 1.1].

It is an intriguing question if the set of regular semisimple elements is the actual purity locus. The methods of Theorem 1.3, at least for low ranks, allow one to see that matrices, whose Jordan normal form over the algebraic closure has an eigenvalue with more than one Jordan block, are too large for square root cancellations as can be seen by inspection. Interestingly matrices with only one block for each eigenvalue do exhibit square root cancellation. Our methods do not yield enough information to decide whether these sums are cohomologically pure; see Section 5E.

Remark 3. Will Sawin suggested to us a geometric approach that could shed even more light on these sums. The sum $K_n(a)$ may be viewed as the trace of Frobenius on the stalk at a of a complex of sheaves (in fact, a perverse sheaf). The geometry of this complex is linked to the Springer correspondence, in the sense that the stalk at a should be related to the cohomology of the fixed points of a acting on various flag varieties. This is an interesting and promising approach whose exact shape can be conjectured from the recursion formulas. While the paper was going through publication this suggestion was elaborated in [Erdélyi et al. 2024a], settling the purity question in the previous remark in the positive.

Our main theorem for bounding $K_n(a)$ is as follows.

Theorem 1.8. *For all $a \in \mathrm{GL}_n(\mathbb{F}_q)$ we have*

$$|K_n(a)| \ll_n q^{(3n^2-\delta(n))/4},$$

where $\delta(n) = 0$ if n is even and $\delta(n) = 1$ if n is odd.

Remark 4. It is possible to refine the statement based on the characteristic polynomial of a. If the characteristic polynomial is $p_a(t) = \prod_{i=1}^r (t - \alpha_i)^{n_i}$ for some pairwise distinct $\alpha_i \in \overline{\mathbb{F}}$, then

$$|K_n(a)| \ll_n \prod_{i=1}^r q^{(3n_i^2-\delta(n_i))/4} \prod_{1 \le i < j \le r} q^{n_i n_j}.$$

In the classical picture it is natural to look at the more general expression

$$K_1(\alpha, \beta) = K(\alpha, \beta, \mathbb{F}_q^*) = \sum_{\gamma \in \mathbb{F}_q^*} \varphi(\alpha\gamma + \beta\gamma^{-1}), \tag{5}$$

which, in the case $\beta \ne 0$, immediately reduces to $K_1(\alpha\beta^{-1})$ by the simple identity

$$K_1(\alpha, \beta) = K_1(\alpha\delta, \beta\delta^{-1})$$

valid for any $\delta \in \mathbb{F}_q^*$. However the case $\beta = 0$ is again interesting in its own way as it is the Fourier transform of the characteristic function of the set of invertible elements. While trivial to evaluate it is also unavoidable in the analytic applications.

We will look at these sums for $n \times n$ matrices and so we let

$$K_n(a, b) = K_n(a, b, \mathbb{F}_q) = \sum_{x \in \mathrm{GL}_n(\mathbb{F}_q)} \psi(ax + bx^{-1}). \tag{6}$$

It is easy to see that just like as above, we have $K_n(a, b) = K_n(b, a)$ and for any invertible g, h

$$K_n(gah^{-1}, hbg^{-1}) = K_n(a, b). \tag{7}$$

Therefore if $\det a$ or $\det b$ is nonzero, we may use the results above to obtain the bound $|K_n(a, b)| \ll q^{3n^2/4}$. However unlike in dimension 1, the other cases are not settled completely by Theorem 1.2, as the orbit structure of pairs of matrices (a, b) under the $\mathrm{GL}_n \times \mathrm{GL}_n$-action given in (7) gets more intricate when $n > 1$. We start with the most natural case of $b = 0$.

Theorem 1.9. *Assume that $a \in M_n(\mathbb{F}_q)$ has rank r. Then*

$$K_n(a, 0, \mathbb{F}_q) = (-1)^r q^{-r(r+1)/2} q^{rn} |\mathrm{GL}_{n-r}(\mathbb{F}_q)|.$$

Since for an r-dimensional subspace the Moebius function of the poset of subspaces evaluates to $(-1)^r q^{r(r-1)/2}$ by the q-binomial theorem [Stanley 1986, Formula (1.87)], this result is vaguely reminiscent to the evaluation of the Ramanujan sums

$$\sum_{\substack{1 \le x \le q \\ (x,q)=1}} e^{2\pi i a x/q} = \sum_{d | (a,q)} \mu(d) \frac{q}{d}.$$

One can see that the sums $K_n(a, b)$ can get significantly larger than $q^{3n^2/4}$ since for rank $a = 1$ we have $|K_n(a, 0)| \gg q^{n^2-n}$. The next theorem shows that this is close to the worst case scenario.

Theorem 1.10. *Let $a, b \in M_n(\mathbb{F}_q)$ be singular $n \times n$ matrices such that*

$$r = \mathrm{rk}(b) \ge s = \mathrm{rk}(a).$$

Then

$$|K_n(a, b, \mathbb{F}_q)| \le 2q^{n^2 - rn + r^2 + \binom{\min(r, n-r)}{2}}.$$

Corollary 1.11. *If $a, b \in M_n(\mathbb{F}_q)$ are not both 0, we have the general estimate*

$$|K_n(a, b, \mathbb{F}_q)| \le 2q^{n^2 - n + 1}.$$

Remark 5. Apart from the constant 2 which can probably be replaced by $1 + o(1)$ as $q^n \to \infty$, this bound is sharp, since

$$K_n(\boldsymbol{e}_{1,n}, \boldsymbol{e}_{1,n}) = q^{2n-2} |\mathrm{GL}_{n-2}(\mathbb{F}_q)| + (q-1)q^{n-1} |\mathrm{GL}_{n-1}(\mathbb{F}_q)| \sim q^{n^2-n+1}. \tag{8}$$

(Here again $\boldsymbol{e}_{i,j}$ stands for the elementary matrix with 1 in the (i, j)-th position and 0 everywhere else.)

1C. *The organization of the paper.* The paper naturally splits into four parts. The first deals with matrices a all of whose eigenvalues are in the ground field $\mathbb{F}_q$ with the aim of evaluating the generalized Kloosterman sums in terms of classical ones. Later sections deal with the nonsplit case using cohomology, and the entirely different and more combinatorial case of degenerate matrices. These are included because they are needed for the equidistribution problem in [El-Baz et al. 2022]. The final part provides some examples and highlights some open questions. The material bifurcates on several occasions and this may somewhat obscure the insight one gains from the results. Therefore we first highlight the generic case of regular semisimple matrices before further details.

To treat regular semisimple elements in concrete terms we need to assume that they are split over the ground field. The calculation to reduce to block upper diagonal a with the blocks having no common eigenvalue is presented in Section 2A. We show here that in that case the Kloosterman sum $K_n(a)$ factors as a product of Kloosterman sums of smaller ranks corresponding to the diagonal blocks. This is elementary and leads immediately to Theorem 1.1. If one is merely interested in this generic case then

one can jump to Section 3F which shows how to circumvent the problem when the eigenvalues are only defined in a field extension. In this regard one is also naturally lead to Conjecture 5.9 which suggests that an evaluation might be possible even in the nonsplit case.

The finer picture of the non semisimple case is both of natural interest and required by the applications. We spend a great deal of the paper on them. After the proof of Theorem 1.1, Section 2 and parts of Section 3 handle this case still under the assumption that the eigenvalues are in the ground field $\mathbb{F}_q$. To reach our goals we will evaluate the subsums of $K_n(a)$ restricted to Bruhat cells in various decompositions.

While the calculations in Section 2 are somewhat lengthy the overall structure is simple. We use Bruhat decomposition with respect to a maximal parabolic fixing a line. In Sections 2B and 2C we introduce the necessary notation for this task. This decomposition is the first step in the plain old Gauss–Jordan elimination and it is natural to expect that this process should lead to an inductive structure for matrix Kloosterman sums. An elementary computation in Section 2C justifies this expectation.

Since nilpotent matrices can be put into Jordan normal form over any field this step immediately shows that Theorem 1.3 holds, at least if the polynomial in the statement is allowed to depend on the characteristic p. It is easy to see that in the simplest case of Theorem 1.4 there is a recursion that works over all primes simultaneously. The bounds in Section 3 are based only on this result and if one is only interested in the estimates the rest of the section can be skipped.

However from the exponential sum point of view independence on the characteristic is of great interest in itself. Therefore it is important that the recursions to lower rank can be done across all finite fields universally for matrices whose eigenvalues are in that field. The rest of Section 2 is then devoted to show this. Section 2D presents the technical core by showing that one may restrict to matrices with entries 1 and 0. These matrices can be lifted to $\mathbb{Z}$ and can be put into Jordan normal form over $\mathbb{Q}$. In effect this establishes that Theorem 1.3 holds for almost all primes simultaneously. The final step in the proof is then a simple criterion for the existence of a Jordan normal form over $\mathbb{Z}$ in Section 2F. While this Jordan normal form reduction can be made explicit, the rather technical calculations are postponed to Section 5B.

The second main part of the paper, Section 3, is about estimates. This again starts with an elementary but structural observation that repeated applications of the reduction in Section 2 is equivalent to the use of the Bruhat decomposition with respect to a Borel subgroup. We are able to refine the usual Bruhat decompositions just enough to establish the first bound in Theorem 1.6.

While the proof of the above estimates are still group-theoretic in nature, in general one needs methods from cohomology. Conversely the cohomological apparatus relies on these more classical arguments. The connection is given by a key lemma, Lemma 3.16, whose simple proof somewhat disguises its importance. This statement allows one to push "trivial cancellations" over a field extension back to the original field.

In Section 3E we review the cohomological apparatus by listing the necessary tools for the linear algebra that follows. Sections 3F and 3G then give the proofs of the main theorems, including the special case that the sums $K_n(a)$ are "pure" when a is a regular semisimple element. These results are based on the fact that the cohomology groups attached to the subsum restricted to a Bruhat cell vanishes in large enough degree.

The third part deals with $K_n(a, b)$ in the degenerate cases. After dealing with the case $b = 0$ the general situation is reduced to this special case when $b = 0$. Again one needs a double coset decomposition suitable for this task. The machinery here is combinatorial in nature using Gaussian binomials [Stanley 1986].

In the last part we provide examples to illustrate our results. We give explicit evaluations for low ranks. The case $K_2(a, b)$ is explicitly computed for any $a, b \in M_2(\mathbb{F}_q)$ including the nonsplit semisimple case. There are further calculations that illustrate the difficulties to get explicit results for $n \geq 4$. These sections also contain some observations and open questions that are interesting on their own.

Notational conventions. Letters of the Greek alphabet $\alpha, \ldots, \lambda, \ldots, \xi$ will denote scalars, that is, elements of the field $\mathbb{F}_q$, or its algebraic closure. Lower case letters of the Latin alphabet $a, \ldots, g, \ldots, x$ will denote matrices, but their entries, while scalars, will usually be denoted with Latin characters $a_{ij}, \ldots, g_{ij}, \ldots, x_{ij}$ as well. Upper case letters $A, \ldots, G, \ldots, X$ will denote sets of matrices, usually but not necessarily subgroups. While these depend on n, that dependence is usually suppressed for better readability. There are a few exceptions that should not cause confusion: a general partition will be denoted λ, I will denote the identity matrix (I_n if the size is not clear from the context), K will denote Kloosterman sums, and on occasion $S_1, S_2, \ldots$ will denote some auxiliary sums.

For the cohomological arguments in Sections 3E–3G we have to work with algebraic varieties, so then by $A, \ldots, G, \ldots, X$ we will also denote the affine algebraic varieties of M_n defined by simple algebraic equations in the matrix entries.

There are two notions of a trace in the paper: $\mathrm{Tr} = \mathrm{Tr}_{\mathbb{F}_{q^m}/\mathbb{F}_q}$ will denote the trace from an overfield to the ground field, while $\mathrm{tr} = \mathrm{tr}_n$ will denote the trace of a matrix of size $n \times n$.

If u is a unipotent matrix that is upper triangular, we will denote $\bar{u} = u - I$ the strictly upper triangular part of u. Similarly, if U is the group of upper triangular unipotent matrices and $V \subset U$, then $\overline{V}$ denotes the set (or variety) $\{\bar{v} \mid v \in V\}$.

An important role is played by the Weyl group $W \simeq N(T)/T$ of GL_n. Here T is the set of diagonal matrices in GL_n, and $N(T)$ is its normalizer. If S_n is the group of permutations on the letters $\{1, \ldots, n\}$ then $W \simeq S_n$, and we will choose a specific isomorphism in Section 3A.

Cohomology will mean ℓ-adic cohomology with compact support, with some unspecified $\ell \neq p$.

2. Identities via reduction of rank

2A. *Splitting to primary components.* We start with some elementary but important observations. Recall from (5) and (2) that

$$K_n(a, b) = \sum_{x \in \mathrm{GL}_n(\mathbb{F}_q)} \psi(ax + bx^{-1}) \quad \text{and} \quad K_n(a) = K_n(a, I).$$

The sum is clearly unchanged if x is replaced by x^{-1}, or if x is replaced by cx for some $c \in \mathrm{GL}_n(\mathbb{F}_q)$. This leads immediately to:

Lemma 2.1. *If $a, b \in M_n(\mathbb{F}_q)$ and $c \in \mathrm{GL}_n(\mathbb{F}_q)$ then we have*

$$K_n(a, b) = K_n(b, a), \quad K_n(ac, b) = K_n(a, cb) \quad and \quad K_n(a) = K_n(cac^{-1}).$$

Therefore, $K_n(a)$ only depends on the conjugacy class of a and so using conjugation we can put a into Jordan normal form over $\overline{\mathbb{F}}$. To exploit this we recall the following result of Sylvester [1885] whose proof is given here for the convenience of the reader.

Lemma 2.2. *Let $M_{k,l}$ be the vector space of $k \times l$ matrices over an arbitrary field F. If $a_k \in M_k(F)$ and $a_l \in M_l(F)$ are matrices with no common eigenvalue over the algebraic closure $\overline{F}$ of F then the linear endomorphism of $M_{k,l}$ given by*

$$v \mapsto v a_l - a_k v$$

is an isomorphism.

Proof. It suffices to prove that this map is injective. Assume that

$$v a_l - a_k v = 0$$

for some v. Then for any polynomial $p(t)$

$$v p(a_l) = p(a_k) v.$$

Let $p_k(t) \in F[t]$ (resp. $p_l(t)$) be the characteristic polynomial of a_k (resp. a_l). Then by the Cayley–Hamilton theorem we have

$$0 = p_k(a_k) v = v p_k(a_l) \quad and \quad 0 = v p_l(a_l).$$

By our assumption on the eigenvalues we have that $p_k(t)$ and $p_l(t)$ are relatively prime in $F[t]$. Thus there are polynomials $r_k(t), r_l(t)$ such that $p_k(t) r_k(t) + p_l(t) r_l(t) = 1$ which implies $0 = v$; hence our claim. $\square$

To use this observation let $U_{[k,l]}$ be the linear subgroup of GL_n whose set of points for any ring R is

$$U_{[k,l]}(R) = \left\{ \left(\begin{array}{c|c} I_k & v \\ \hline 0 & I_l \end{array} \right) \,\middle|\, v \in M_{k,l}(R) \right\}.$$

The fact that this is a subvariety will play a role in the cohomological arguments, but for now we will only use the particular subgroup $U_{k,l}(\mathbb{F}_q) \subset \mathrm{GL}_n(\mathbb{F}_q)$. As usual since the field $\mathbb{F}_q$ is fixed, for easier reading we will write $K_n(a)$ for $K_n(a, \mathbb{F}_q)$, and G and $U_{k,l}$ for the sets $\mathrm{GL}_n(\mathbb{F}_q)$ and $U_{k,l}(\mathbb{F}_q)$.

Proposition 2.3. *Assume that*

$$a = \left(\begin{array}{c|c} a_k & b \\ \hline 0 & a_l \end{array} \right)$$

with $a_k \in M_k(\mathbb{F}_q)$, $a_l \in M_l(\mathbb{F}_q)$, $b \in M_{k,l}(\mathbb{F}_q)$ for some $k, l \in \mathbb{N}$ such that $k + l = n$ and a_k and a_l have no common eigenvalue in $\overline{\mathbb{F}}$. Then

$$K_n(a) = q^{kl} K_k(a_k) K_l(a_l).$$

Proof. Since tr is invariant under conjugation, we have

$$K_n(a) = \sum_{x \in G} \psi(ax + x^{-1}) = \frac{1}{q^{kl}} \sum_{x \in G} \sum_{u \in U_{[k,l]}} \psi(a(u^{-1}xu) + (u^{-1}xu)^{-1})$$

$$= \frac{1}{q^{kl}} \sum_{x \in G} \sum_{u \in U_{[k,l]}} \psi((uau^{-1})x + x^{-1}).$$

Now

$$uau^{-1} = \left(\begin{array}{c|c} a_k & b + va_l - a_k v \\ \hline 0 & a_l \end{array} \right)$$

and so

$$K_n(a) = \frac{1}{q^{kl}} \sum_{x \in G} \psi(ax + x^{-1}) \left(\sum_{v \in \mathbb{F}_q^{k \times l}} \psi_k((va_l - a_k v)x') \right),$$

where x' is the $l \times k$ matrix which we get by deleting the first k rows and last l columns of x and ψ_k is the $k \times k$ matrix trace composed with φ.

From Lemma 2.2 we have

$$\sum_{v \in \mathbb{F}_q^{k \times l}} \psi_k((va_l - a_k v)x') = \sum_{v \in \mathbb{F}_q^{k \times l}} \psi_k(vx') = \begin{cases} 0 & \text{if } x' \neq 0, \\ q^{kl} & \text{if } x' = 0. \end{cases}$$

This immediately yields

$$K_n(a) = \sum_{\substack{x \in G \\ x' = 0}} \psi(ax + x^{-1}) = q^{kl} K_k(a_k) K_l(a_l). \qquad \square$$

Proof of Theorem 1.1. The first claim was proved above. For the second using the invariance under conjugation we may assume that $a = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$. By Proposition 2.3 we have

$$K_n(a) = q^{n-1} K_1(\alpha_n) K_{n-1}(a'),$$

where $a' = \mathrm{diag}(\alpha_1, \ldots, \alpha_{n-1})$. The result follows by induction. $\qquad \square$

2B. *A parabolic Bruhat decomposition.* We prepare the proofs of Theorems 1.3 and 1.4. Our goal is to express the Kloosterman sum $K_n(a)$ in terms of sums $K_{n-1}(a')$ and $K_{n-2}(a'')$ where the matrices a' and a'' are derived from a by deleting one or two rows and columns.

When a has a single eigenvalue α, it is conjugate to

$$a = \alpha I + \bar{a}, \tag{9}$$

where $\bar{a}$ is strictly upper triangular. Since $K_n(a)$ is conjugacy invariant, we will assume that a itself is in the above form. Our reductions are then based on a parabolic Bruhat decomposition for GL_n. While it can be deduced from general facts — see [Springer 1998, Theorem 8.3.8; Borel 1991, Proposition IV.14.21(iii)] — it is easier to work them out explicitly for the special case at hand.

Let P be the closed subgroup of GL_n defined by the vanishing of $g_{n,1}, \ldots, g_{n,n-1}$. If F is a field, P may be described alternatively as follows. Let $\mathrm{GL}_n(F)$ act on row vectors by multiplication on the right: $v \mapsto vg$. Then $P(F)$ is the stabilizer of the line $\{\lambda \, \boldsymbol{e}_n : \lambda \in F\}$ where $\boldsymbol{e}_n$ is the row vector $(0, \ldots, 0, 1)$:

$$P(F) = \{g \in \mathrm{GL}_n(F) \mid \boldsymbol{e}_n g = \lambda \boldsymbol{e}_n, \ \lambda \in F^*\}. \tag{10}$$

Since the arguments in this and the following sections will not be used in the cohomological proofs we will only concentrate on the set $P(\mathbb{F}_q)$.

Then as sets $G = \mathrm{GL}_n(\mathbb{F}_q) = \bigsqcup_{k=1}^n P(\mathbb{F}_q) w_{(kn)} P(\mathbb{F}_q)$, where $w_{(kn)}$ is the permutation matrix corresponding to the transposition (kn). To make this a parameterization let

$$U_k = \left\{ I + \sum_{j=k+1}^n u_j \boldsymbol{e}_{k,j} \ \middle| \ u_{k,j} \in F \right\} \tag{11}$$

be the set of unipotent matrices with nonzero elements only in the k-th row. (While this is again an algebraic group scheme, this fact will not play any role.)

We will only deal with $F = \mathbb{F}_q$ and from here on we will write P and U_k for $P(\mathbb{F}_q)$ and $U_k(\mathbb{F}_q)$. We then have the following decomposition into disjoint sets.

Lemma 2.4. *Let $X_k = U_k w_{(kn)} P$. The map $U_k \times P \to X_k$, $(u, g) \mapsto u w_{(kn)} g$, is a bijection and $G = \bigsqcup_{k=1}^n U_k w_{(kn)} P$.*

Proof. Let x be a matrix in $G = \mathrm{GL}_n(\mathbb{F}_q)$ with rows $\boldsymbol{x}_i$, and write

$$\boldsymbol{e}_n = \sum_{j=1}^n u_j \boldsymbol{x}_j,$$

where $\boldsymbol{e}_n = (0, \ldots, 0, 1)$. We claim that

$$X_k = \left\{ x \in G \mid \min\{j \mid u_j \neq 0\} = k \right\}.$$

It is clear that $X_k P = X_k$ and that $U_k w \subset X_k$ with $w = w_{(kn)}$. Conversely if we let

$$u = I + \sum_{j=k+1}^n (u_j/u_k) \boldsymbol{e}_{k,j}$$

then we have $ux \in wP$.

Finally it is enough to show that if $u_1 w p_1 = u_2 w p_2$, then $u_1 = u_2$. To this effect note that the above implies that $\boldsymbol{e}_n w u_2^{-1} u_1 = \boldsymbol{e}_n p_2 p_1^{-1} w$. However $\boldsymbol{e}_n w = \boldsymbol{e}_k$ and so the k-th rows of u_1 and u_2 are the same, which implies $u_1 = u_2$. $\square$

By the lemma we have

$$K_n(a) = \sum_{g \in \mathrm{GL}_n(\mathbb{F}_q)} \psi(ag + g^{-1}) = \sum_{k=1}^n \sum_{x \in X_k} \psi(ax + x^{-1}). \tag{12}$$

When summing over X_k, the case of $k = n$, when $X_n = P$, is trivial. To see this we will give an explicit Levi decomposition of $P = P(\mathbb{F}_q)$. This fixes notation and will also be used in our further calculations on X_k for $k < n$. Note that again we will be working not with the algebraic groups but the fixed finite groups that they give rise to for $\mathbb{F}_q$.

For $h \in \mathrm{GL}_{n-1}(\mathbb{F}_q)$ and $\mu \in \mathrm{GL}_1(\mathbb{F}_q)$ let

$$[h, \mu] = \begin{bmatrix} h & 0 \\ 0 & \mu \end{bmatrix} \in \mathrm{GL}_n(\mathbb{F}_q) \tag{13}$$

and let

$$L = \{[h, \mu] : h \in \mathrm{GL}_{n-1}(\mathbb{F}_q), \ \mu \in \mathrm{GL}_1(\mathbb{F}_q)\} \subset \mathrm{GL}_n(\mathbb{F}_q). \tag{14}$$

Also let

$$V = \left\{ I + \sum_{l=1}^{n-1} v_l e_{l,n} : v_l \in \mathbb{F}_q \right\}. \tag{15}$$

Then

$$P = LV = VL.$$

Proposition 2.5. *If a is as in* (9) *then*

$$\sum_{x \in X_n} \psi(ax + x^{-1}) = q^{n-1} K_1(\alpha) K_{n-1}(a'),$$

where a' is the matrix one gets by deleting the last row and column of a.

Proof. Since

$$\sum_{x \in X_n} \psi(ax + x^{-1}) = \sum_{\substack{g \in L \\ v \in V}} \psi(agv + (gv)^{-1}) = q^{n-1} \sum_{g \in L} \psi(ag + g^{-1}),$$

the claim follows from the description of L in (14) and that

$$\mathrm{tr}(agv + (gv)^{-1}) = \mathrm{tr}(ag + g^{-1}) = \mathrm{tr}(a'h + h^{-1}) + \alpha\lambda + \lambda^{-1}. \qquad \square$$

2C. *The sum over the nontrivial cells.* We continue to assume that $a = \alpha I_n + \bar{a}$, with $\bar{a}$ strictly upper triangular, so that a has a unique eigenvalue α. We will show that for $k < n$ the sum

$$\sum_{x \in X_k} \psi(ax + x^{-1})$$

can be expressed as a sum over the subvariety

$$L_k(\alpha) = \{g \in L \mid g_{k,j} = 0 \text{ for all } j \neq k \text{ and } \alpha g_{k,k} = g_{n,n}^{-1}\}. \tag{16}$$

However we will only work over the set of points in $\mathbb{F}_q$ and write $L_l(\alpha)$ for $L_k(\alpha)(\mathbb{F}_q)$. For $\alpha = 0$ these sets are empty, while for $\alpha \in \mathbb{F}_q^*$ they are subvarieties of L isomorphic to $\mathrm{GL}_{n-2} \times \mathrm{GL}_1 \times \mathbb{A}^{n-2}$ that can

be visualized as elements $g \in L$ of the form

$$g = \begin{bmatrix} h_{11} & * & h_{12} & 0 \\ 0 & \mu & 0 & 0 \\ h_{21} & * & h_{22} & 0 \\ 0 & 0 & 0 & 1/(\alpha\mu) \end{bmatrix}. \tag{17}$$

Here the blocks correspond to the partition

$$\{1, \ldots, n\} = \{1, \ldots, k-1\} \sqcup \{k\} \sqcup \{k+1, \ldots, n-1\} \sqcup \{n\}$$

for $1 < k < n-1$, while for $k = 1$ and $n-1$ we have to adapt (17) to 3×3 blocks

$$g = \begin{bmatrix} \mu & 0 & 0 \\ * & h'' & 0 \\ 0 & 0 & 1/(\alpha\mu) \end{bmatrix}, \quad g = \begin{bmatrix} h'' & * & 0 \\ 0 & \mu & 0 \\ 0 & 0 & 1/(\alpha\mu) \end{bmatrix}. \tag{18}$$

This is merely a preliminary step in the reduction to rank $n-2$, but is already quite useful, a fact that we will illustrate by proving Theorems 1.2 and 1.4.

The reduction to the special form in (17) is based on the following calculation.

Proposition 2.6. *For a fixed $g \in L$,*

$$\sum_{v \in V} \psi(\alpha w_{(kn)} g v + (w_{(kn)} g v)^{-1}) = 0$$

unless

$$g_{k,j} = 0 \quad \text{for all } j \neq k \quad \text{and} \quad \alpha g_{k,k} = g_{n,n}^{-1}.$$

When these conditions hold

$$\sum_{v \in V} \psi(\alpha w_{(kn)} g v + (w_{(kn)} g v)^{-1}) = q^{n-1} \psi(\alpha w_{(kn)} g + (w_{(kn)} g)^{-1}).$$

Proof. Let $\overline{V} = \{\bar{v} = \sum_{l=1}^{n-1} v_l e_{l,n} : v_l \in \mathbb{F}_q\}$. If $v = I + \bar{v} \in V$ then $v^{-1} = I - \bar{v}$. The sum in question then becomes

$$\sum_{v \in V} \psi(\alpha w_{(kn)} g v + (w_{(kn)} g v)^{-1}) = \psi(\alpha w_{(kn)} g + (w_{(kn)} g)^{-1}) \sum_{\bar{v} \in \overline{V}} \psi(\alpha w_{(kn)} g \bar{v} - \bar{v}(w_{(kn)} g)^{-1}),$$

which vanishes unless the linear function

$$v \mapsto \mathrm{tr}(\alpha w_{(kn)} g \bar{v} - \bar{v}(w_{(kn)} g)^{-1}) = \alpha \sum_{l=1}^{n-1} g_{k,l} v_l - g_{n,n}^{-1} v_k$$

is trivial. $\qquad\square$

We can now prove the following claim about the sum over X_k.

Proposition 2.7. *Let $a = \alpha I + \bar{a}$, where $\bar{a}$ is strictly upper triangular, $k < n$, X_k is as in Lemma 2.4 and $L_k(\alpha)$ is as in (16). Then we have*

$$\sum_{x \in X_k} \psi(ax + x^{-1}) = q^{n-1} \sum_{g \in L_k(\alpha)} \sum_{u \in U_k} \psi(a^u w_{(kn)} g + (w_{(kn)} g)^{-1}),$$

where $a^u = u^{-1} a u$.

Proof. Since $Pu = P$ for any $u \in U_k$, we have

$$X_k = \bigsqcup_{u \in U_k} u w_{(kn)} P = \bigsqcup_{u \in U_k} u w_{(kn)} P u^{-1},$$

and so by $u^{-1} a u = \alpha I + u^{-1} \bar{a} u$ we have

$$\sum_{x \in X_k} \psi(ax + x^{-1}) = \sum_{\substack{g \in L \\ v \in V \\ u \in U_k}} \psi(\alpha w_{(kn)} g v + (w_{(kn)} g v)^{-1}) \psi(u^{-1} \bar{a} u w g v).$$

A direct calculation, based on the fact that the last row of $\bar{a} = a - \alpha I$ is identically 0, then shows that

$$\operatorname{tr}(u^{-1} \bar{a} u w g v) = \operatorname{tr}(u^{-1} \bar{a} u w g) \tag{19}$$

is independent of v. Therefore

$$\sum_{x \in X_k} \psi(ax + x^{-1}) = \sum_{\substack{g \in L \\ u \in U_k}} \psi(u^{-1} \bar{a} u w g) \sum_{v \in V} \psi(\alpha w_{(kn)} g v + (w_{(kn)} g v)^{-1}).$$

The inner sum is identical to the one in Proposition 2.6; thus the proposition is proven. $\qquad \square$

Remark 6. We comment briefly on identity (19). The calculations are simplified by using $M_n(\mathbb{F}_q)$, writing $v = I + \bar{v}$, and observing that $\operatorname{tr} x \bar{v} = \sum_{l=1}^{n} x_{n,l} v_l$ which clearly vanishes if the last row of the matrix x is identically 0.

However one may argue alternatively via interpreting these matrices as linear transformations as follows. The group P is the parabolic subgroup fixing the 1-dimensional subspace $M = \{\lambda e_n \mid \lambda \in \mathbb{F}_q\}$ and so its elements also preserve the flag $\{0\} \subset M \subset \mathbb{F}_q^n$. Any element g of P then gives rise to a linear transformation g' of $M' = \mathbb{F}_q^n / M$. The subgroup V itself is characterized by the property that its elements act trivially both on M and on M'. Let $x = u^{-1} \bar{a} u w g$. Since $e_n x = 0$, the linear transformation x also induces a map x' on M' and $\operatorname{tr} x = \operatorname{tr} x'$. Since $e_n x v = 0$ as well, $\operatorname{tr} x v = \operatorname{tr}(x v)' = \operatorname{tr} x' v' = \operatorname{tr} x'$.

In general all our calculations can easily be proved using block matrices, either by hand or by using a symbolic algebra package. Since this gives an easy way to check the validity of these statements, we will present most of the identities in this matrix interpretation.

As a corollary to Proposition 2.7 we immediately have:

Proof of Theorem 1.2. If $\alpha = 0$ then the set $L_k(\alpha)$ is empty, and so for $k < n$

$$\sum_{x \in X_k} \psi(ax + x^{-1}) = 0.$$

Since $K_1(0) = -1$, Proposition 2.5 gives

$$K_n(a) = -q^{n-1} K_{n-1}(a'),$$

where a' is the matrix one gets by deleting the last rows and columns of a. Since by assumption a is upper triangular nilpotent, so is a' and the theorem follows by induction. $\square$

Proof of Theorem 1.4. If $a = \alpha I$, with $\alpha \in \mathbb{F}_q^*$ a scalar matrix, then $\bar{a} = 0$ and $a^u = \alpha I$. If $1 < k < n-1$ and g is as in (17) then g is invertible if and only if $g'' = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$ is invertible, in which case $(g^{-1})'' = (g'')^{-1}$. It follows that for such k

$$\sum_{x \in X_k} \psi(ax + x^{-1}) = q^{2n-3}(q-1)K_{n-2}(\alpha I)q^{n-k},$$

and it is easy to see that this relation holds for $k = 1, n-1$ as well. This gives

$$K_n(\alpha I) = \sum_{k=1}^{n} \sum_{x \in X_k} \psi(\alpha x + x^{-1}) = q^{n-1}K_1(\alpha)K_{n-1}(\alpha I) + q^{2n-3}(q-1)K_{n-2}(\alpha I)\sum_{k=1}^{n-1} q^{n-k}$$

from which the desired formula follows. $\square$

2D. *Reduction to* $\mathbf{GL_{n-2}}$. Assume that $a = \alpha I + \bar{a}$ has a unique eigenvalue $\alpha \neq 0$, and that $\bar{a}$ is strictly upper triangular. Since the results of the previous section take care of the case when $n = 2$ or $a = \alpha I$, we will assume that $n \geq 3$ and that $\bar{a} \neq 0$.

Recall that $\sum_{x \in X_k} \psi(ax + x^{-1}) = q^{n-1} \sum_{u,g} \psi(a^u w_{(kn)}g + (w_{(kn)}g)^{-1})$, the sums over $u \in U_k$ and $g \in L_k(\alpha)$. We will use the fact that as a variety $L_k(\alpha)$ is isomorphic to $\mathrm{GL}_{n-2}(\mathbb{F}_q) \times \mathbb{F}_q^* \times \mathbb{F}_q^{n-2}$ to express $\sum_{x \in X_k} \psi(ax + x^{-1})$ as a linear combination of Kloosterman sums of rank $n - 2$.

To state the reduction step we denote by $m''_{\not{k},\not{n}}$ the matrix one gets by deleting the k-th and n-th rows and columns of a matrix m. For us n will be fixed, and the value of k will be clear from the context, in which case we will often simply write m''. For any matrix m let $m_{(k)}$ denote its k-th row, and $m^{(l)}$ denote its l-th column.

Proposition 2.8. *Assume that $n > 2$, $a = \alpha I + \bar{a}$ with $\bar{a}$ strictly upper triangular and let $u = I + \bar{u} \in U_k$. Then we have that*

$$\sum_{g \in L_k(\alpha)} \psi(a^u w_{(kn)}g + (w_{(kn)}g)^{-1}) = 0$$

unless $\bar{u}_{(k)}\bar{a}^{(j)} = \bar{a}_{k,j}$ for $j = k+1, \ldots, n-1$. When this condition holds

$$\sum_{g \in L_k(\alpha)} \psi(a^u w_{(kn)}g + (w_{(kn)}g)^{-1}) = q^{n-1}K_{n-2}(a'' + z)\sum_{\lambda \in \mathbb{F}_q^*} \varphi(\lambda\xi),$$

where $z = (\bar{a}^{(k)}\bar{u}_{(k)})'' \in M_{n-2}$ and $\xi = a_{k,n} - \bar{u}_{(k)}\bar{a}^{(n)}$.

Remark 7. When $k = 1$ or $n - 1$ the perturbation z vanishes for any u.

Proof. A direct calculation shows that

$$a^u = (I - \bar{u})a(I + \bar{u}) = a - \bar{u}\bar{a} + \bar{a}\bar{u}. \tag{20}$$

First assume that $1 < k < n - 1$ and that

$$g = \begin{bmatrix} g_{11} & y_1 & g_{12} & 0 \\ 0 & \mu & 0 & 0 \\ g_{21} & y_2 & g_{22} & 0 \\ 0 & 0 & 0 & 1/(\alpha\mu) \end{bmatrix} \in L_k(\alpha)$$

as in (17), in which case $g'' = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$ is invertible. Let tr_n denote the $n \times n$ matrix trace. Then

$$\mathrm{tr}_n((wg)^{-1}) = \mathrm{tr}_{n-2}((g'')^{-1}).$$

Moreover

$$\mathrm{tr}_n(awg) = \mathrm{tr}_{n-2}(a''g'') + \mu a_{k,n} \quad \text{and} \quad \mathrm{tr}_n(\bar{a}\bar{u}wg) = \mathrm{tr}_{n-2}((a^{(k)}u_{(k)})''g''),$$

where we have used the fact that $\bar{u}$ has only one nonzero row $\bar{u}_{(k)}$.

Finally note that $\mathrm{tr}(a^u wg + (wg)^{-1})$ does not depend on the $(k-1) \times 1$ matrix y_1, and as a function of y_2 only depends on $\mathrm{tr}(\bar{u}\bar{a}wg)$. The function

$$y_2 \mapsto \mathrm{tr}(\bar{u}\bar{a}wg)$$

is constant if and only if

$$\bar{u}_{(k)}\bar{a}^{(j)} = 0$$

for $j = k+1, \ldots, n-1$, and if this condition does not hold the sum over $g \in L_k(\alpha)$ vanishes. This proves the claim for $1 < k < n - 1$. The remaining cases are treated similarly. $\square$

We will now specify the result in the case $a = \alpha I + \bar{a}$ is in Jordan normal form. There is a partition λ associated to a, i.e., a sequence of positive integers $n_1 \leq n_2 \leq \cdots \leq n_l$, such that $n_1 + n_2 + \cdots + n_l = n$. Conversely, given a partition $\lambda = [n_1, \ldots, n_l]$ let

$$N_i = n_1 + \cdots + n_i \tag{21}$$

so that $1 \leq N_1 < N_2 < \cdots < N_l = n$, and form

$$\bar{a}(\lambda) = \sum_{j=1}^{n-1} \varepsilon_j(\lambda)e_{j,j+1}, \quad \text{where} \quad \varepsilon_j(\lambda) = \begin{cases} 0 & \text{if } j = N_i, \text{ for some } i, \\ 1 & \text{otherwise.} \end{cases} \tag{22}$$

Any a with a single eigenvalue α is conjugate to one of the matrices $\alpha I + \bar{a}(\lambda)$ and we will assume from now on that a is already in that form. Since scalar matrices were already dealt with, we will also assume that $\lambda \neq [1, 1, \ldots, 1]$, which ensures that $\varepsilon_{n-1} = 1$.

Theorem 2.9. *Assume that $n > 2$, $\alpha \neq 0$, $a = \alpha I + \bar{a}(\lambda)$ with $\varepsilon_j = \varepsilon_j(\lambda) \in \{0, 1\}$ as in (22) with $\varepsilon_{n-1} = 1$. Then*:

(i) *We have*

$$\sum_{x \in X_{n-1}} \psi(ax + x^{-1}) = -q^{2n-2} K_{n-2}(a''),$$

where $a'' = a''_{\not{n-1}, \not{n}}$ — the matrix obtained by deleting the last two rows and columns of a.

(ii) *If $k \leq n - 2$ then*

$$\sum_{x \in X_k} \psi(ax + x^{-1}) = 0$$

unless $k = N_i$ for one of the N_i in (21) for which $n_i > 1$.

(iii) *When $k = N_i = n_1 + \cdots + n_i < n - 1$ and $n_i > 1$, we have*

$$\sum_{x \in X_k} \psi(ax + x^{-1}) = q^{2n-2} \sum_{\substack{z \in Z \\ \mu \in \mathbb{F}_q^*}} K_{n-2}(a'' + z)\varphi(\mu\xi_l), \tag{23}$$

where $a'' = a''_{\not{k}, \not{n}}$ and

$$Z = \left\{ \sum_{j=i+1}^{l-1} \xi_j e_{k-1, N_j-1} + \xi_l e_{k-1, n-2} \,\Big|\, \xi_j \in \mathbb{F}_q \text{ for } i+1 \leq j \leq l \right\} \subset M_{n-2}.$$

In Z the elementary matrices $e_{*,*}$ are of size $(n-2) \times (n-2)$.

Proof. The statements are easy corollaries of Proposition 2.8 except for the fact in (ii) that n_i must be greater than 1, which is equivalent to $\varepsilon_{k-1} \neq 0$. Since $k < n - 1$ and $\varepsilon_{k-1} = 0$ imply that the parameters in Proposition 2.8 are very simple, all $z = 0$, and $\xi = -\bar{u}_{k,n-1}$. Therefore

$$\sum_{x \in X_k} \psi(ax + x^{-1}) = q^{2n-2} \sum_u K_{n-2}(a'') \sum_{\mu \in \mathbb{F}_q^*} \varphi(-\mu\bar{u}_{k,n-1})$$

vanishes.   □

While the matrices $a'' + z$ are not in Jordan normal form, they are again matrices with a single eigenvalue α. Therefore collecting them according to their conjugacy classes gives a reduction algorithm, in fact, a characteristic p version of Theorem 1.3 (see Proposition 2.13). In the next two sections we will explicate this strategy and prove that the polynomials that arise this way do not depend on p.

2E. *Jordan normal forms over $\mathbb{Z}$.* The proof of Theorem 1.3 will be based on proving that the perturbations arising from the reduction to M_{n-2} can be collected into Jordan normal forms that do not depend on the characteristic p. For this we will need some details about Jordan normal forms for integral nilpotent matrices. This of course is a trivial task over $\mathbb{Q}$, but requires a little care when one works over $\mathbb{Z}$.

Assume, for example, that x is an $n \times n$ nilpotent matrix, and $g \in \mathrm{GL}_n(\mathbb{Z})$ is such that gxg^{-1} is in Jordan normal form as in (22). A moment's thought reveals that $\{vx \mid v \in \mathbb{Z}^n\}$, the row space of x, must

be a direct summand of $\mathbb{Z}^n$, which we also think of as row vectors. This by itself is not sufficient for the existence of a Jordan normal form over $\mathbb{Z}$ but we have the following.

Theorem 2.10. *Let $\bar{a}$ be an $n \times n$ nilpotent matrix with integral entries. There exist $g \in \mathrm{GL}_n(\mathbb{Z})$ such that $g\bar{a}g^{-1} = \sum_{j=1}^{n-1} \varepsilon_j e_{j,j+1}$, $\varepsilon_j \in \{0, 1\}$ if and only if*

$$\{v\bar{a}^k \mid v \in \mathbb{Z}^n\}$$

is a direct summand of $\mathbb{Z}^n$ (as an abelian group) for any $k \in \mathbb{N}$.

This is equivalent to the conditions that

$$\{\bar{a}^k v^T \mid v \in \mathbb{Z}^n\}$$

is a direct summand of $(\mathbb{Z}^n)^T$ for any k (here $\cdot^T$ is the matrix transpose).

The following examples illustrate the situation.

Example 2.11. Let $x = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}$. Its Jordan normal form over $\mathbb{Q}$ is $y = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. If $g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the equation $gx = yg$ leads to $c = 0$ and $d = 2a$. Therefore the equation $gxg^{-1} = y$ has no solution in $\mathrm{SL}_2(\mathbb{Z})$, or even $\mathrm{SL}_2(\mathbb{Q})$.

Example 2.12. Let

$$x = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

with normal form

$$y = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The $\mathbb{Z}$-span of the rows of x is clearly a direct summand. If $gxg^{-1} = y$ then also $gx^2 = y^2 g$, but

$$x^2 = \begin{bmatrix} 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad y^2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and so a Jordan normal form over $\mathbb{Z}$ does not exist.

Theorem 2.10 follows along the lines of the standard proofs in the case of a vector space over a field. For completeness we present such a proof below, but for both this proof and the applications of the theorem it is more convenient to work with linear transformations than matrices.

Let R be either $\mathbb{F}_q$ or $\mathbb{Z}$ and V a free R-module of finite rank n. If $A : V \to V$ is an R-homomorphism, then it gives rise to an $R[T]$-module structure on V, where $R[T]$ is the polynomial ring over R, and

$Tv := A(v)$. If needed we will denote these $R[T]$-modules by V_A to distinguish modules corresponding to different transformations.

We will be interested in the situation when A is nilpotent: $A^n = 0$. If $v \in V$ let k be minimal such that $A^k v = 0$ and let

$$\langle v \rangle = Rv + R(Av) + \cdots + R(A^{k-1}v) \tag{24}$$

denote the cyclic submodule generated by v. In this case we will call V cyclic if $V_A = \langle v \rangle$ for some $v \in V_A$. This happens exactly when the $R[T]$ module V_A is isomorphic to $\mathcal{C}_n = R[T]/(T^n)$. If R is a field, then any V_A is a direct sum of cyclic modules, but this fails for $R = \mathbb{Z}$, and in general when $R[T]$ is not a principal ideal domain.

Proof of Theorem 2.10. The theorem is equivalent to the following statement: if $V \simeq \mathbb{Z}^n$, and $A : V \to V$ is a nilpotent homomorphism, then the $\mathbb{Z}[T]$ module V_A is a direct sum of cyclic modules if and only if $A^k(V)$ is a direct summand of V (as an abelian group) for all $k \in \mathbb{N}$.

By the structure theorem for finitely generated abelian groups a subgroup V' of V is a direct summand if and only if for $k \in \mathbb{Z}$, $v \in V$, $kv \in V'$ implies that $v \in V'$. This immediately shows that $\ker A$ is a direct summand of V, and $A(V) \cap \ker A$ is a direct summand of $\ker A$. Let V_0 be a complementary direct summand so that

$$\ker A = V_0 \oplus (A(V) \cap \ker A). \tag{25}$$

Since A is nilpotent, the rank of $A(V)$ is strictly less than that of V. The condition on A descends to $A(V)$, and so by induction we have that $A : A(V) \to A(V)$ has a cyclic basis, i.e., there are $v_1, \ldots, v_l$ such that $A(V) \simeq \langle Av_1 \rangle \oplus \cdots \oplus \langle Av_l \rangle$. If we let d_i be the smallest integer k such that $A^k v_i = 0$, then we have that the set

$$\{A^j v_i \mid i = 1, \ldots, l, \ j = 1, \ldots, d_i - 1\} \tag{26}$$

is a basis of the free abelian group $A(V)$.

Let $v_{l+1}, \ldots, v_r$ be such that $V_0 = \bigoplus_{i=l+1}^r \mathbb{Z} v_i$, where V_0 is as in (25). Extending the notation from above we let $d_i = 1$ for $i = l+1, \ldots, r$.

We claim that

$$V_A \simeq \bigoplus_{j=1}^r \langle v_j \rangle.$$

We need to prove that for each $v \in V$, there is a unique choice of $\alpha_{i,j} \in \mathbb{Z}$ such that

$$v = \sum_{i=1}^r \sum_{j=0}^{d_i-1} \alpha_{i,j} A^j v_i. \tag{27}$$

To see uniqueness assume that $v = 0$ is expressed this way. Then $Av = 0$ as well, and the linear independence of the set in (26) shows that $\alpha_{i,j} = 0$ for all $i = 1, \ldots, l$ and $j = 0, \ldots, d_i - 2$. Since by (25) we have that $A^{d_i-1}v_i$, for $i = 1, \ldots, l$, and $v_{l+1}, \ldots, v_r$ are linearly independent, this shows that $\alpha_{i,d_i-1} = 0$ as well for all $i = 1, \ldots, r$.

It remains to show that every $v \in V$ can be expressed as an integral linear combination as in (27). By (26) this is clearly true for Av:

$$Av = \sum_{i=1}^{l} \sum_{j=0}^{d_i-2} \alpha_{i,j} A^i (Av_j)$$

for some $\alpha_{i,j} \in \mathbb{Z}$. Let $v' = \sum_{i=1}^{l} \sum_{j=0}^{d_i-2} \alpha_{i,j} A^i v_j$. Then $v - v' \in \ker A$; thus proving the claim. $\qquad\square$

2F. *The proof of Theorem 1.3.* The proof of Theorem 1.3 relies on the following two propositions.

Proposition 2.13. *Assume that $n \geq 2$, $\alpha \in \mathbb{F}_q^*$, $a = \alpha I + \bar{a}(\lambda)$ with $\varepsilon_j \in \{0, 1\}$ as in (22) with $\varepsilon_{n-1} = 1$. Also assume that $z_1 = \sum_{j=i+1}^{l-1} \xi_j \boldsymbol{e}_{k-1, N_j - 1} + \xi_l \boldsymbol{e}_{k-1, n_2}$, and that*

$$z_2 = \sum_{j=i+1}^{l-1} \eta_j \boldsymbol{e}_{k-1, N_j - 1} + \eta_l \boldsymbol{e}_{k-1, n_2}, \quad \text{where } \eta_j = \begin{cases} 0 & \text{if } \xi_j = 0, \\ 1 & \text{if } \xi_j \neq 0. \end{cases}$$

Then $a'' + z_1$ and $a'' + z_2$ are conjugate in $\mathrm{GL}_{n-2}(\mathbb{F}_q)$, and so $K_{n-2}(a'' + z_1) = K_{n-2}(a'' + z_2)$.

As a corollary of Proposition 2.13 one immediately has that for $k = N_i$ as above

$$\sum_{x \in X_k} \psi(ax + x^{-1}) = q^{2n-2} \sum_{z \in Z_0} (q-1)^{J+1} (K_{n-2}(a'' + z) - K_{n-2}(a'' + z + \boldsymbol{e}_{k-1, n-2})), \qquad (28)$$

where $a'' = a''_{\not k, \not \eta}$, and where z runs over

$$Z_0 = \left\{ z = \sum_{j=i+1}^{l-1} \eta_j \boldsymbol{e}_{k-1, N_j - 1} \,\middle|\, \eta_j \in \{0, 1\} \text{ for } i+1 \leq j \leq l \right\}, \qquad (29)$$

with $J = J(z) = |\{j \mid \eta_j = 1\}|$.

This in itself proves a version of Theorem 1.3 valid for almost all primes. The matrices $a'' + z$, $a'' + z + \boldsymbol{e}_{k-1, n-2}$ can obviously be lifted to $\mathbb{Z}$ where they can be put into Jordan normal form after a rational change of basis. This shows there are well-defined partitions $\lambda''(z)$, $\lambda''(z + \boldsymbol{e}_{k, n-2})$ of $n - 2$ associated to the partition λ, the value $k = N_i$ and z which are independent of p except for the finitely many primes dividing the determinant of the change of base matrix. The next proposition shows that such exceptions do not arise.

Proposition 2.14. *Let $\bar{a} \in M_n(\mathbb{Z})$ be as in Proposition 2.13 and $z \in Z_0$ be as in (29). There exists a unique partition $\lambda'' \vdash n - 2$ such that for any prime p and finite field $\mathbb{F}_q$ of characteristic p the partition $\lambda''(z) \vdash n - 2$ associated to the nilpotent matrix $\bar{a}'' + z$ equals λ''.*

Similarly there exist a partition $\mu'' \vdash n - 2$ such that the block partition of the matrix $\bar{a}'' + z + \boldsymbol{e}_{k-1, n-2}$ is μ''.

Proof of Proposition 2.13. Again we will use the language of linear transformations. Let V be a finite-dimensional vector-space over $\mathbb{F}_q$, and $A : V \to V$ a nilpotent linear transformation such that

$$V_A \simeq \langle v_0 \rangle \oplus \langle v_1 \rangle \oplus \cdots \oplus \langle v_l \rangle.$$

Let $d_i = \dim \langle v_i \rangle$. We are interested in perturbations $A + Z_1$, $A + Z_2$ of A, where Z_1, Z_2 are such that

$$Z_1(A^j v_i) = Z_2(A^j v_i) = 0 \quad \text{for } i = 1, \ldots, l, \, j = 0, \ldots, d_i - 1,$$

$$Z_1(A^j v_0) = Z_2(A^j v_0) = 0 \quad \text{for } j = 0, \ldots, n_0 - 2,$$

and

$$Z_1(A^{d_0-1} v_0) = \sum_{i=1}^{l} \xi_i A^{d_i-1} v_i \quad \text{and} \quad Z_2(A^{d_0-1} v_0) = \sum_{i=1}^{l} \eta_i A^{d_i-1} v_i,$$

where $\eta_i = 0$ or 1 depending on whether $\xi = 0$ or not.

Let $\phi : V_A \to V_A$ be the A-linear ($\phi \circ A = A \circ \phi$) isomorphism for which

$$\phi(v_i) = \begin{cases} \frac{1}{\xi_i} v_i & \text{for } 1 \leq i \leq l, \, \xi_i \neq 0, \\ v_i & \text{for } 1 \leq i \leq l, \, \xi_i = 0. \end{cases}$$

Then $\phi(Z_1(v)) = Z_2(\phi(v))$ as well, showing that the modules V_{A+Z_1} and V_{A+Z_2} are isomorphic. $\qquad\square$

Proof of Proposition 2.14. The statement is only meaningful if $n > 2$. Lift the matrix $\bar{a}''$ which only has entries 0 and 1 to a matrix $\tilde{a}$ over $\mathbb{Z}$ by lifting $0_{\mathbb{F}_q}$ to $0_{\mathbb{Z}}$ and $1_{\mathbb{F}_q}$ to $1_{\mathbb{Z}}$. Identify $\mathbb{Z}^{n-2}$ with $1 \times (n-2)$ matrices (row vectors) and let $A : \mathbb{Z}^{n-2} \to \mathbb{Z}^{n-2}$ be the linear transformation

$$v \mapsto v\tilde{a}.$$

In a similar fashion we may lift the matrix z or $z + e_{k-1,n-2}$ to a linear transformation $Z : \mathbb{Z}^{n-2} \to \mathbb{Z}^{n-2}$.

In both cases a simple change of the standard generators shows that $(A + Z)(\mathbb{Z}^{n-2}) = A(\mathbb{Z}^{n-2})$ is a direct summand. One also has that $AZ = Z^2 = 0$ and so

$$(A + Z)^k = (A + Z)A^{k-1}.$$

It also follows that $(A + Z)^k(\mathbb{Z}^{n-2})$ is a direct summand and by Theorem 2.10 has a Jordan normal form over $\mathbb{Z}$. $\qquad\square$

Proof of Theorem 1.3. Assume that $a = \alpha I + \bar{a}$, where $\bar{a} = \bar{a}(\lambda)$ as in (22). By (12),

$$K_n(a) = \sum_{k=1}^{n} \sum_{x \in X_k} \psi(ax + x^{-1}).$$

It is enough to prove that each of the sums $\sum_{x \in X_k} \psi(ax + x^{-1})$ is expressible as a polynomial in q, $q - 1$ and $K = K_1(\alpha)$. This was already established when $\bar{a} = 0$, so we may assume that $\varepsilon_{n-1} \neq 0$.

Proposition 2.5 and (i) of Theorem 2.9 take care of the cells X_n and X_{n-1}. For the other cells we may refer to statements (ii) and (iii) in Theorem 2.9 which reduce the sum down to the case when $k = N_i$ for one of the $N_i = n_1 + \cdots + n_i$, in which case Proposition 2.13 gives (28). Induction on the rank and Proposition 2.14 then shows that the resulting Kloosterman sums of rank $n - 1$ and $n - 2$ can be expressed as polynomials in q, $q - 1$ and K independently of $p = \operatorname{char} \mathbb{F}_q$. $\qquad\square$

It is possible to make this recursion more concrete; see Section 5B for examples.

3. Estimates

3A. *Kloosterman sums over Bruhat cells: reduction to involutions.* We will move on to setup the technical background for the proof of Theorem 1.6. We will pursue a path which establishes both of the estimates in Theorem 1.6 as well as Theorem 1.5 simultaneously and which will also allow us to analyze these sums cohomologically.

This approach is based on the Bruhat decomposition of the algebraic group GL_n as well as its specialization in the finite group $G = \mathrm{GL}_n(\mathbb{F}_q)$ with respect to the Borel subgroup B_n of upper triangular matrices, with

$$B_n = T_n U_n,$$

where U_n is the algebraic subgroup of upper triangular unipotent matrices, and T_n is the maximal torus (diagonal matrices). Since in this section n is fixed we will simply write B, T and U. Let W be the Weyl group of GL_n. We then have

$$G = \mathrm{GL}_n(\mathbb{F}_q) = \bigsqcup_{w \in W} C_w(\mathbb{F}_q), \quad C_w(\mathbb{F}_q) = U^\flat(\mathbb{F}_q) w B(\mathbb{F}_q), \tag{30}$$

with $U^\flat = \{u \in U \mid w^{-1}uw \in U^T\}$, where U^T is the unipotent subgroup of the opposite Borel subgroup of lower triangular matrices. It is clear from Gaussian elimination that for a field F all $x \in C_w(F)$ have a unique decomposition $x = uwb$ with $u \in U^\flat(F)$ and $b \in B(F)$. This remains true for the algebraic variety GL_n canonically [Springer 1998, Chapter 8]. (Alternatively, given any field F one may work over an algebraic closure of F, say $\bar{F}$, as in [Borel 1991, IV.14.12, Theorem (a)]. It is clear from the uniqueness that once representatives for W as permutation matrices are fixed, the decomposition above is invariant under any Galois automorphism of $\bar{F}$ fixing F. The same argument shows that the map $U^\flat \times B \to C_w$ is an isomorphism of algebraic varieties, defined over F.)

We will work with

$$K_n^{(w)}(a, \mathbb{F}_q) = \sum_{x \in C_w(\mathbb{F}_q)} \psi(ax + x^{-1}). \tag{31}$$

As a first step we will prove that the above sum vanishes unless $w^2 = I$ and analyze the cells C_w to simplify these sums for $w^2 = I$.

The Weyl group, $W = W_n$, is isomorphic to S_n, the permutation group on n letters. For later calculations we make this identification explicit as follows. For a permutation matrix w we associate the permutation π such that $i = \pi(j)$ if $w_{i,j} = 1$. Conversely given π, we let

$$w_\pi = \sum_{j=1}^n e_{\pi(j),j} \tag{32}$$

so that $w_{\pi_1} w_{\pi_2} = w_{\pi_1 \pi_2}$. To ease reading the arguments that will follow, we overload the notation and write $w(i)$ for $\pi(i)$ if $w = w_\pi$. This convention leads to

$$(gw)_{ij} = g_{i,w(j)} \quad \text{and} \quad (wg)_{ij} = g_{w^{-1}(i),j} \tag{33}$$

for any $g \in G$, which will be frequently used without further mention.

Proposition 3.1. *Let $a \in M_n(\mathbb{F}_q)$ be an upper triangular matrix, such that* $\det a \neq 0$, *and assume $w \in W$ is such that* $w^2 \neq I$. *Then*

$$K_n^{(w)}(a, \mathbb{F}_q) = 0.$$

Proof. Let i be minimal such that $i \neq w^2(i)$ and let $j = w(i)$. If $k = w(j)$ then $w^2(k) \neq k$ and so $k > i$.

Consider now the one parameter subgroup

$$\{x_{i,j}(s) = I + s e_{i,j} : s \in \mathbb{F}_q\} \subset B.$$

Clearly we have $x_{ij}(s) B(\mathbb{F}_q) = B(\mathbb{F}_q)$ and so

$$K_n^{(w)}(a, \mathbb{F}_q) = \sum_{\substack{u \in U^\flat(\mathbb{F}_q) \\ b \in B(\mathbb{F}_q)}} \psi(auwb + (uwb)^{-1}) = \frac{1}{q} \sum_{s \in \mathbb{F}_q} \sum_{\substack{u \in U^\flat(\mathbb{F}_q) \\ b \in B(\mathbb{F}_q)}} \psi(auwx_{i,j}(s)b + (uwx_{i,j}(s)b)^{-1}).$$

Note that

$$\frac{\partial}{\partial s} \operatorname{Tr}(auwx_{i,j}(s)b) = \operatorname{Tr}(bauwe_{i,j}) = (bauw)_{j,i} = (bau)_{j,j}$$

since $j = w(i)$. By the assumption $\det a \neq 0$ we have $(bau)_{j,j} \neq 0$.

On the other hand

$$\frac{\partial}{\partial s} \operatorname{Tr}(b^{-1}x_{i,j}(-s)w^{-1}u^{-1}) = -\operatorname{Tr}(e_{i,j}w^{-1}u^{-1}b^{-1}) = -(w^{-1}u^{-1}b^{-1})_{j,i} = -(u^{-1}b^{-1})_{w(j),i} = 0$$

since $w(j) > i$. Since $\psi(b^{-1}x_{i,j}(-s)w^{-1}u^{-1})$ is linear in s, it is equal to $\psi(b^{-1}w^{-1}u^{-1})$ for any s.

Writing $\psi(auwx_{i,j}(s)b) = \psi(auwb)\psi(auwe_{i,j}bs)$ and using that ψ is invariant under conjugation we have $\psi(auwe_{i,j}bs) = \psi(bauwe_{i,j}s) = \psi((bau)_{j,j}s)$. This gives

$$\sum_{s \in \mathbb{F}_q} \psi(auwx_{i,j}(s)b + b^{-1}x_{i,j}(-s)w^{-1}u^{-1}) = \psi(auwb + b^{-1}w^{-1}u^{-1}) \sum_{s \in \mathbb{F}_q} \psi((bau)_{j,j}s) = 0. \quad \square$$

We can also prove Theorems 1.1 and 1.2 this way. For example, for nilpotent matrices we may prove that $K_n^{(w)}(a) = 0$, unless $w = I$. First we may assume that a is strictly upper triangular.

Let $j = \max(k \mid w(k) \neq k)$ and $i = w(j) < j$ and consider the one parameter subgroup

$$\{x_{i,j}(s) = I + s e_{i,j} \mid s \in \mathbb{F}_q\} \leq B(\mathbb{F}_q).$$

We have $x_{i,j}(s) B(\mathbb{F}_q) = B(\mathbb{F}_q)$ and so

$$K_n^{(w)}(a) = \sum_{\substack{u \in U_w(\mathbb{F}_q) \\ b \in B(\mathbb{F}_q)}} \psi(auwb + (uwb)^{-1}) = \frac{1}{q} \sum_{s \in \mathbb{F}_q} \sum_{\substack{u \in U_w(\mathbb{F}_q) \\ b \in B(\mathbb{F}_q)}} \psi(auwx_{i,j}(s)b + (uwx_{i,j}(s)b)^{-1}).$$

By the definition of j for any $u \in U_w(\mathbb{F}_q)$ and $b \in B(\mathbb{F}_q)$

$$\frac{\partial}{\partial s} \operatorname{tr}(auwx_{i,j}(s)b) = \operatorname{tr}(e_{i,j}bauw) = 0.$$

On the other hand

$$\frac{\partial}{\partial s} \operatorname{tr}((uwx_{i,j}(s)b)^{-1}) = \operatorname{tr}(\boldsymbol{e}_{i,j} w^{-1} u^{-1} b^{-1}) = b_{i,i}^{-1} \neq 0.$$

This gives

$$\sum_{s \in \mathbb{F}_q} \psi(auwx_{i,j}(s)b + b^{-1}x_{i,j}(-s)w^{-1}u^{-1}) = \psi(auwb + b^{-1}w^{-1}u^{-1}) \sum_{s \in \mathbb{F}_q} \psi(b_{i,i}^{-1}s) = 0.$$

So

$$K_n(a, \mathbb{F}_q) = K_n^{(I)}(a, \mathbb{F}_q) = \sum_{b \in B(\mathbb{F}_q)} \psi(ab + b^{-1}) = \sum_{b \in B(\mathbb{F}_q)} \psi(b^{-1}) = |U| \sum_{t \in T(\mathbb{F}_q)} \prod_{i=1}^{n} \varphi(t_{i,i}^{-1}) = (-1)^n q^{n(n-1)/2}$$

as $\psi(b^{-1}) = \psi(t^{-1}) = \prod_{i=1}^{n} \varphi(t_{i,i}^{-1})$ if $b = tu \in TU = B$.

3B. *Finer decomposition of individual Bruhat cells.* We will give a decomposition of the algebraic group U of unipotent upper triangular matrices in GL_n. To show that the underlying maps are morphisms we will work over a general commutative ring R. Therefore the letter U will denote the algebraic group itself, and not its set of points $U(\mathbb{F}_q)$. Similarly further subsets of U denoted with various markings will define affine subvarieties of U and not their set of points in $\mathbb{F}_q$.

The motivation for this refinement of the Bruhat decomposition is as follows. The Bruhat cells $C_w(\mathbb{F}_q) = U^\flat(\mathbb{F}_q) w B(\mathbb{F}_q)$ are already of the form $\mathbb{F}_q^d \times (\mathbb{F}_q^*)^e$ when using the entries of the matrices in $U^\flat(\mathbb{F}_q)$, $T(\mathbb{F}_q)$ and $U(\mathbb{F}_q)$ as coordinates, and exponential sums over such spaces have a well-established theory [Adolphson and Sperber 1989; Denef and Loeser 1991]. Also, $\operatorname{tr}(ax + x^{-1})$ as a function on $\mathbb{F}_q^d$ (using the entries as coordinates of x) is degree 1 in each of these variables. However they have to be collected the right way to use this observation. Therefore we will fiber up the cells C_w into finer (affine) subspaces that are more suitable for this purpose. For later arguments that rely on cohomology it will be important that this refinement gives an isomorphism of affine varieties. While for this purpose one could restrict to $\mathbb{F}_q$-algebras, the results are true over an arbitrary commutative ring. Also this refinement has potential applications elsewhere and so we first set it up for a general element $w \in W$ which is assumed to be fixed.

This section mainly consists in developing a nomenclature for these simple but numerous subsets (many of them subgroups). These subsets are defined under various actions of W on the affine variety $\overline{U}$ of strictly upper triangular nilpotent matrices, whose set of points in any ring is

$$\overline{U}(R) = \{y \in M_n(R) \mid I + y \in U(R)\} \tag{34}$$

Clearly $\overline{U}$ is isomorphic to $\mathbb{A}^{n(n-1)/2}$ as an algebraic variety.

The subsets we are going to define depend on w but we will not work with more than one w at a time, so we drop w from the notation. For example, the subgroups

$$U^\flat = \{u \in U \mid w^{-1}uw \in U^T\} \quad \text{and} \quad U^\sharp = \{u \in U \mid w^{-1}uw \in U\}$$

that satisfy $U^\sharp \cap U^\flat = \{I\}$ and $U = U^\sharp U^\flat = U^\flat U^\sharp$ can be defined as

$$U^\flat = I + \overline{U}^\flat \quad \text{and} \quad U^\sharp = I + \overline{U}^\sharp$$

where

$$\overline{U}^\flat = \overline{U} \cap w \overline{U}^T w^{-1} \quad \text{and} \quad \overline{U}^\sharp = \overline{U} \cap w \overline{U} w^{-1}.$$

The further refinement comes from exploiting the action of W, the group of permutation matrices on M_n via left multiplication and so it is tied to the standard representation of GL_n.

Definition 3.2. For a fixed $w \in W$ let

$$\overline{U}_a = \overline{U} \cap w^{-1}\overline{U}, \quad \overline{U}_b = \overline{U} \cap w^{-1}\overline{U}^T \quad \text{and} \quad \overline{U}_o = \overline{U} \cap w^{-1}D, \tag{35}$$

where D is the set of possibly singular diagonal matrices. Using these subvarieties define

$$U_a = I + \overline{U}_a, \quad U_b = I + \overline{U}_b \quad \text{and} \quad U_o = I + \overline{U}_o. \tag{36}$$

Remark 8. The subscripts "a", "b" and "o" denote the fact that these are the elements $\bar{u} \in \overline{U}$ for which the nonzero entries in $w\bar{u}$ are strictly "above", "below" or "on" the diagonal. Clearly, for any ring R, $\overline{U}(R) = \overline{U}_a(R) \oplus \overline{U}_b(R) \oplus \overline{U}_o(R)$.

We will see below that the map $x \mapsto \mathrm{tr}(ax + x^{-1})$ is linear on $\overline{U}_a$ which leads to immediate cancellations. However for this purpose we first need to setup further notation.

Definition 3.3. Let

$$U_a^\sharp = U^\sharp \cap U_a, \quad U_b^\sharp = U^\sharp \cap U_b, \quad U_a^\flat = U^\flat \cap U_a, \quad U_b^\flat = U^\flat \cap U_b.$$

Remark 9. These subvarieties are defined via their nonvanishing entries. In general let

$$\mathcal{I} = \{(i, j) \mid 1 \leq i < j \leq n\},$$

and for $\mathcal{J} \subset \mathcal{I}$ define the affine variety

$$\overline{U}_{\mathcal{J}} = \{\bar{u} \in \overline{U} \mid \bar{u}_{i,j} \neq 0 \implies (i, j) \in \mathcal{J}\}.$$

Note that $U_{\mathcal{J}} = I + \overline{U}_{\mathcal{J}}$ is an algebraic subgroup of U if and only if $\mathcal{J}$ is transitive in the sense that

$$(i, j), (j, k) \in \mathcal{J} \implies (i, k) \in \mathcal{J}.$$

For example, $U_a = U_{\mathcal{J}_a}$ where $\mathcal{J}_a = \{(i, j) \in \mathcal{I} \mid w(i) < j\}$, which is transitive, and so U_a is a subgroup.

Example 3.4. Consider the case $n = 5$. We indicate below the indices with the different properties for two involutions:

$$w = (14)(25) \text{ gives } \begin{bmatrix} b & b & o & a \\ & b & b & o \\ & & a & a \\ & & & a \end{bmatrix}, \quad \text{while } w = (15)(23) \text{ gives } \begin{bmatrix} b & b & b & o \\ & o & a & a \\ & & a & a \\ & & & a \end{bmatrix}.$$

To state the main result of this section we need to introduce one more piece of notation. Let

$$\bar{U}_{b/o} = \bar{U}_b \oplus \bar{U}_o$$

be the subset of $\bar{U}$ of elements $\bar{u}$ for which $w\bar{u}$ has nonzero elements only below or on the diagonal, and let

$$U_{b/o}^\sharp = I + \bar{U}_{b/o}^\sharp \quad \text{and} \quad U_{b/o}^\flat = I + \bar{U}_{b/o}^\flat,$$

where $\bar{U}_{b/o}^\sharp = \bar{U}^\sharp \cap \bar{U}_{b/o}$, and $\bar{U}_{b/o}^\flat = \bar{U}^\flat \cap \bar{U}_{b/o}$.

Proposition 3.5. (i) U_a, $U^\sharp U_a$ and $U_{b/o}^\flat$ are algebraic subgroups of U and $U^\sharp U_a \cap U_{b/o}^\flat = \{I\}$.

(ii) *The morphism*

$$U_{b/o}^\sharp \times U_a \times U_{b/o}^\flat \to U, \quad (u_1, y, u_2) \mapsto u_1 y u_2,$$

is an isomorphism.

(iii) *If $w^2 = I$, then $U_o \subset U^\flat$ is a subgroup, and the morphism*

$$U_b^\sharp \times U_a \times U_o \times U_b^\flat \to U, \quad (u_1, y, u_o, u_2) \mapsto u_1 y u_o u_2,$$

is an isomorphism.

There are many alternative versions of the statement in (i) as, for example, $U_{a/o} = U_a U_o = U_o U_a$ is a subgroup of U as well. Also if one is merely interested in a bijection over $\mathbb{F}_q$, the statements in (ii) and (iii) can easily be proved by a counting argument. One may argue similarly to see that the set $U_{b/o}^\sharp$ is a complement of U_a in $U^\sharp U_a$. To prove that these maps are isomorphisms it is possible to adapt the reasoning of Lemma 8.2.2 in [Springer 1998] but given the concrete nature of the statement we give a self-contained proof here, based on the fact that affine varieties and their maps are determined by their functor of points.

Lemma 3.6. *We have the following*:

 (i) *If $u \in U(R)$ and $x \in \bar{U}_a(R)$ then $xu \in \bar{U}_a(R)$.*

 (ii) *If $u \in U^\sharp(R)$ and $x \in \bar{U}_a(R)$ then $ux \in \bar{U}_a(R)$.*

(iii) *If $u_1 \in U^\sharp(R)$ and $u_2 \in U^\flat(R)$ then*

$$u_1 u_2 + \bar{U}_a(R) = u_1 U_a(R) u_2.$$

Proof. Since $\bar{U}_a(R) = \bar{U}(R) \cap w^{-1}\bar{U}(R)$, the first and second claims are obvious from the fact that $U(R)\bar{U}(R) = \bar{U}(R)U(R) = \bar{U}(R)$, and that for $u \in U^\sharp(R)$, $w^{-1}uw \in U(R)$.

By the first two claims, if $u_1 \in U^\sharp(R)$ and $u_2 \in U^\flat(R)$ then $u_1\bar{U}_a(R)u_2 = \bar{U}_a(R)$, from which the third claim is obvious. $\qquad\square$

Proof of Proposition 3.5. It is enough to prove that for any commutative ring R the sets $U_a(R)$, $U^\sharp(R)U_a(R)$ and $U_{b/o}^\flat(R)$ are subgroups of $U(R)$, and that the maps

$$U_{b/o}^\sharp(R) \times U_a(R) \times U_{b/o}^\flat(R) \to U(R), \quad (u_1, y, u_2) \mapsto u_1 y u_2,$$

and

$$U_b^\sharp(R) \times U_a(R) \times U_o(R) \times U_b^\flat(R) \to U(R), \quad (u_1, y, u_o, u_2) \mapsto u_1 y u_o u_2,$$

are bijections.

Define on $U(R)$ the equivalence relation

$$u_1 \sim_{\overline{U}_a} u_2 \iff u_1 - u_2 \in \overline{U}_a(R). \tag{37}$$

We start with the proof of the claim about the group property of the three sets in (i). By (i) of the previous lemma, if $u_1 \sim_{\overline{U}_a} u_2$, and $u \in U(R)$, then $u_1 u \simeq_{\overline{U}_a} u_2 u$. Therefore $U_a(R)$, the stabilizer of the equivalence class of the identity I, is a subgroup.

By (ii) of the same lemma $U^\sharp(R)$ normalizes $U_a(R)$ and so $U^\sharp(R)U_a(R)$ is a subgroup. Finally,

$$\mathcal{J} = \{(i, j) \mid j \le w(i), \; w(j) \le w(i)\} \quad \text{and} \quad \mathcal{J}' = \mathcal{I} \setminus \mathcal{J} = \{(i, j) \in \mathcal{I} \mid w(i) < j \text{ or } w(i) < w(j)\},$$

are disjoint transitive subsets. This shows that $U_{b/o}^\flat(R) = U_{\mathcal{J}}(R)$ is a subgroup, and since $U^\sharp(R)$, $U_a(R) \subset U_{\mathcal{J}'}$ we have that $U^\sharp(R)U_a(R) \cap U_{b/o}^\flat(R) = \{I\}$.

Now to prove the claim that every $u \in U(R)$ can be represented in a unique way as

$$u = u_1 y u_2, \quad u_1 \in U_{b/o}^\sharp(R), \;\; y \in U_a(R), \;\; u_2 \in U_{b/o}^\flat(R),$$

we will first show that $U(R) = U_{b/o}^\sharp(R)U_a(R)U_{b/o}^\flat(R)$. Let $u \in U(R)$. We know [Springer 1998, Chapter 8, Proposition 8.2.1] that there are $v_1 \in U^\sharp(R)$ and $v_2 \in U^\flat(R)$ such that $u = v_1 v_2$.

Let $u_1 \in U_{b/o}^\sharp(R)$ be the matrix whose entries agree with v_1 for $(i, j) \in \mathcal{I}$ when $w(i) \ge j$ and are 0 otherwise. Then we have that

$$v_1 \in u_1 + \overline{U}_a(R) \quad \text{and so} \quad v_1 U_a(R) = u_1 U_a(R).$$

Similarly let $u_2 \in U_b^\flat(R)$ be such that $v_2 \in u_2 + \overline{U}_a(R)$, so that $U_a(R)u_2 = U_a(R)v_2$. We have that

$$u = v_1 v_2 \in v_1 U_a(R)v_2 = u_1 U_a(R)u_2.$$

Now for the injectivity assume that $u_1, u_1' \in U_{b/o}^\sharp(R)$, $y, y' \in U_a(R)$, and $u_2, u_2' \in U_{b/o}^\flat(R)$ are such that

$$u_1 y u_2 = u_1' y' u_2'.$$

Since $U^\sharp(R)U_a(R) \cap U_{b/o}^\flat(R) = \{I\}$ we have that $u_2 = u_2'$ and so $u_1 y = u_1' y'$. By (iii) of the previous lemma $u_1 \sim_{\overline{U}_a} u_1'$ which can only happen in $U_{b/o}^\sharp(R)$ if $u_1 = u_1'$.

The very last claim about the case $w^2 = I$ is elementary. $\qquad\square$

For convenience we will parameterize C_w using the following lemma.

Lemma 3.7. *Assume that $w^2 = I$. Then $U_o = U_o^\flat$ and the morphisms*

(i) $U^\flat \times T \times U \to C_w, (v, t, u) \mapsto vwtuv^{-1}$,

(ii) $U_b^\sharp \times \overline{U}_a \times U_o^\flat \times U_b^\flat \to U, (u_1, y, u_o, u_2) \mapsto u_1(I + y)^{-1}u_o u_2$,

are isomorphisms.

Proof. The first claim follows from (30) and the fact that the morphism $U^\flat \times T \times U \to U^\flat \times T \times U$, $(v, t, u) \mapsto (v, t, uv^{-1})$ is an isomorphism of varieties. The second claim is merely a restatement of (iii) of Proposition 3.5. $\square$

The advantage of the parameterization in (i) is obvious: if $x = vwtuv^{-1} \in C_w(\mathbb{F}_q)$ then

$$\mathrm{tr}(ax + x^{-1}) = \mathrm{tr}(a^v wtu + (wtu)^{-1}),$$

where $a^v = v^{-1}av = \alpha I + \bar{a}^v$, with $\bar{a}^v = v^{-1}\bar{a}v$ still strictly upper triangular.

3C. *Some trace calculations.* From now on we will specify to the finite field $\mathbb{F}_q$ and so B, U, $\overline{U}$ etc. will mean $B(\mathbb{F}_q)$, $U(\mathbb{F}_q)$, $\overline{U}(\mathbb{F}_q)$ etc. Assume again that a is a matrix with a unique eigenvalue $\alpha \in \mathbb{F}_q^*$:

$$a = \alpha I + \bar{a} \in M_n(\mathbb{F}_q), \quad \text{where } \bar{a} \text{ is strictly upper triangular.} \tag{38}$$

With (ii) of Lemma 3.7 it is now easy to bound $\sum_{x \in C_w} \psi(ax + x^{-1})$. However the cohomological methods used in the proofs of Theorems 1.7 and 1.8 require repeated use of Lemma 3.16, which in turn relies on understanding

$$x \mapsto \mathrm{tr}(ax + x^{-1})$$

itself as a function on C_w. This is achieved in Propositions 3.8 and 3.9.

Proposition 3.8. *Assume that $w \in W$ satisfies $w^2 = I$ and let $\overline{U}_a$, $U_b^\sharp$, $U_b^\flat$, U_o be as in Definitions 3.2 and 3.3 with values in $\mathbb{F}_q$.*

Assume that $a^v = \alpha I + \bar{a}' \in M_n(\mathbb{F}_q)$ with $\alpha \in \mathbb{F}_q^*$, $\bar{a}' \in \overline{U}$, and that $u_1 \in U_b^\sharp$, $u_o \in U_o$ and $u_2 \in U_b^\flat$. If $y \in \overline{U}_a$, let $g(y) = u_1(I + y)^{-1}u_ou_2$. Then the function

$$y \mapsto \mathrm{tr}(a^v wtg(y) + (wtg(y))^{-1})$$

is affine linear on $\overline{U}_a$. This map is nonconstant unless $u_2 = I$.

When the map is constant its value is

$$\mathrm{tr}(a^v wtu_1u_o + (wtu_1u_o)^{-1}).$$

Further reductions are given by:

Proposition 3.9. *Assume $w \in W$ satisfies $w^2 = I$, and that $a' = \alpha I + \bar{a}' \in M_n(\mathbb{F}_q)$ with $\alpha \in \mathbb{F}_q^*$, $\bar{a}' \in \overline{U}$. Let $u_1 \in U_b^\sharp$ and $t = \mathrm{diag}(t_1, \ldots, t_n) \in T$. For $u_o = I + \bar{u}_o \in U_o$ the map*

$$u_o \mapsto \mathrm{tr}(a'wtu_1(I + \bar{u}_o) + (wtu_1(I + \bar{u}_o))^{-1})$$

is affine linear on U_0. This map is nonconstant unless $t_{w(i)}^{-1} = \alpha t_i$ for all $i \neq w(i)$. When the map is constant its value is

$$\sum_{i = w(i)} (\alpha t_i + t_i^{-1}) + \mathrm{tr}(\bar{a}'wdu_1),$$

where $d = \sum_{i < w(i)} t_i e_{i,i}$.

The proof of the above facts rely on some simple lemmas that we present first.

Lemma 3.10. *If a' is upper triangular, $y \in \bar{U}_a$, $u_1 \in U^\sharp$ and $u_2 \in U$, then*

$$\operatorname{tr}(a'wtu_1(I+y)^{-1}u_2) = \operatorname{tr}(a'wtu_1u_2) \tag{39}$$

is independent of y.

Proof. By Proposition 3.5 $(I+y)^{-1} = I + y'$ for some $y' \in \bar{U}_a$ and so wy' is strictly upper triangular. By the assumptions on a', u_1 and u_2 we have that a', wtu_1w^{-1} and u_2 are upper triangular, so

$$\operatorname{tr}(a'wtu_1y'u_2) = \operatorname{tr}(a'(wtu_1w^{-1})(wy')u_2) = 0. \qquad \square$$

Lemma 3.11. *For any $u \in U$, $t \in T$*

$$y \mapsto wyu^{-1}t^{-1}$$

is a linear automorphism of $\bar{U}_a$.

Proof. The map $y \mapsto wyu^{-1}t^{-1}$ is clearly linear, and a bijection to its image. Since $w\bar{U}_a = \bar{U}_a$ and $\bar{U}_aut = \bar{U}_a$ this image is $\bar{U}_a$. $\qquad \square$

Proof of Proposition 3.8. First by Lemma 3.10 we have

$$\operatorname{tr}(a^v wtg(y) + (wtg(y))^{-1}) = \operatorname{tr}(a^v wtu_1u_ou_2) + \operatorname{tr}((wtu_1u_ou_2)^{-1}) + \operatorname{tr}((u_2u_o)^{-1}yu_1^{-1}t^{-1}w)$$

showing instantly that the map $y \mapsto \operatorname{tr}(a^v wtg(y) + (wtg(y))^{-1})$ is affine linear. Also

$$\operatorname{tr}((u_2u_o)^{-1}yu_1^{-1}t^{-1}w) = \operatorname{tr}(w(u_2u_o)^{-1}wy'),$$

where $y' = wyu_1^{-1}t^{-1} \in U$, and so by Lemma 3.11 it is enough to show that the map

$$y' \mapsto \operatorname{tr}(w(u_2u_o)^{-1}wy')$$

is nonconstant unless $u_2 = I$.

Assume that $u_2 \neq I$, and let $z = (u_2u_o)^{-1}$. Since $U_b^\flat$ and $U_o^\flat$ are subgroups, z is in $U_b^\flat U_o^\flat$, but not in $U_o^\flat$ and so there exist (i, j) such that

$$i < j, \quad w(i) > j \quad \text{and} \quad z_{i,j} \neq 0.$$

Also, by the fact that $U_b^\flat \subset U^\flat$, we have that $w(i) > w(j)$, and so

$$e_{w(j),w(i)} \in \bar{U}_a.$$

Now let

$$Y_0 = \{y \in \bar{U}_a : y_{w(j),w(i)} = 0\}$$

so that any element $y \in \bar{U}_a$ may be written as $y = y_0 + se_{w(j),w(i)}$, where $y_0 \in Y_0$. Then

$$\operatorname{tr}(wzw(y_0 + se_{w(j),w(i)})) = \operatorname{tr}(wzwy_0) + s\,\operatorname{tr}(wzwe_{w(j),w(i)}) = \operatorname{tr}(wzwy_0) + sz_{i,j}.$$

This proves the proposition. $\qquad \square$

For the proof of Proposition 3.9 we need an explicit evaluation:

Lemma 3.12. *Assume* $w \in W$ *satisfies* $w^2 = I$, $\alpha \in \mathbb{F}_q^*$, $\bar{a}' \in \bar{U}$ *and that* $t = \mathrm{diag}(t_1, \ldots, t_n) \in T$ *satisfies* $t_{w(i)}^{-1} = \alpha t_i$ *for all* $i < w(i)$. *Then*

$$\mathrm{tr}(w(\alpha t + t^{-1})) = \sum_{i = w(i)} \alpha t_i + t_i^{-1}.$$

If $u \in U$, *then*

$$\mathrm{tr}(\bar{a}' wtu) = \mathrm{tr}(\bar{a}' wdu),$$

where $d = \sum_{i < w(i)} t_i e_{i,i}$.

Proof. First observe that

$$wt = \sum_{i = w(i)} t_i e_{i,i} + \sum_{i < w(i)} (t_i e_{w(i),i} + (\alpha t_i)^{-1} e_{i,w(i)}),$$

from which the first claim follows immediately. The second is a slight variant, using that $\bar{a}'_{i,i} = 0$, from the assumption $\bar{a}' \in \bar{U}$. $\square$

Proof of Proposition 3.9. It is easy to check that $U_o = U_o^\flat = \left\{ I + \sum_{i < w(i)} s_i e_{i,w(i)} \mid s_i \in \mathbb{F}_q \right\}$ and it is abelian. Using that we have that $(I + \bar{u}_o)^{-1} = I - \bar{u}_o$. Therefore

$$\mathrm{tr}(a' wtu_1(I + \bar{u}_o) + (wtu_1(I + \bar{u}_o))^{-1}) = \mathrm{tr}(a' wtu_1 + (wtu_1)^{-1}) + \mathrm{tr}((a' wtu_1 - (wtu_1)^{-1})\bar{u}_o)$$

is clearly affine linear as a function of $\bar{u}_o$. To see when it is nonconstant write $\bar{u}_o$ as $\sum_{i < w(i)} s_i e_{i,w(i)}$ so that by the conventions in (33)

$$\bar{u}_o w = \sum_{i < w(i)} s_i e_{i,i} \quad \text{and} \quad w\bar{u}_o = \sum_{i < w(i)} s_i e_{w(i),w(i)}.$$

This leads to

$$\mathrm{tr}(a' wtu_1 \bar{u}_o) = \mathrm{tr}((a' wtu_1 w)w\bar{u}_o) = \sum_{i < w(i)} s_i(a' wtu_1 w)_{w(i),w(i)} = \sum_{i < w(i)} s_i \alpha t_i$$

since $wu_1 w \in U$, and $w \, \mathrm{diag}(t_i) w = \mathrm{diag}(t_{w(i)})$. In a similar manner

$$\mathrm{tr}((wtu_1)^{-1}\bar{u}_o) = \mathrm{tr}((tu_1)^{-1}w\bar{u}_o) = \sum_{i < w(i)} s_i((tu_1)^{-1})_{w(i),w(i)} = \sum_{i < w(i)} s_i t_{w(i)}^{-1}.$$

Finally, since $u_1 \in U_b^\sharp \subset U^\sharp$, $u_1^{-1} \in U^\sharp$ as well, and for any $u \in U^\sharp$ we have

$$u_{i,w(i)} = 0 \quad \text{in the case } i < w(i).$$

Hence for $u = I + \bar{u} \in U^\sharp$, $\mathrm{tr}(wt\bar{u}) = 0$, and $\mathrm{tr}(tw\bar{u}) = 0$. This gives

$$\mathrm{tr}(a' wtu_1 + wu_1^{-1} t^{-1}) = \mathrm{tr}(\alpha wt + wt^{-1}) + \mathrm{tr}(\bar{a}' wtu_1).$$

This finishes the proof of the proposition in view of Lemma 3.12. $\square$

3D. *The proofs of Theorems 1.5 and 1.6.* We again interpret the notation for all affine varieties as the set of their $\mathbb{F}_q$-rational points, so C_w stands for $C_w(\mathbb{F}_q)$, etc. Recall that

$$K_n^{(w)}(a) = \sum_{x \in C_w} \psi(ax + x^{-1}).$$

Proposition 3.13. *Assume that $a = \alpha I + \bar{a} \in M_n(\mathbb{F}_q)$ with $\alpha \in \mathbb{F}_q^*$, $\bar{a} \in \overline{U}$. Then*

$$K_n^{(w)}(a) = q^{n_a} \sum_v \sum_{t, u_o, u_1} \psi(a^v w t u_1 u_o + (w t u_1 u_o)^{-1}),$$

where $v \in U^\flat$, $a^v = v^{-1} a v$, $t \in T$, $u_o \in U_o$, $u_1 \in U_b^\sharp$ and

$$n_a = \dim U_a = |\{\mathcal{J}_a\}| = \left| \{(i, j) \mid i, w(i) < j\} \right|.$$

Proof. By Lemma 3.7(ii)

$$K_n^{(w)}(a) = \sum_v \sum_{t, u_o, u_1, u_2} \sum_y \psi(a' w t g(y) + (w t g(y))^{-1}),$$

where $g(y) = u_1 (I + y)^{-1} u_0 u_2$ and the inner sum is

$$\sum_y \psi(a' w t g(y) + (w t g(y))^{-1}) = \begin{cases} 0 & \text{if } u_2 \neq I, \\ q^{n_a} \psi(a' w t u_1 u_o + (w t u_1 u_o)^{-1}) & \text{if } u_2 = I, \end{cases}$$

by Proposition 3.8. $\square$

Proposition 3.14. *We have*

$$K_n^{(w)}(a) = q^{n_a + n_o} K_1(\alpha)^f \sum_{v, d, u} \psi(\bar{a}^v w d u),$$

where $v \in U^\flat$, $u \in U_b^\sharp$, $\bar{a}^v = v^{-1} \bar{a} v$, $n_o = e = \left| \{i \mid i < w(i)\} \right|$ is the number of involution pairs in w, $f = \left| \{i \mid w(i) = i\} \right|$ is the number of fixed points of w and $d \in D(w) = \left\{ \sum_{i < w(i)} t_i e_{i,i} \mid t_i \in \mathbb{F}_q^* \right\}$.

Proof. Let

$$T(w) = \{t \in T \mid t_{w(i)}^{-1} = \alpha t_i \text{ if } i < w(i)\}.$$

By Proposition 3.9,

$$\sum_{u_o} \psi(a^v w t u_1 u_o + (w t u_1 u_o)^{-1}) = 0$$

unless $t \in T'(w)$, in which case

$$\sum_{u_o} \psi(a^v w t u_1 u_o + (w t u_1 u_o)^{-1}) = q^{n_o} \prod_{i = w(i)} \phi(\alpha t_i + t_i^{-1}) \psi(\bar{a}^v w d u_1),$$

with $d = \sum_{i < w(i)} t_i e_{i,i}$.

The proposition follows after summing over $t \in T(w)$. $\square$

Proof of Theorem 1.5. We will show that for $\alpha \neq 0$ and $w^2 = I$ we have

$$K_n^{(w)}(\alpha I) = q^{n(n-1)/2+N}(q-1)^e K_1(\alpha)^f, \tag{40}$$

where $N = n_{a/o}^{\flat} = \dim U_a^{\flat} + \dim U_o^{\flat}$.

Since $\bar{a} = 0$ now, $\bar{a}^v = v^{-1}\bar{a}v = 0$ as well, and so $\psi(\bar{a}^v w d u_1) = \psi(0) = 1$. To get the exponent of q note that $n_a + n_o + n^{\flat} + n_b^{\sharp} = n_a + n_o + n_b + n_{a/o}^{\flat}$, where these are denoting the dimension of the corresponding subspaces of U. Then we also have $n_a + n_o + n_b = n(n-1)/2$. $\qquad \square$

Proof of Theorem 1.6. If $a = \alpha I + \bar{a}$, $\bar{a} \in \bar{U}$, $w^2 = I$, then by Proposition 3.14

$$|K_n^{(w)}(a)| = q^{n_a+n_o}|K_1(\alpha)^f| \left| \sum_{v,d,u} \psi(\bar{a}^v w d u) \right| \leq q^{n_a+n_o}|K_1(\alpha)^f| \sum_{v,d,u} |\psi(\bar{a}^v w d u)| = |K_n^{(w)}(\alpha I)|$$

since $|\psi(\bar{a}^v w d u_1)| = 1$.

Since $w^2 = I$, every element is either fixed by w or is in an involution pair, and so $n = f + 2n_o$. So while f depends on w,

$$K_1(\alpha)^f = K_1(\alpha)^n K_1(\alpha)^{-2n_o} = \text{sign}(K_1(\alpha))^n |K_1(\alpha)|^n K_1(\alpha)^{-2n_o} = \varepsilon |K_1(\alpha)|^f,$$

with the sign

$$\varepsilon = (\text{sign}(K_1(\alpha)))^n$$

independent of w. Here we have used the fact that $K_1(\alpha)$ is real. Thus we immediately have that

$$|K_n^{(w)}(\alpha I)| = \varepsilon K_n^{(w)}(\alpha I).$$

Therefore

$$|K_n(a)| \leq \sum_{w^2=I} |K_n^{(w)}(a)| \leq \sum_{w^2=I} |K_n^{(w)}(\alpha I)| = \varepsilon \sum_{w^2=I} K_n^{(w)}(\alpha I) = \varepsilon K_n(\alpha I) = |K_n(\alpha I)|.$$

It remains to prove that $c_n = |K_n(\alpha I)|/q^{(3n^2-\delta(n))/4} \leq 4$. By the recursion formula (4) we have

$$c_{n+1} \leq c_n |K_1(\alpha)| q^{-n/2} q^{(\delta(n+1)-\delta(n)-3)/4} + c_{n-1}.$$

Recall that

$$\delta(n) = \begin{cases} 0 & \text{if } n \text{ is even,} \\ 1 & \text{if } n \text{ is odd.} \end{cases}$$

Thus

$$\delta(n+1) - \delta(n) = \begin{cases} 1 & \text{if } n \text{ is even,} \\ -1 & \text{if } n \text{ is odd.} \end{cases}$$

If $n = 2k$ this gives $c_{2k+1} - c_{2k-1} \leq 2c_{2k}q^{-k}$, and so

$$c_{2k+1} \leq c_1 + 2 \sum_{j=1}^{k} c_{2j} q^{-j}.$$

Similarly

$$c_{2k} \leq 2 \sum_{j=1}^{k-1} c_{2j+1} q^{-j}$$

from which the claim follows easily for $q \geq 3$. The case $q = 2$ is easily checked by hand. $\qquad\square$

The sum $\sum_{v,d,u} \psi(\bar{a}^v w d u)$ gives rise to some intriguing questions on its own; see the problems mentioned Section 5C.

3E. ***Review of cohomology.*** With the results of the previous section, for a fixed $a \in M_n$ the bounds can be proven over those extensions of $\mathbb{F}_q$ in which a can be conjugated to Jordan normal form. However, to get the general result, we need to understand certain cohomology groups attached to the sum — which are independent of the field extension. In the rest of this section we will consider a subset of the matrix group $X \subset M_n$ defined by algebraic equations of the matrix entries as the corresponding algebraic variety.

We first introduce the notation and the main tools, then prove cohomological versions of Propositions 2.3 and 3.14. This enables us to prove Theorems 1.7 and 1.8.

Let $\ell \neq p$ be a prime, and $\overline{\mathbb{Q}}_\ell$ be an algebraic closure of the field $\mathbb{Q}_\ell$ of ℓ-adic numbers, such that there is a p-th primitive root of unity ζ contained in $\overline{\mathbb{Q}}_\ell$. Fix the field embedding $\iota_0 : \mathbb{Q}(\zeta) \to \mathbb{C}$ which sends ζ to $e(1/p)$ and let $\mathcal{L}_\varphi$ be the Artin–Schreier sheaf on $\mathbb{A}^1 = \mathbb{A}^1_{\mathbb{F}_q}$ corresponding to the additive character φ.

For a quasiprojective scheme $X/\mathbb{F}_q$ and a morphism $f : X \to \mathbb{A}^1$ the Grothendieck trace formula [1965] yields

$$\sum_{x \in X(\mathbb{F}_{q^m})} \varphi(f(x)) = \sum_{i=0}^{2 \dim X} (-1)^i \operatorname{Tr}(\operatorname{Frob}_q^m, H_c^i(\overline{X}, f^*\mathcal{L}_\varphi)),$$

where $\overline{X} = X \otimes_{\mathbb{F}_q} \overline{\mathbb{F}}$, H_c^i is the ℓ-adic cohomology group with compact support in degree i. We use the notation $H_c^\bullet$ for the "complex" of cohomologies. These cohomology groups are finite-dimensional $\overline{\mathbb{Q}}_\ell$-vector spaces and $\operatorname{Frob}_q \in \operatorname{Gal}(\overline{\mathbb{F}}/\mathbb{F}_q)$ is the geometric Frobenius acting on them. By Deligne's work [1980] (see also [Kiehl and Weissauer 2001; Milne 1980; 2016]) we know that each Frobenius eigenvalue λ_k^i on H_c^i (for $1 \leq k \leq d_i = \dim H_c^i$) is a Weil number of weight j for some $j \leq i$, that is,

$$|\iota(\lambda_k^i)| = q^{j/2} \tag{41}$$

for all embeddings $\iota : \mathbb{Q}(\lambda) \to \mathbb{C}$, and thus

$$\sum_{x \in X(\mathbb{F}_{q^m})} \varphi(f(x)) = \sum_{i=0}^{2 \dim X} (-1)^i \sum_{j=1}^{d_i} (\lambda_j^i)^m.$$

To simplify the notation we write $H_c^i(Y, f) = H_c^i(\overline{Y}, (f^*\mathcal{L}_\varphi)|_{\overline{Y}})$ for arbitrary subschemes $Y \leq X$. As a corollary of the above we have that if

$$H_c^i(Y, f) = 0 \quad \text{for } i > d$$

then

$$\left| \sum_{x \in X(\mathbb{F}_{q^m})} \varphi(f(x)) \right| \leq C q^{md/2}, \tag{42}$$

with $C = \sum_{i=0}^{d} \dim H_c^i(Y, f)$. We will consider the cohomologies of the sums in the previous section: the sum $K_n(a)$ corresponds to the scheme $X = G = \mathrm{GL}_n$ and the morphism $f = g : x \mapsto \mathrm{tr}(ax + x^{-1})$ (and also the embedding of $\iota_0 : \mathbb{Q}(\zeta) \to \mathbb{C}$).

In the previous sections we derived bounds for the general Kloosterman sums over a finite extension where the eigenvalues of the coefficient matrix are defined. However bounds over an extension field do not imply that the weights are small. Consider, for example, $X = (\mathbb{A}^1 \setminus \{1\}) \sqcup \mathbb{A}^0$ and the regular function $f : X \to \mathbb{A}^1$ defined by

$$f(x) = \begin{cases} x & \text{if } x \in \mathbb{A}^1 \setminus \{1\}, \\ 0 & \text{if } x \in \mathbb{A}^0. \end{cases}$$

Then $\sum_{x \in X(\mathbb{F}_{q^m})} \varphi(f(x)) = 1 - \zeta^m$ which vanishes if $p \mid m$ but only in that case.

The reason for this phenomenon is that the Frobenius eigenvalues on different cohomologies differ by a multiple of a root of unity, and thus cancel in some extensions: here $\dim H_c^1(X, f) = \dim H_c^0(X, f) = 1$, $\dim H_c^2(X, f) = 0$ and the Frobenius eigenvalues are $\lambda_1^1 = \zeta$ and $\lambda_1^0 = 1$.

Thus, to get the general bound, we will prove that the cohomologies $H_c^i(G, g)$ vanish if i is large enough; hence the weights are not too large.

We will use the following properties of $H_c^\bullet$ (for an overview, see, e.g., [Katz 1980] especially Chapters 3.5. and 4.1-3 and [Laumon 2000; Fresán and Jossen 2020]):

Excision. If $f : X \to \mathbb{A}^1$ is a regular function, $Z \to X$ is a closed immersion and $U \to X$ is the complementary open immersion, then there exists a long exact sequence in the form

$$\cdots \to H_c^i(U, f) \to H_c^i(X, f) \to H_c^i(Z, f) \to H_c^{i+1}(U, f) \to \cdots.$$

Künneth formula. If $f_i : X_i \to \mathbb{A}^1$ for $i = 1, 2$, π_i is the canonical map $X = X_1 \times_{\mathrm{Spec}(\mathbb{F}_q)} X_2 \to X_i$, and $f_1 + f_2 := \pi_1^* f_1 + \pi_2^* f_2$, then

$$(f_1^* \mathcal{L}_\varphi) \boxtimes (f_2^* \mathcal{L}_\varphi) \simeq (f_1 + f_2)^* \mathcal{L}_\varphi$$

and

$$H_c^\bullet(X, f_1 + f_2) \simeq H_c^\bullet(X_1, f_1) \otimes H_c^\bullet(X_2, f_2),$$

that is, for all i

$$H_c^i(X, f_1 + f_2) \simeq \bigoplus_{j+k=i} H_c^j(X_1, f_1) \otimes H_c^k(X_2, f_2).$$

We will also need some knowledge of the cohomologies in simple situations. They are listed in the following theorem.

Theorem 3.15 (cohomology of some basic sheaves). (i) *Cohomology of some basic sheaves on $\mathbb{A}^1$.*

 (a) $H_c^i(\mathbb{A}^1, \mathrm{id}) = 0$ *for all i.*

 (b) *If $0 : \mathbb{A}^1 \to \mathbb{A}^1$ is the zero map, then $\mathcal{L}_0 = f_0^* \mathcal{L}_\varphi$ is the constant sheaf and*

$$\dim H_c^i(\mathbb{A}^1, 0) = \begin{cases} 1 & \text{if } i = 2, \\ 0 & \text{if } i \neq 2. \end{cases}$$

 The Frobenius eigenvalue on H_c^2 is q (which is of weight 2).

(ii) *Cohomology of some basic sheaves on $\mathbb{A}^1 \setminus \mathbb{A}^0$.*

 (a) *We have*

$$\dim H_c^i(\mathbb{A}^1 \setminus \mathbb{A}^0, \mathrm{id}) = \begin{cases} 1 & \text{if } i = 1, \\ 0 & \text{if } i \neq 1. \end{cases}$$

 The Frobenius eigenvalue on H_c^1 is 1 (which is of weight 0).

 (b) *We have*

$$\dim H_c^i(\mathbb{A}^1 \setminus \mathbb{A}^0, 0) = \begin{cases} 1 & \text{if } i = 1, 2, \\ 0 & \text{if } i \neq 1, 2. \end{cases}$$

 The Frobenius eigenvalue on H_c^2 is q (which is of weight 2) and on H_c^1 is 1 (weight 0).

(iii) *Cohomology of sheaves corresponding to Kloosterman sums; see [Weil 1948a]. If $\alpha \in \mathbb{F}_q^*$ and $f_\alpha : \mathbb{G}_m = \mathbb{A}^1 \setminus \mathbb{A}^0 \to \mathbb{A}^1$ is the morphism which corresponds to the map*

$$f_\alpha(t) = \alpha \cdot t + 1/t$$

then

$$\dim H_c^i(\mathbb{A}^1 \setminus \mathbb{A}^0, f_\alpha) = \begin{cases} 2 & \text{if } i = 1, \\ 0 & \text{if } i \neq 1, \end{cases}$$

and on H_c^1 both weights are 1.

   The next observation is essential in what follows.

Lemma 3.16. *Let $f, g : X \to \mathbb{A}^1$ be regular functions, $X_0 = f^{-1}(\{0\})$ and consider $f \cdot \mathrm{id}_{\mathbb{A}^1} + g : X \times_{\mathrm{Spec}(\mathbb{F}_q)} \mathbb{A}^1 \to \mathbb{A}^1$. Then*

$$H_c^\bullet(X \times_{\mathrm{Spec}(\mathbb{F}_q)} \mathbb{A}^1, f \cdot \mathrm{id}_{\mathbb{A}^1} + g) \simeq H_c^\bullet(X_0, g) \otimes H_c^\bullet(\mathbb{A}^1, 0);$$

thus

$$H_c^{i+2}(X \times_{\mathrm{Spec}(\mathbb{F}_q)} \mathbb{A}^1, f \cdot \mathrm{id}_{\mathbb{A}^1} + g) \simeq H_c^i(X_0, g) \otimes H_c^2(\mathbb{A}^1, 0) \quad \text{for all } i. \tag{43}$$

Proof. Let $V = X \setminus X_0$ and consider the morphism $j = \mathrm{id}_V \otimes (\mathrm{id}_{\mathbb{A}^1} - g)/f : V \times \mathbb{A}^1 \to V \times \mathbb{A}^1$. This is clearly an isomorphism and $j \circ (f \cdot \mathrm{id}_{\mathbb{A}^1} + g) = 0_V + \mathrm{id}_{\mathbb{A}^1}$; thus by the Künneth formula, $H_c^\bullet(V \times \overline{\mathbb{A}^1}, j \circ (f \cdot \mathrm{id}_{\mathbb{A}^1} + g)) \equiv 0$.

   Let $Z = X_0 \times \mathbb{A}^1$ and $U = X \times \mathbb{A}^1 \setminus Z$. From the previous argument, $H_c^\bullet(U, f \cdot \mathrm{id}_{\mathbb{A}^1} + g) \equiv 0$. By excision $H_c^i(X \times_{\mathrm{Spec}(\mathbb{F}_q)} \mathbb{A}^1, f \cdot \mathrm{id}_{\mathbb{A}^1} + g) \simeq H_c^i(Z, f \cdot \mathrm{id}_{\mathbb{A}^1} + g)$, but $(f \cdot \mathrm{id}_{\mathbb{A}^1})|_Z \equiv 0$; hence $(f \cdot \mathrm{id}_{\mathbb{A}^1} + g)|_Z = g|_Z = g|_{X_0} + 0|_{\mathbb{A}^1}$ and applying the Künneth formula we get the lemma. $\qquad\square$

Remark 10. This lemma is the cohomological form of the straightforward computation

$$\sum_{(x,t)\in X(\mathbb{F})\times F} \varphi(tf(x)+g(x)) = \sum_{x\in X(\mathbb{F})} \varphi(g(x))\sum_{t\in\mathbb{F}}\varphi(tf(x)) = q\sum_{x\in X_0(\mathbb{F})}\varphi(g(x)).$$

A similar argument as in the proof appears in motivic context in [Fresán and Jossen 2020, Lemma 6.5.3 and Remark 6.5.4].

Applying the lemma repeatedly we get:

Corollary 3.17. *Let* $\pi_j : \mathbb{A}^m \to \mathbb{A}^1$ *be the projection*

$$(x_1, x_2, \ldots, x_m) \mapsto x_j.$$

For $f_j, g : X \to \mathbb{A}^1$, $1 \le j \le m$, *let* $h : X \times_{\mathrm{Spec}(\mathbb{F}_q)} \mathbb{A}^m \to \mathbb{A}^1$ *be defined by*

$$h = \sum_{j=1}^{m} f_j \cdot \pi_j + g.$$

Consider $X_0 = \{x \in X \mid h(x, \cdot) \equiv 0\} \le X$ *as a subscheme. Then*

$$H_c^\bullet(X \times_{\mathrm{Spec}(\mathbb{F}_q)} \mathbb{A}^m, h) \simeq H_c^\bullet(X_0, g) \otimes \left(\bigotimes_{j=1}^{m} H_c^\bullet(\mathbb{A}^1, 0) \right);$$

thus

$$H_c^{i+2m}(X \times_{\mathrm{Spec}(\mathbb{F}_q)} \mathbb{A}^m, h) \simeq H_c^i(X_0, g) \otimes \left(\bigotimes_{j=1}^{m} H_c^2(\mathbb{A}^1, 0) \right) \quad \text{for all } i.$$

We will now show the vanishing of cohomologies of high enough degree for the exponential sums that were used in the previous sections.

3F. *The proof of Theorem 1.7.* We first start with the reduction to Jordan blocks as in Proposition 2.3. Let $G = \mathrm{GL}_n$, $G_k = \mathrm{GL}_k$, $G_l = \mathrm{GL}_l$ for some $n = k + l$. Let

$$a = \left(\begin{array}{c|c} a_k & b \\ \hline 0 & a_l \end{array} \right) \in \mathbb{A}^{n \times n}$$

be block upper triangular with $a_k \in \mathbb{A}^{k \times k}$, $a_l \in \mathbb{A}^{l \times l}$. Let $g : G \to \mathbb{A}^1$, $x \mapsto \mathrm{tr}(ax + x^{-1})$, and for the diagonal blocks denote by g_k and g_l the morphisms $x_k \mapsto \mathrm{tr}_k(a_k x_k + x_k^{-1})$ and $x_l \mapsto \mathrm{tr}_l(a_l x_l + x_l^{-1})$, respectively. Let $H^\bullet = H_c^\bullet(G, g)$ and similarly $H_k^\bullet = H_c^\bullet(G_k, g_k)$, $H_l^\bullet = H_c^\bullet(G_l, g_l)$.

Proposition 3.18. *If* a_k *and* a_l *have no common eigenvalues, then*

$$H^\bullet \simeq H^\bullet(\mathbb{A}^{k \times l}, 0) \otimes H_k^\bullet \otimes H_l^\bullet, \quad \text{that is,} \quad H^i \simeq H^{2kl}(\mathbb{A}^{k \times l}, 0) \otimes \left(\bigoplus_{j+j'=i-2kl} H_k^j \otimes H_l^{j'} \right),$$

where $H^\bullet(\mathbb{A}^{k \times l}, 0) = \left(\bigotimes_{i=1}^{kl} H_c^\bullet(\mathbb{A}^1, 0) \right)$

Proof. The morphism $U_{[k,l]} \to \mathbb{A}^{kl}$, $u \mapsto (u_{i,j})_{1 \leq i \leq k, 1 \leq j \leq l}$, is an isomorphism, so we apply Corollary 3.17 with $m = kl$ and $X = G$ on $X \times_{\mathrm{Spec}(\mathbb{F}_q)} U_{[k,l]}$ and

$$h : (x, u) \mapsto \mathrm{tr}(a(u^{-1}xu) + (u^{-1}xu)^{-1}).$$

From the proof of Proposition 2.3 it is clear that $h(x, \cdot)$ is cohomologically nontrivial if and only if $x' = 0$, that is, $x \in X_0$ with $X_0 \leq X$ the subscheme of "block upper triangular" matrices and by Corollary 3.17

$$H_c^\bullet(X \otimes_{\mathrm{Spec}(\mathbb{F}_q)} U_{[k,l]}, h) \simeq H_c^\bullet(X_0, g) \otimes H^\bullet(\mathbb{A}^{k \times l}, 0) \simeq H_c^\bullet(G_k, g_k) \otimes H_c^\bullet(G_l, g_l) \otimes H^\bullet(\mathbb{A}^{k \times l}, 0),$$

where the second isomorphism is a consequence of a being a block matrix. Thus

$$X_0 \simeq G_k \times_{\mathrm{Spec}(\mathbb{F}_q)} G_l \times_{\mathrm{Spec}(\mathbb{F}_q)} \mathbb{A}^{k \times l}, \quad g|_{X_0} = g_k + g_l + 0_{\mathbb{A}^{k \times l}}$$

and the Künneth formula.

On the other hand $j : X \to X$, $(u, x) \mapsto (u, uxu^{-1})$, is an isomorphism and $j^*h = 0_{U_k} + g$ so

$$H_c^\bullet(X \otimes_{\mathrm{Spec}(\mathbb{F}_q)} U_{[k,l]}, h) \simeq H_c^\bullet(G, g) \otimes \left(\bigotimes_{i=1}^{kl} H_c^\bullet(\mathbb{A}^1, 0) \right);$$

hence the proposition. $\square$

Proof of Theorem 1.7. Since cohomology does not depend on the finite field in question we may assume that a is a diagonal matrix with nonzero and unequal entries α_i on the diagonal. As above let $g : G \to \mathbb{A}^1$, $x \mapsto \mathrm{tr}(ax + x^{-1})$. Also let $H_{K(\alpha_i)}^\bullet = H_c^\bullet(\mathbb{A}^1 \setminus \mathbb{A}^0, f_\alpha)$. Repeated applications of the proposition above gives

$$H_c^\bullet(G, g) = H^{n(n-1)}(\mathbb{A}^{n(n-1)/2}, 0) \otimes \bigotimes_{i=1}^n H_{K(\alpha_i)}^\bullet.$$

By Theorem 3.15(i) and the Künneth formula, $\dim H^{n(n-1)}(\mathbb{A}^{n(n-1)/2}, 0) = 1$, with Frobenius eigenvalue $q^{n(n-1)}$. Similarly by Theorem 3.15(iii) $\bigotimes_{i=1}^n H_{K(\alpha_i)}^\bullet$ is concentrated in degree n where it equals $\bigotimes_{i=1}^n H_{K(\alpha_i)}^1$.

The claim now follows from the purity of the classical Kloosterman sums $K_1(\alpha_i)$. $\square$

3G. *Bounding the weights in the nonsplit case.*

For the proof of Theorem 1.8 we need to bound the degrees of the nontrivial cohomology groups. Recall that for $w^2 = I$ we defined in Theorem 1.5

$$N = N(w) = \left| \{(i, j) \mid 1 \leq i < j \leq n, \ w(j) < w(i) \leq j\} \right|.$$

Theorem 3.19. *Assume that $a = \alpha I + \bar{a}$, with $\alpha \in \mathbb{F}_q^*$ and $\bar{a}$ nilpotent, and consider the cohomology $H_c^\bullet(C_w, x \mapsto \mathrm{tr}(ax + x^{-1}))$ associated to the exponential sum* (31).

(i) *If $w^2 \neq I$, then $H_c^\bullet(C_w, x \mapsto \mathrm{tr}(ax + x^{-1})) \equiv 0$.*

(ii) *If $w^2 = I$, then $H_c^i(C_w, x \mapsto \mathrm{tr}(ax + x^{-1})) = 0$ for $i > n^2 + 2N(w)$; thus all weights of the sum $\sum_{x \in C_w} \psi(ax + x^{-1})$ are at most $n^2 + 2N(w)$.*

Proof. (i) We may assume that $\bar{a}$ is upper triangular. The case $w^2 \neq I$ follows from the proof of Proposition 3.1. Let the pair (i, j) be chosen as in there: i is minimal such that $i \neq w^2(i)$ and $j = w(i)$. Consider the subvarieties $Y = \{I + s e_{i,j} \mid s \in \mathbb{F}_q\} \simeq \mathbb{A}^1$, $X_1 = U_b$ and $X_2 = \{x \in B \mid x_{ij} = 0\}$. Then we have the decomposition $C_w = X_1 \times Y \times X_2$ by mapping $x_1 \in X_1$, $x_2 \in X_2$, $s \in \mathbb{F}_q$ to $g = x_1 w(I + s e_{i,j}) x_2$ (this is indeed an isomorphism of algebraic varieties). If $X = X_1 \times X_2$ then the proof of Proposition 3.1 shows that the map $g \operatorname{tr} a + (uwb)^{-1}$ is of the form $f(x)s + g(x)$ with $X_0 = \{x \in X \mid f(x) = 0\}$ empty. Hence Lemma 3.16 implies the vanishing of cohomology.

Assume now that $w^2 = I$. By Lemma 3.7, $C_w \simeq U^\flat \times T \times U_b^\sharp \times \bar{U}_a \times U_o \times U_b^\flat$ and this is an isomorphism of algebraic varieties.

We can apply Corollary 3.17 in the setting of Proposition 3.8. We have $\mathbb{A}^m = \bar{U}_a$ with $m = n_a$, $h : x \mapsto \operatorname{tr}(ax + x^{-1})$ and

$$X_0 = U_b \times T \times U_b^\sharp \times U_o \hookrightarrow X = U^\flat \times T \times U_b^\sharp \times U_o \times U_b^\flat,$$

where the embedding maps the last coordinate to I. We obtain

$$H_c^\bullet(C_w, x \mapsto \operatorname{tr}(ax + x^{-1})) \simeq H_c^\bullet(U^\flat \times T \times U_b^\sharp \times U_o, g_1) \otimes \left(\bigotimes_{i=1}^{n_a} H_c^\bullet(\mathbb{A}^1, 0) \right),$$

where $g_1 = \operatorname{tr}(a^v w t u_1 u_o + (w t u_1 u_o)^{-1})$ with the notation of Section 3C.

Applying Corollary 3.17 in the setting of Proposition 3.9 with $\mathbb{A}^m \simeq U_o$, $m = n_o$, $h : x \mapsto \operatorname{tr}(ax + x^{-1})$ and

$$X_0 = U^\flat \times T(w) \times U_b^\sharp \hookrightarrow X = U^\flat \times T \times U_b^\sharp,$$

where $T(w)$ is as in the proof of Proposition 3.14, we obtain

$$H_c^\bullet(C_w, x \mapsto \operatorname{tr}(ax + x^{-1})) \simeq H_c^\bullet(U^\flat \times T(w) \times U_b^\sharp, g_2) \otimes \left(\bigotimes_{i=1}^{n_a + n_o} H_c^\bullet(\mathbb{A}^1, 0) \right),$$

where $g_2 = \sum_{i=w(i)}(\alpha t_i + t_i^{-1}) + \operatorname{tr}(\bar{a}^v w d u_1)$.

The Künneth formula yields

$$H_c^\bullet(C_w, x \mapsto \operatorname{tr}(ax + x^{-1}))$$

$$\simeq H_c^\bullet(U^\flat \times U_b^\sharp, g_3) \otimes \left(\bigotimes_{i=1}^{n_a + n_o} H_c^\bullet(\mathbb{A}^1, 0) \right) \otimes \left(\bigotimes_{j=1}^{e} H_c^\bullet(\mathbb{A}^1 \setminus \mathbb{A}^0, 0) \right) \otimes \left(\bigotimes_{k=1}^{f} H_c^\bullet(\mathbb{A}^1 \setminus \mathbb{A}^0, f_\alpha) \right),$$

with $g_3 = \operatorname{tr}(\bar{a}^v w d u_1)$ and $f_\alpha : x \mapsto \alpha x + x^{-1}$.

Now it is enough to observe that the maximal nontrivial cohomology group is of degree

$$2 \dim U^\flat \times U_b^\sharp + 2(n_a + n_o) + 2e + f,$$

where e is the number of involution pairs, and f is the number fixed elements in w. It is clear that $n = 2e + f$ and the calculation in the proof of (40) shows that the rest is equal to $n(n-1) + 2N(w)$. $\square$

To show the vanishing of cohomologies of the Bruhat cells in higher degrees, we need one more combinatorial lemma. For this the Weyl group W is identified with the symmetric group S_n viewed as the group of permutations of the set $\{1, \ldots, n\}$ as in (32). In the notation of Remark 9 we let for any $w \in W$

$$\mathcal{J}_{a/o}^{\flat}(w) = \{(i, j) \mid 1 \le i < j \le n,\ w(j) < w(i) \le j\}$$

Recall that $N(w) = |\mathcal{J}_{a/o}^{\flat}(w)| = \dim \overline{U}_{a/o}^{\flat}$.

Lemma 3.20. *Let e be a positive integer such that $e \le \lfloor n/2 \rfloor$ and $w_e \in S_n$ be the involution for which*

$$w_e(j) = \begin{cases} n - j + 1 & \text{if } j = 1, \ldots, e, \\ j & \text{if } j = e+1, \ldots, n-e. \end{cases}$$

We then have the following:

(i) *If $w^2 = I$ and w is a product of e disjoint transpositions then $N(w) \le N(w_e)$ with equality only if $w = w_e$.*

(ii) *$N(w_e) = e(n - e)$. In particular $N(w)$ is maximal for the long element $w_{\lfloor n/2 \rfloor}$, and we have $N(w_{\lfloor n/2 \rfloor}) = (n^2 - \delta(n))/4$.*

Proof. We proceed by induction on e. Let $k = w(n)$ and first assume that $k > 1$. Let $v = (12 \ldots k) \in S_n$, that is,

$$v(j) = \begin{cases} j+1 & \text{if } j < k, \\ 1 & \text{if } j = k, \\ j & \text{if } j > k, \end{cases}$$

and $w' = v^{-1} w v \in S_n$. We claim that $(i, j) \in \mathcal{J}_{a/o}^{\flat}(w) \Rightarrow (v(i), v(j)) \in \mathcal{J}_{a/o}^{\flat}(w')$; thus $N(w) \le N(w')$.

To see this, first assume that $\{i, j\} \cap \{k, n\} = \varnothing$. In this case the claim is clear since v respects the ordering of $X = \{1, 2, \ldots, n\} \setminus \{k, n\}$ and $w(X) \subseteq X$.

Now the case $(i, k) \in \mathcal{J}_{a/o}^{\flat}(w)$ does not arise since $w(k) = n$. There is a single j such $(k, j) \in \mathcal{J}_{a/o}^{\flat}(w)$, namely $j = n$, but then $(v(k), v(n)) = (1, n) \in \mathcal{J}_{a/o}^{\flat}(w')$. Finally $(i, n) \in \mathcal{J}_{a/o}^{\flat}(w')$ for all i; thus for all $(i, n) \in \mathcal{J}_{a/o}^{\flat}(w)$ we have $(v(i), v(n)) \in \mathcal{J}_{a/o}^{\flat}(w')$.

If $w(n) \ne 1$, then the last case of the above argument shows $N(w)$ is strictly smaller than $N(w')$.

We now move to the induction step. Let w' be as above if $w(n) \ne 1$ and $w' = w$ otherwise. Let also $w'' \in S_{n-2}$ be the element which arises from the permutation w' restricted to $\{2, \ldots, n-1\}$ which we identify with $\{1, \ldots, n-2\}$ using $j \mapsto j - 1$. Then w'' is a product of $e - 1$ transpositions and the induction hypothesis shows that $N(w'')$ is maximal if and only if w'' arises from $w' = w_e$.

To prove the second part note that if $w(n) = 1$, then $(i, n) \in \mathcal{J}_{a/o}^{\flat}(w)$ for all $i < n$. Again let now $w'' \in S_{n-2}$ be the element which we obtain by deleting the first and last rows and columns of w'. Then $N(w'') = N(w) + n - 1$. This time induction again shows that $N(w_e) = (n-1) + (n-3) + \cdots + (n-2e+1)$ from which the statements follow. $\qquad\square$

The following lemma enables us to work on the individual groups GL_{n_j}:

Lemma 3.21. *Let $G = \mathrm{GL}_m$ for some m and $g : \mathrm{GL}_m(\mathbb{F}_q) \to \mathbb{A}^1$, $x \mapsto \mathrm{tr}(ax + x^{-1})$, where $a = \alpha I + \bar{a}$, with $\bar{a}$ nilpotent. Then $H^i(\mathrm{GL}_m, g)$ vanishes if $i > m^2 + (m^2 - \delta(m))/2$.*

Proof. We may assume that $\bar{a}$ is upper triangular. For an involution $w^2 = I$ again let $e = e(w)$ be the number of involution pairs in w. By Theorem 3.19 and Lemma 3.20

$$H_c^i(C_w, g) = \begin{cases} 0 & \text{for any } i, \text{ if } w^2 \neq I, \\ 0 & \text{for } i > m^2 + 2e(m - e), \text{ if } w \neq I, w^2 = I, \\ 0 & \text{for } i \neq m^2, \text{ if } w = I. \end{cases} \qquad (44)$$

Let l be the standard length function on $W = S_m$ [Borel 1991, 21.21] and consider

$$Y_l = \bigsqcup_{l(w)=l} C_w.$$

We also let $Y_0 = B$ corresponding to the unit element of W.

If $l(w) = l$ then C_w is an open subscheme of

$$X_l = \bigsqcup_{l(w) \leq l} C_w = Y_l \sqcup X_{l-1}.$$

Clearly for any w we have $H_c^i(C_w, g) = 0$ if $i > m^2 + (m^2 - \delta(m))/2$ and so this remain true for Y_l:

$$H_c^i(Y_l, g) = 0 \quad \text{if } i > m^2 + (m^2 - \delta(m))/2.$$

We will now apply induction in the excision long exact sequence on the disjoint union $X_l = Y_l \sqcup X_{l-1}$:

$$\cdots \to H_c^i(X_{l-1}, f) \to H_c^i(X_l, f) \to H_c^i(Y_l, f) \to H_c^{i+1}(X_{l-1}, f) \to \cdots.$$

For $l = 0$ we have that $X_0 = Y_0$ and so

$$H_c^i(X_0, g) = 0$$

already for $i > m^2$. Assume now that $H_c^i(X_{l-1}, g) = 0$ if $i > m^2 + (m^2 - \delta(m))/2$. From the excision long exact sequence we then also have that $H_c^i(X_l, g) = 0$ for $i > m^2 + (m^2 - \delta(m))/2$. $\qquad\square$

Proof of Theorem 1.8. Fix $a \in M_n$. As the weights do not change after base change, we might work over a sufficiently large finite extension of $\mathbb{F}_q$, say $\mathbb{F}_{q^{m(a)}}$, over which a is conjugate to a block diagonal matrix, where the blocks on the diagonal are square matrices $a_j \in M_{n_j}$ in Jordan normal form with a unique eigenvalue α_j.

Then by Proposition 3.18 we have

$$H_c^\bullet(G, g) \simeq \left(\bigotimes_{j=1}^r H_c^\bullet(\mathrm{GL}_{n_j}, g_j) \right) \otimes \left(\bigotimes_{1 \leq i < j \leq r} \bigotimes_{k=1}^{n_i n_j} H_c^\bullet(\mathbb{A}^1, 0) \right),$$

where $g_j : \mathrm{GL}_{n_j}(\mathbb{F}_q) \to \mathbb{A}^1$, $x \mapsto \mathrm{tr}(a_j x + x^{-1})$.

We apply the lemma with $m = n_i$ and $g = g_i$ for $1 \leq i \leq r$, respectively. By (42) and Lemma 3.21 we have that $K_n(a) \ll q^d$, with

$$d = \sum_{i=1}^{r} (3n_i^2 - \delta(n_i))/4 + \sum_{1 \leq i < j \leq r} n_i n_j.$$

To conclude the proof note that

$$\max\left(\sum_{i=1}^{r} (3n_i^2 - \delta(n_i))/4 + \sum_{1 \leq i < j \leq r} n_i n_j \,\Big|\, r, n_i \in \mathbb{N} : \sum_{i=1}^{r} n_i = n \right) = (3n^2 - \delta(n))/4. \qquad \square$$

4. Degenerate cases

4A. *Preliminary observations on $K_n(a, b)$.* The results in this section are combinatorial in nature and will not require cohomology. Therefore from now on we work solely over $\mathbb{F}_q$ and $M_n = M_n(\mathbb{F}_q)$.

Let a and b singular $n \times n$ matrices such that

$$r = \mathrm{rk}(b) \geq s = \mathrm{rk}(a). \tag{45}$$

This section contains some elementary observations about the generalized Kloosterman sums

$$K_n(a, b) = \sum_{x \in \mathrm{GL}_n(\mathbb{F}_q)} \psi(ax + bx^{-1}).$$

First, we clearly have

$$K_n(a, b) = K_n(c_1 a c_2^{-1}, c_2 b c_1^{-1}) \quad \text{for any } c_1, c_2 \in \mathrm{GL}_n(\mathbb{F}_q). \tag{46}$$

In this, and the following sections, when we write a block matrix $a = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \in \mathbb{F}_q^{n \times n}$, we always mean the blocks to correspond to the partition $\{1, 2, \ldots, n\} = \{1, \ldots, r\} \sqcup \{r+1, \ldots, n\}$, r as in (45). For example, by (46) we may assume that $b = E_r$, where

$$E_r = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \tag{47}$$

is a standard idempotent, but an exact description of the equivalence classes is a delicate question. However all we need is a reasonable set of representatives for the action $(a, b) \mapsto (c_1 a c_2^{-1}, c_2 b c_1^{-1})$ that are suitable for handling the Kloosterman sums. This is most conveniently achieved via a parabolic Bruhat decomposition of $G = \mathrm{GL}_n(\mathbb{F}_q)$ with respect to the subgroups $P_r = P_r(\mathbb{F}_q)$ consisting of elements that when acting on row vectors map the subspace $V_r = \langle e_{r+1}, \ldots, e_n \rangle$ to itself. In block matrix notation we have

$$P_r = P_r(\mathbb{F}_q) = \left\{ g \in \mathrm{GL}_n(\mathbb{F}_q) \,\Big|\, g = \begin{pmatrix} g_{11} & g_{12} \\ 0 & g_{22} \end{pmatrix} \right\}.$$

Let $Q_r = P_r^T = \left\{ g \in \mathrm{GL}_n(\mathbb{F}_q) \mid g = \begin{pmatrix} g_{11} & 0 \\ g_{21} & g_{22} \end{pmatrix} \right\}$ be the stabilizer of the columns space $\langle e_1^T, \ldots, e_r^T \rangle$, the subspace of linear functionals vanishing on V_r. Then we have the following Bruhat decomposition for the group $G = \mathrm{GL}_n(\mathbb{F}_q)$.

Proposition 4.1. *Let $G = \mathrm{GL}_n(\mathbb{F}_q)$ with P_r, Q_r as above. Then*

$$G = \bigcup_{w \in W_P \backslash W / W_P} Q_r w P_r,$$

where $W_P = W \cap P_r = W \cap Q_r$ and $W_P \backslash W / W_P$ denotes the set of double cosets.

Proof. From $G = \bigcup_{w \in W} B w B$ it is clear that $G = \bigcup_{w \in W} P_{n-r} w P_r$. Let $w_l = \sum_{i=1}^{n} e_{i,n-i}$ be the matrix that corresponds to the longest element $(1, n)(2, n-1) \cdots \in W = S_n$. Then $w_l P_{n-r} w_l = Q_r$, from which $G = w_l G = \bigcup_{w \in W} Q_r w P_r$. It is obvious that if $w' = w_1 w w_2$ with $w_1, w_2 \in W_P$, then $Q_r w' P_r = Q_r w P_r$. $\quad\square$

The following lemma is well known; see, for example, Section 1.3 in [James and Kerber 1981].

Lemma 4.2. *Let $W = S_{1,2,\dots,n}$, $W_P = S_{1,2,\dots,r} \times S_{r+1,r+2,\dots,n}$ and $m = \min(r, n-r)$. A set of double coset representatives of $W_P \backslash W / W_P$ is*

$$\overline{W}_r = \{w_k \mid k = 1, \dots, m\},$$

where for $k \leq m$, w_k is the permutation matrix

$$w_k = \sum_{i=1}^{k} e_{i,i+r} + \sum_{i=k+1}^{r} e_{i,i} + \sum_{i=r+k+1}^{n} e_{i,i}, \tag{48}$$

which we also identify with $w_k = (1, r+1)(2, r+2) \cdots (k, r+k) \in W = S_n$.

Recall that we have $E_r = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$.

Proposition 4.3. *Let a and b as in (45). Then there exist matrices d and $w \in \overline{W}_r$ such that*

$$K_n(a, b) = K_n(E_r \cdot d, E_r \cdot w).$$

Proof. First we can write $a = c_1 E_s d_1$ and $b = d_2 E_r c_2$ for some $c_i, d_j \in \mathrm{GL}_n$, and thus $K_n(a, b) = K_n(E_r d_0, E_r c_0)$, where $d_0 = E_s d_1 d_2$, $c_0 = c_2 c_1$. Here we have used that $E_r E_s = E_s$, since $r \geq s$.

Now we have that $c_0 = q w p$, where $q \in Q_r$, $p \in P_r$ are as in Proposition 4.1, and $w \in \overline{W}_r$ as in Lemma 4.2. Let

$$U_r = \left\{ g \in \mathrm{GL}_n(\mathbb{F}_q) \,\middle|\, g = \begin{pmatrix} I_r & g_{12} \\ 0 & I_{n-r} \end{pmatrix} \right\} \tag{49}$$

be the unipotent radical of P_r, and

$$L_r = P_r \cap Q_r = \left\{ g \in \mathrm{GL}_n(\mathbb{F}_q) \,\middle|\, g = \begin{pmatrix} g_{11} & 0 \\ 0 & g_{22} \end{pmatrix} \right\},$$

so that $P_r = L_r U_r$ and $Q_r = L_r U_r^T$.

We have $L_r = H_1 H_2$, where

$$H_1 = \left\{ g \in \mathrm{GL}_n(\mathbb{F}_q) \,\middle|\, g = \begin{pmatrix} g_{11} & 0 \\ 0 & I_{n-r} \end{pmatrix} \right\}, \quad H_2 = \left\{ g \in \mathrm{GL}_n(\mathbb{F}_q) \,\middle|\, g = \begin{pmatrix} I_r & 0 \\ 0 & g_{22} \end{pmatrix} \right\}. \tag{50}$$

For $g \in L_r$,

$$E_r g = g E_r,$$

and for $u \in U_r$ we have

$$u E_r = E_r \quad \text{and} \quad E_r u^T = E_r.$$

Therefore writing $q = g_1 u_1^T$, $p = g_2 u_2$ we have

$$K_n(E_r d_0, E_r g_1 u_1^T w g_2 u_2) = K_n(g_2 u_2 E_r d_0, g_1 E_r u_1^T w) = K_n(E_r d, w),$$

where $d = g_2 d_0 g_1$.   □

4B. *The proof of Theorem 1.9 and a preliminary bound.* We will give a proof of Theorem 1.9 on $K_n(a, 0)$ with a of rank r, namely

$$K(a, 0) = K_n(E_r, 0) = (-1)^r q^{-\binom{r+1}{2}} q^{rn} |\mathrm{GL}_{n-r}(\mathbb{F}_q)|. \tag{51}$$

While this evaluation is trivial, for singular a, b it is the basis of our bounds for the general Kloosterman sums given in Proposition 4.4 below.

Proof of Theorem 1.9. We have that

$$K_n(E_r, 0) = \sum_{x \in \mathrm{GL}_n(\mathbb{F}_q)} \psi(E_r x) = \frac{1}{q^{r(n-r)}} \sum_{u \in U_r} \sum_{x \in \mathrm{GL}_n(\mathbb{F}_q)} \psi(E_r u x),$$

where $U_r = \left\{ u \in \mathrm{GL}_n(\mathbb{F}_q) \mid u = \left(\begin{smallmatrix} I_r & u_{12} \\ 0 & I_{n-r} \end{smallmatrix} \right) \right\}$ as in (49). Let $x = \left(\begin{smallmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{smallmatrix} \right)$. It is clear that when summing over u first,

$$\sum_{u \in U_r} \psi(E_r u x) = \begin{cases} 0 & \text{if } x_{21} \neq 0, \\ q^{r(n-r)} & \text{if } x_{21} = 0. \end{cases}$$

Therefore

$$K_n(E_r, 0) = q^{r(n-r)} |\mathrm{GL}_{n-r}(\mathbb{F}_q)| K_r(I_r, 0),$$

which leads immediately to the claim, in view of Theorem 1.2.   □

Proposition 4.4. *Let $w = w_k \in \overline{W}_r$ be as in Lemma 4.2, and $d \in M_n$. Then*

$$|K_n(E_r d, E_r w)| \leq \sum_{j=0}^{k} q^{j(n-r-j)-\binom{j}{2}} \frac{|\mathrm{GL}_{n-r-j}(\mathbb{F}_q)|}{|\mathrm{GL}_{n-r}(\mathbb{F}_q)|} R_w(j),$$

where $R_w(j) = \left| \left\{ x \in \mathrm{GL}_n(\mathbb{F}_q) \mid x E_r w = \left(\begin{smallmatrix} y_{11} & y_{12} \\ 0 & y_{22} \end{smallmatrix} \right), \mathrm{rk}(y_{22}) = j \right\} \right|$.

Proof. First swap the parameters

$$K_n(E_r d, E_r w) = K_n(E_r w, E_r d)$$

and then use the action of U_r as above to get

$$K_n(E_r w, E_r d) = \sum_{x \in \mathrm{GL}_n(\mathbb{F}_q)} \psi(E_r w x + E_r d x^{-1})$$

$$= \frac{1}{q^{r(n-r)}} \sum_{u \in U_r} \sum_{x \in \mathrm{GL}_n(\mathbb{F}_q)} \psi(E_r w u x + E_r d x^{-1} u^{-1})$$

$$= \frac{1}{q^{r(n-r)}} \sum_{x \in \mathrm{GL}_n(\mathbb{F}_q)} \sum_{u \in U_r} \psi(E_r w u x + E_r d x^{-1}).$$

This shows that

$$K_n(E_r w, E_r d) = \sum_{x \in \mathcal{R}_w} \psi(E_r w x + E_r d x^{-1}),$$

where $\mathcal{R}_w = \left\{ x \in \mathrm{GL}_n(\mathbb{F}_q) \mid x E_r w = \begin{pmatrix} y_{11} & y_{12} \\ 0 & y_{22} \end{pmatrix} \text{ for some } y_{11}, y_{12}, y_{22} \right\}$.

Let

$$\mathcal{R}_w(j) = \left\{ x \in \mathrm{GL}_n(\mathbb{F}_q) \,\middle|\, x E_r w = \begin{pmatrix} y_{11} & y_{12} \\ 0 & y_{22} \end{pmatrix}, \ \mathrm{rk}(y_{22}) = j \right\} \tag{52}$$

so that $\mathcal{R}_w = \bigsqcup_j \mathcal{R}_w(j)$. Since in $x E_r w$ the last $n-r-k$ columns are 0, $\mathcal{R}_w(j)$ is empty if $j > k$ and so

$$K_n(E_r w, E_r d) = \sum_{j=1}^{k} \sum_{x \in \mathcal{R}_w(j)} \psi(E_r w x + E_r d x^{-1}).$$

Clearly if $x \in \mathcal{R}_w(j)$ then $gx \in \mathcal{R}_w(j)$ for any $g \in P_r$. Therefore let

$$H_2 = \left\{ g \in L_r \,\middle|\, g = \begin{pmatrix} I_r & 0 \\ 0 & h \end{pmatrix}, \ h \in \mathrm{GL}_{n-r}(\mathbb{F}_q) \right\}$$

as in (50) and note that for $g \in H_2$, $x \in \mathcal{R}_w(j)$

$$\mathrm{tr}(E_r w g x) = \mathrm{tr}^{(r)}(y_{11}) + \mathrm{tr}^{(n-r)}(y_{22} h) \quad \text{and} \quad \mathrm{tr}(E_r d (gx)^{-1}) = \mathrm{tr}(E_r d x^{-1}),$$

where $\mathrm{tr}^{(j)}$ is the $j \times j$ matrix trace and y_{11}, y_{22} are as in (52). This immediately implies that for $x \in \mathcal{R}_w(j)$

$$\sum_{g \in H_2} \psi(E_r w g x + E_r d (gx)^{-1}) = K_{n-r}(E_j, 0) \varphi(\mathrm{tr}^{(r)}(y_{11}) + \mathrm{tr}(E_r d x^{-1})),$$

and this gives

$$\sum_{x \in \mathcal{R}_w(j)} \psi(E_r w x + E_r d x^{-1}) = \frac{1}{|\mathrm{GL}_{n-r}(\mathbb{F}_q)|} \sum_{g \in H_2} \sum_{x \in \mathcal{R}_w(j)} \psi(E_r w x g + E_r d (xg)^{-1})$$

$$= \sum_{x \in \mathcal{R}_w(j)} \varphi(\mathrm{tr}^{(r)}(y_{11}) + \mathrm{tr}(E_r d x^{-1})) \frac{K_{n-r}(E_j, 0)}{|\mathrm{GL}_{n-r}(\mathbb{F}_q)|}.$$

The proposition follows from trivially estimating the last sum using $|\varphi(\cdot)| \leq 1$ and the evaluation (51). $\square$

4C. *The proof of Theorem 1.10.* We restate the theorem and its corollary. We need to prove that if a and b are singular $n \times n$ matrices such that $s = \mathrm{rk}(a) \le r = \mathrm{rk}(b) < n$ and $m = \min(r, n - r)$, then

(i) $K_n(a, b) \le 2q^{n^2 - rn + r^2 + \binom{m}{2}}$,

(ii) if a, b are not both 0 then $K_n(a, b) \le 2q^{n^2 - n + 1}$, and

(iii) this bound is sharp, since

$$K_n(e_{1,n}, e_{1,n}) = q^{2n-2}|\mathrm{GL}_{n-2}(\mathbb{F}_q)| + (q-1)q^{n-1}|\mathrm{GL}_{n-1}(\mathbb{F}_q)| \sim q^{n^2 - n + 1}.$$

Here (ii) is an obvious corollary of (i), and we start with the proof of that claim (Theorem 1.10). By Proposition 4.4 this will require some estimates for the number $R_w(j) = |\mathcal{R}_w(j)|$.

Lemma 4.5. *Let* $w = w_k \in \overline{W}_r$ *and* $\mathcal{R}_w(j)$ *be as in* (52). *Then*

$$\mathcal{R}_w(j) = P_r w_j P_s,$$

where $P_r = L_r U_r$ *as in* (49), *and* $P_s = H_1' H_2 U_r$ *with* H_1, H_2 *as in* (50) *and* $H_1' = H_1 \cap w_k H_1 w_k$.

Proof. First, if $y \in \mathcal{R}_w(j) E_r w$ then $gy \in \mathcal{R}_w(j) E_r w$ for any $g \in P_r$. This immediately shows that $P_r \mathcal{R}_w(j) = \mathcal{R}_w(j)$.

On the other hand if $g = h_1 h_2 u$ with $h_1 \in H_1'$, $h_2 \in H_2$ and $u \in U_r$ then

$$xg E_r w = xh_1 E_r w = x E_r h_1 w = (x E_r w) w h_1 w = y h_1'$$

for some $h_1' \in H_1$ and $y = \left(\begin{smallmatrix} y_{11} & y_{12} \\ 0 & y_{22} \end{smallmatrix}\right)$, $\mathrm{rk}(y_{22}) = j$ as in (52), which shows that $\mathcal{R}_w(j) P_s = \mathcal{R}_w(j)$ as well.

The fact that there is a unique orbit represented by w_j is a direct calculation based on the definition of $\mathcal{R}_w(j)$ which implies that x_{21} has rank j, and the last $r - k$ columns of x_{21} are identically 0. $\qquad \square$

Lemma 4.6. *In the notation above*

$$R_w(j) = |\mathcal{R}_w(j)| = c_{n-r,k}(j) c_{r,r-j}(r - j) q^{r(j+n-r)} |\mathrm{GL}_{n-r}(\mathbb{F}_q)|,$$

where $c_{k,l}(j) = \left|\{x \in \mathbb{F}_q^{k \times l} \mid \mathrm{rk}(x) = j\}\right|$.

Proof. Assume that $x = \left(\begin{smallmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{smallmatrix}\right) \in \mathcal{R}_w(j)$, and that $x' = \left(\begin{smallmatrix} x_{11} & x_{12}' \\ x_{21} & x_{22}' \end{smallmatrix}\right) \in \mathrm{GL}_n(\mathbb{F}_q)$. Then $x' \in \mathcal{R}_w(j)$ as well, and there is $u \in U_r$, $h \in H_2$, such that $x' = xuh$.

It follows that $R_w(j) = q^{r(n-r)} |\mathrm{GL}_n(\mathbb{F}_q)| \cdot |\mathcal{R}_w'(j)|$, where $\mathcal{R}_w'(j)$ consists of those $n \times r$ matrices $\left\{\left(\begin{smallmatrix} x_{11} \\ x_{21} \end{smallmatrix}\right)\right\}$, which have rank r, and for which the $(n-r) \times r$ matrix x_{21} is such that it has rank j and its last $r - k$ columns are identically 0.

The number of choices for x_{21} for $x = \left(\begin{smallmatrix} x_{11} \\ x_{21} \end{smallmatrix}\right) \in \mathcal{R}_w'(j)$ is $c_{n-r,k}(j)$.

From the transitivity $\mathcal{R}_w(j) = P_r \mathcal{R}_w(j)$ in Lemma 4.5 for each x_{21} there are the same number of possible x_{11}. For x_{21} the matrix with a I_j in the top left corner and zeros everywhere else, it is readily seen that the number of possible x_{11}'s is $q^{rj} c_{r,r-j}(r - j)$ which proves the claim. $\qquad \square$

In view of the description in Lemma 4.5 it may seem that Proposition 4.4 is wasteful and a more exact evaluation is possible. While it is true that one can push this approach to get more precise information, we will see below that there are cases when the estimates are of the right order of magnitude. Still we will use some enumerative combinatorics, but merely for getting a good constant to match the q-power in the estimate that arises from Proposition 4.4. To do this it is convenient to use the Gaussian binomial coefficients (q binomials) [Cameron 2017]. For $k \in \mathbb{N}$ let

$$[k]_q = \frac{q^k - 1}{q - 1}, \quad [k]_q! = \prod_{j=1}^{k} [j]_q, \quad \text{and} \quad \binom{k}{l}_q = \frac{[k]_q!}{[l]_q! \cdot [k - l]_q!}.$$

With this notation we have $|\mathrm{GL}_k(\mathbb{F}_q)| = (q - 1)^k q^{\binom{k}{2}} [k]_q!$ and the number of matrices of fixed size and rank ([Landsberg 1893, Formula $(\mathcal{B})$]; see also [Morrison 2006, Section 1.7])

$$c_{k,l}(j) = \left|\{x \in \mathbb{F}_q^{k \times l} \mid \mathrm{rk}(x) = j\}\right| = (q - 1)^j q^{\binom{j}{2}} \frac{[k]_q! \cdot [l]_q!}{[k - j]_q! \cdot [l - j]_q! \cdot [j]_q!}. \tag{53}$$

We may therefore rephrase Lemma 4.6 as

$$R_w(j) = |\mathcal{R}_w(j)| = (q - 1)^n q^{\binom{n}{2} + j^2} \frac{[k]_q! \cdot [r]_q! \cdot [n - r]_q!^2}{[k - j]_q! \cdot [n - r - j]_q! \cdot [j]_q!^2}. \tag{54}$$

(Here $w = w_k \in \overline{W}_r$ and $\mathcal{R}_w(j)$ is as in (52).)

Proof of Theorem 1.10. By Lemma 4.6, (53) and (54) we have that the summands in Proposition 4.4 are equal to

$$(q - 1)^{n-j} q^{\binom{n}{2} + j^2} \frac{[k]_q! \cdot [r]_q! \cdot [n - r]_q!}{[k - j]_q! \cdot [j]_q!^2},$$

and thus

$$|K_n(E_r d, E_r w)| \leq (q - 1)^{n-r} q^{\binom{n}{2}} [n - r]_q! \sum_{j=0}^{k} q^{j^2} \binom{k}{j}_q \prod_{j'=j+1}^{r} (q^{j'} - 1).$$

Using the trivial identity $q^{j'} - 1 \leq q^{j'}$ we have for the inner sum

$$\sum_{j=0}^{k} q^{j^2} \binom{k}{j}_q \prod_{j'=j+1}^{r} (q^{j'} - 1) < \sum_{j=0}^{k} q^{j^2} \binom{k}{j}_q q^{\binom{r+1}{2} - \binom{j+1}{2}} = q^{\binom{r+1}{2}} \sum_{j=0}^{k} q^{\binom{j}{2}} \binom{k}{j}_q = 2q^{\binom{r+1}{2}} \prod_{j=1}^{k-1} (1 + q^j)$$

by the q-binomial theorem [Stanley 1986, Formula (1.87)]. From this,

$$|K_n(E_r d, E_r w)| \leq 2q^{\binom{n}{2} + \binom{r+1}{2}} \prod_{j=1}^{k-1} (q^{2j} - 1) \prod_{j'=k}^{n-r} (q^{j'} - 1)$$

$$< 2q^{\binom{n}{2} + \binom{r+1}{2} + 2\binom{k}{2} + k(n-r-k+1) + \binom{n-r-k+1}{2}} = 2q^{n^2 - rn + r^2 + \binom{k}{2}}.$$

Recall that $m = \min(r, n - r)$, and thus by Lemma 4.2 we have $1 \leq k \leq m$, so

$$|K_n(E_r d, E_r w)| \leq 2q^{n^2 - rn + r^2 + \binom{m}{2}}. \qquad \square$$

Finally we prove the claim about $K_n(e_{1,n}, e_{1,n})$. Using the Bruhat decomposition with the maximal parabolic subgroup P as in (10) we get

$$K_n(e_{1,n}, e_{1,n}) = \sum_{k=1}^{n} \sum_{u \in U_k} \sum_{p \in P} \varphi((uw_{(kn)}p)_{n,1} + (uw_{(kn)}p)_{n,1}^{-1}).$$

Write $p = \begin{pmatrix} h & v \\ 0 & \lambda \end{pmatrix}$ with $h \in \mathrm{GL}_{n-1}(\mathbb{F}_q)$, $v \in \mathbb{F}_q^{(n-1) \times 1}$ and $\lambda \in \mathbb{F}_q^*$. Then

$$(uw_{(kn)}p)_{n,1} = \begin{cases} h_{k1} & \text{if } k < n, \\ 0 & \text{if } k = n, \end{cases} \qquad (uw_{(kn)}p)_{n,1}^{-1} = \begin{cases} \lambda^{-1} & \text{if } k = 1, \\ 0 & \text{if } k > 1, \end{cases}$$

and thus

$$\sum_{g \in U_k w_{(kn)} P} \varphi((x)_{n,1} + (x)_{n,1}^{-1}) = \begin{cases} -q^{n-1}|U_1| K_{n-1}(e_{1,1}, 0) & \text{if } k = 1, \\ (q-1)q^{n-1}|U_k| K_{n-1}(e_{1,k}, 0) & \text{if } 1 < k < n, \\ |P| & \text{if } k = n. \end{cases}$$

Since $K_{n-1}(e_{1,k}, 0) = K_{n-1}(e_{1,1}, 0) = -q^{n-2}|\mathrm{GL}_{n-2}(\mathbb{F}_q)|$ by Theorem 1.9, we get

$$K_n(e_{1,n}, e_{1,n}) = -q^{2n-3}|\mathrm{GL}_{n-2}(\mathbb{F}_q)| \left(-q^{n-1} + (q-1) \sum_{k=2}^{n-1} q^{n-k} \right) + (q-1)q^{n-1}|\mathrm{GL}_{n-1}(\mathbb{F}_q)|$$

$$= q^{2n-2}|\mathrm{GL}_{n-2}(\mathbb{F}_q)| + (q-1)q^{n-1}|\mathrm{GL}_{n-1}(\mathbb{F}_q)| \sim q^{n^2-n+1}.$$

5. Examples

5A. *Kloosterman sums of 2×2 matrices.* Let $a \in M_2(\mathbb{F}_q)$. Since the Kloosterman sum is invariant under conjugation, $K_2(a) = K_2(gag^{-1})$ for any $g \in \mathrm{GL}_2(\mathbb{F}_q)$, we may assume that a is in Frobenius normal form, $a = \begin{pmatrix} 0 & 1 \\ -d & t \end{pmatrix}$, where $t = \mathrm{tr}(a)$, $d = \det a$. The Kloosterman sum is then

$$K_2(a) = \sum_{x_{11}x_{22} - x_{12}x_{21} \neq 0} \varphi(-dx_{12} + tx_{22})\varphi(x_{21} + (x_{11} + x_{22})/(x_{11}x_{22} - x_{12}x_{21})).$$

From this presentation it is not at all clear that this sum should behave differently depending on whether $t^2 - 4d$ is or is not a nonzero square or 0, showing that brute force calculations without using the finer group structure are unlikely to highlight any of the features of these sums.

For $n = 2$ the maximal parabolic of Section 2B is also the Borel subgroup, and the two approaches given earlier are the same. They lead in an elementary way to the following evaluations:

a	$\begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$, $\alpha \neq \beta$	$\begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$, $\alpha \neq 0$	$\begin{pmatrix} \alpha & 1 \\ 0 & \alpha \end{pmatrix}$, $\alpha \neq 0$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$
$K_2(a)$	$q K_1(\alpha) K_1(\beta)$	$q^3 - q^2 + K_1(\alpha)^2 q$	$-q^2 + K_1(\alpha)^2 q$	q	q

The nonsplit case, $a = \begin{pmatrix} \alpha & \delta\beta \\ \beta & \alpha \end{pmatrix}$, with $\beta \neq 0$ and $\delta \notin (\mathbb{F}_q^*)^2$, can also be evaluated explicitly, although this requires some effort. We have

$$K_2(a) = -q K_1\big(\alpha + \beta\sqrt{\delta}, \mathbb{F}_{q^2}^*\big).$$

See the proof of Proposition 5.10 below.

It is also possible to deal directly with the more general sum $K_2(a, b)$. Of course if either rk(a) or rk(b) is 2, say rk$(b) = 2$, this leads to the previous evaluation by

$$K_2(a, b) = K_2(ab^{-1}, I_2).$$

If one of them, but not both are 0, say $b = 0$, and rk$(a) = 1$, then

$$K_2(a, 0) = -q(q - 1).$$

If both of a and b have rank 1, then we may assume that $a = e_1 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, and that b is one of $b_1 = \left(\begin{smallmatrix} \alpha & 0 \\ 0 & 0 \end{smallmatrix}\right)$ for which

$$K_2(e_1, b_1) = K_1(\alpha)q(q - 1),$$

or $b_2 = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ for which

$$K_2(e_1, b_2) = -q(q - 1),$$

or $b_3 = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ for which

$$K_2(e_1, b_3) = q^3 - q^2 + q.$$

Finally one trivially has that $K_2(0, 0) = (q^2 - 1)(q^2 - q)$.

5B. *The recursion in closed form.* We will describe an algorithm for calculating the polynomials that express $K_n(a)$ when $a = \alpha I_n + \bar{a}(\lambda)$ for some fixed $\alpha \neq 0$ and some partition $\lambda = [n_1, \ldots, n_l]$. As usual the fact that $n_1 + \cdots + n_l = n$ is denoted by $\lambda \vdash n$. We will rely on the notation of Sections 2B and 2C, where we made the assumption that $n_i \leq n_{i+1}$. If an element n_i repeats k times we will write $[\ldots, n_i^k, \ldots]$ instead of $[\ldots, n_i, \ldots, n_i, \ldots]$, so, for example, we will write $[1^n]$ for the partition that corresponds to the matrix αI.

Also we will denote the polynomials by K_λ as well, instead of the notation P_λ in Theorem 1.3. However we will still write K for $K_{[1]} = K_{[1]}(\alpha) = K_1(\alpha)$, so, for example, Theorem 1.4 can be stated as

$$K_{[1^n]} = q^{n-1} K K_{[1^{n-1}]} + q^{2n-2}(q^{n-1} - 1)K_{[1^{n-2}]}.$$

It is clear from the proof of Theorem 5.1 below that if $n_{l-1} < n_l$, the recursion is particularly simple, and has only two terms corresponding to the sums over the cells X_{n-1} and X_n. Therefore in the case $n_{l-1} < n_l$

$$K_\lambda = q^{n-1} K K_{\lambda'} - q^{2n-2} K_{\lambda''},$$

where

$$\lambda' = [n_1, \ldots, n_{l-1}, n_l - 1] \quad \text{and} \quad \lambda'' = [n_1, \ldots, n_{l-1}, n_l - 2]$$

possibly reordered into a monotonic sequence, if $n_{l-1} = n_l - 1$. If $n_l = 2$, then the entries corresponding to $n_l - 2 = 0$ are simply deleted. For example, for $\lambda = [1, 2]$ this gives

$$K_{[1,2]} = q^2 K K_{[1^2]} - q^4 K.$$

The situation is more interesting when the last entry is repeated. This can be handled by the following explicit recursion, which gives an alternative proof of Theorem 1.3.

Theorem 5.1 (recursion algorithm). *Assume that* $\lambda = [n_1^{k_1}, \ldots, n_{l-1}^{k_{l-1}}, n_l^{k_l}]$, *with* $k_l > 1$. *Then*

$$K_\lambda = q^{n-1} K K_{\lambda'} - q^{2n-2} K_{\lambda''} - (q^{k_l-1} - 1)q^{2n-2}(K_{\lambda''} - K_{\lambda'''}),$$

where

$$\lambda' = [n_1^{k_1}, \ldots, n_{l-1}^{k_{l-1}}, n_l - 1, n_l^{k_l-1}],$$
$$\lambda'' = [n_1^{k_1}, \ldots, n_{l-1}^{k_{l-1}}, n_l - 2, n_l^{k_l-1}],$$
$$\lambda''' = [n_1^{k_1}, \ldots, n_{l-1}^{k_{l-1}}, (n_l - 1)^2, n_l^{k_l-2}],$$

reordered into a monotonic sequence, if needed.

Note that $\lambda' \vdash n - 1$ and $\lambda'', \lambda''' \vdash n - 2$. It is essential that the "reduced" partitions $\lambda', \lambda'', \lambda'''$ are put into the canonical nondecreasing form we are using. This process is somewhat inconvenient to express in notation, but easy to do so in practice. For example, if $\lambda = [1, 2, 3^2]$ then $\lambda' = [1, 2^2, 3]$, $\lambda'' = [1^2, 2, 3]$ and $\lambda''' = [1, 2^3]$, while for $\lambda = [1, 2, 4^2]$, $\lambda' = [1, 2, 3, 4]$, $\lambda'' = [1, 2^2, 4]$ and $\lambda''' = [1, 2, 3^2]$, etc.

The algorithm can be used to express $K_n(a)$ for any a with a split characteristic polynomial when n is small, by first using Theorem 1.1. For example, for $n = 3$, $a = \alpha I + \bar{a}(\lambda)$, $\alpha \neq 0$, it gives

λ	$[1^3]$	$[1, 2]$	$[3]$
K_λ	$q^3 K^3 + (q^5 + 2q^4)(q-1)K$	$q^3 K^3 + q^4(q-2)K$	$q^3 K^3 - 2q^4 K$

The first nontrivial example when λ''' appears is $[2^2] \vdash 4$, for which

$$K_{[2^2]} = q^3 K K_{[1,2]} - q^6 K_{[2]} - q^6(q-1)(K_{[2]} - K_{[1^2]}) = q^6 K^4 + q^7(q-3)K^2 + q^8(q^2 - q + 1).$$

To see a more intricate situation we illustrate the algorithm for $n = 6$ and $\lambda = [1, 2, 3]$, when we have

$$K_{[1,2,3]} = q^5 K K_{[1,2^2]} - q^{10} K_{[1^2,2]}, \quad K_{[1,2^2]} = q^4 K_{[1^2,2]} - 2q^8 K_{[1,2]}$$

and so on.

There are many families where the recursion may be stated in simple terms, for example, if there is only one block, $\lambda = [n]$, when we have

$$K_{[n]} = K q^{n-1} K_{[n-1]} - q^{2n-2} K_{[n-2]}.$$

From this one can get a closed formula for $K_{[n]}$; see Section 5E below.

Theorem 5.1 is an easy corollary of the following two propositions. As usual we assume $\alpha \neq 0$, and $a = \alpha I + \bar{a}(\lambda)$ with $\varepsilon_j \in \{0, 1\}$ as in (22) with $\varepsilon_{n-1} = 1$ and use a'' to denote $a''_{k,\eta'}$.

Proposition 5.2. *Let* $Z = \left\{ z = \sum_{j=i+1}^{l-1} \xi_j e_{k-1, N_j - 1} \mid \xi_j \in \mathbb{F}_q \right\}$. *In the above notation,*

  (i) *$a'' + z$ and a'' are conjugate over $\mathbb{F}_q$,*

 (ii) *$a'' + z + e_{k-1, n-2}$ and $a'' + e_{k-1, n-2}$ are conjugate over $\mathbb{F}_q$.*

As an immediate corollary of (28) we get that $\sum_{x \in X_k} \psi(ax + x^{-1}) = 0$, unless $k = N_i$ for some $k < l$ such that $n_k = n_l$ and then

$$\sum_{x \in X_k} \psi(ax + x^{-1}) = q^{2n+k-l-3}(q-1)(K_{n-2}(a'') - K_{n-2}(a'' + e_{k-1,n-2})).$$

The second step in the proof is the following:

Proposition 5.3. (i) *If $n_i < n_l$ then $a'' + e_{k-1,n-2}$ and a'' are conjugate over $\mathbb{F}_q$.*

(ii) *If $n_i = n_l$, $a'' + e_{k-1,n-2}$ is conjugate to a''', where a''' is built from the partition λ''' as in (22).*

While not needed, we remark that the proposition remains true over $\mathbb{Z}$.

The proof of the claims in the propositions will use the linear transformation interpretation from Section 2F. We start with an easy observation.

Lemma 5.4. *Let $V_A = \langle v_0 \rangle \oplus \cdots \oplus \langle v_l \rangle \simeq \mathcal{C}_{n_0} \oplus \cdots \oplus \mathcal{C}_{n_l}$. If $v \in V_A$ is such that $A^k v = 0$ for $k < n_0$, then $V_A = \langle v_0 + v \rangle \oplus \cdots \oplus \langle v_l \rangle$ as well.*

Proof. One easily checks that $\langle v_0 + v \rangle \cap (\langle v_1 \rangle \oplus \cdots \oplus \langle v_l \rangle) = \{0\}$ and that $\langle v_0 + v \rangle + (\langle v_1 \rangle \oplus \cdots \oplus \langle v_l \rangle) = V_A$. $\square$

Remark 11. It follows that there is an isomorphism $\phi : V_A \to V_A$ which is trivial on $\langle v_1 \rangle \oplus \cdots \oplus \langle v_l \rangle$ and extends $v_0 \mapsto v_0 + v$. Clearly, it satisfies $A\phi = \phi A$.

The question for us is to determine how a module structure given by $A : V \to V$ changes if A is perturbed by another map $Z : V \to V$. For example, Propositions 5.2 and Proposition 5.3(i) are easy consequences of the following lemma.

Lemma 5.5. *Assume that $V_A \simeq \langle v_0 \rangle \oplus \langle v_1 \rangle \oplus \cdots \oplus \langle v_l \rangle \simeq \mathcal{C}_{n_0} \oplus \mathcal{C}_{n_1} \oplus \cdots \oplus \mathcal{C}_{n_l}$, that $Z : V \to V$ is such that for $i = 1, \ldots, l$*

$$Z(A^j v_i) = 0 \quad for \ j = 0, \ldots, n_i - 1$$

and that for $i = 0$ we have

$$Z(A^j v_0) = 0 \quad for \ j = 0, \ldots, n_0 - 2,$$

$$Z(A^{n_0 - 1} v_0) = \sum_{i=1}^{l} \xi_i A^{n_i - 1} v_i.$$

If $n_0 < n_i$ for all $i = 1, \ldots, l$ then $V_{A+Z} \simeq V_A$ as $\mathbb{F}_q[T]$-modules.

Proof. Let $v = \sum_{i=1}^{l} \xi_i A^{n_i - n_0} v_i$. Clearly $A^{n_0} v = 0$, and so by Lemma 5.4 there is an A-linear isomorphism ϕ, for which $\phi(v_0) = v_0 - v$. One easily checks that $Z \circ \phi = 0$ and so ϕ provides the claimed isomorphism. $\square$

Lemma 5.6. *Assume that $V_A \simeq \langle v_0 \rangle \oplus \langle v_1 \rangle \simeq \mathcal{C}_m \oplus \mathcal{C}_m$ and that $Z : V \to V$ is such that*

$$Z(A^j v_1) = 0 \quad \text{for } j = 0, \dots, m-1,$$

$$Z(A^j v_0) = 0 \quad \text{for } j = 0, \dots, m-2,$$

$$Z(A^{m-1} v_0) = \xi A^{m-1} v_1 \quad \text{for some } \xi \in \mathbb{F}_q^*.$$

Then $V_{A+Z} \simeq \mathcal{C}_{m-1} \oplus \mathcal{C}_{m+1}$ as $\mathbb{F}_q[T]$-modules.

Proof. It is easy to see that $(A + Z)^j v_0 = A^j v_0$, for $j = 0, \dots, m-1$ and that $(A + Z)^m v_0 = \xi A^{m-1} v_1$. Moreover if we replace v_1 by $v_1' = v_1 - \frac{1}{\xi} v_0$, then $(A + Z)^{m-1} v_1' = 0$. □

Proofs of Propositions 5.2 and 5.3. The two lemmas above give exactly this. □

5C. *Examples for the Kloosterman sums over Borel Bruhat cells.* Let B be the standard Borel subgroup of invertible upper triangular matrices, $w \in W$ an element of the Weyl group, and $C_w = BwB$. We will consider here the sums

$$K_n^{(w)}(a) = \sum_{x \in C_w} \psi(ax + x^{-1}),$$

where $a = \alpha I + \bar{a}$, with $\bar{a} \in \bar{U}$, where $\bar{U}$ is the set of strictly upper triangular matrices. We will first comment on the nature of these sums and then derive some of the properties that will be used below in the section on purity.

As in the proof of Theorem 1.3 in Section 2D one can show that these sums satisfy a recursion that connects them to similar sums of rank $n - 1$ or $n - 2$, depending on whether $w(n) = n$ or $w(n) < n$. To see this, recall from Proposition 3.14 that

$$K_n^{(w)}(a) = q^{n_a + n_o} K_1(\alpha)^f S_n^{(w)}(\bar{a}), \tag{55}$$

where n_o is the number of involution pairs in w, f is the number of fixed points of w and where the auxiliary sum is given by

$$S_n^{(w)}(\bar{a}) = \sum_{v, d, u_1} \psi(\bar{a}^v w d u),$$

with $v \in U^\flat$, $u \in U_b^\sharp$, $d \in D(w) = \{ \sum_{i < w(i)} t_i \boldsymbol{e}_{i,i} \mid t_i \in \mathbb{F}_q^* \}$ and $\bar{a}^v = v^{-1} \bar{a} v$. The recursion then proceeds on the sum $S_n^{(w)}(a)$. For example, when $w(n) = k < n$ we again let $m'' = m''_{\not{k}, \not{n}}$ denote the matrix one gets by deleting the k-th and n-th rows and columns of an $n \times n$ matrix m. To describe the set of perturbations $Z \subset M_{n-2}$ that arise in the reduction, we let $\bar{a}_{(i)}$ and $\bar{a}^{(i)}$ denote the i-th row, respectively the i-th column, of $\bar{a}$ and consider the set

$$Y = \{ y \in \mathbb{F}_q^n \mid y_i = 0 \text{ for } i \leq k \text{ and } y\bar{a} = \bar{a}_{(k)} \}.$$

Using this notation we have the following proposition whose proof goes along the lines of Proposition 2.8 and is omitted here.

Proposition 5.7. *The following equation holds*:

$$S_n^{(w)}(\bar{a}) = q^{n-k-1} \sum_{y \in Y} S_{n-2}^{(w'')}(\bar{a}'' + (\bar{a}^{(k)}y)'') \sum_{t \in \mathbb{F}_q^*} \varphi(ty\bar{a}^{(n)}). \tag{56}$$

The set Y may be empty, and then so is the set of perturbations which arise from the collection of rank-1 matrices

$$Z = \{\bar{a}^{(k)}y \in M_n \mid y \in Y\},$$

in which case the sum is interpreted as 0.

What makes the sums $S_n^{(w)}(\bar{a})$ harder to deal with is that as functions of a they are no longer invariant under conjugation by GL_n. They are invariant under conjugation by elements of B in virtue of (55) and the fact that

$$K_n^{(w)}(a) = K_n^{(w)}(b^{-1}ab)$$

for any $b \in B$, since $b^{-1}C_w b = C_w$. However the B-orbits in the set $\overline{U}$ of strictly upper triangular matrices under the adjoint action are not well understood. It is easy to see that one can no longer straighten out partitions into a nondecreasing order, or even assume that an orbit is represented by a matrix in Jordan normal form. For example, $\{te_{1,3} \mid t \neq 0\}$ is one of the orbits in the 3×3 case. When $n = 6$, there is even a one-parameter family of orbits, found by Kashin [1990]. In general a full description of the orbits is hard even in low ranks [Bürgstein and Hesselink 1987; Hille and Röhrle 1997].

Returning to the sums $K_n^{(w)}(a)$, it is still quite likely that these can be expressed as polynomials in q and K, independently of the characteristic p since the perturbations arising in the reduction calculations are of a very special nature. In what follows we will provide some low-rank examples, when the independence is easy to establish directly. We will do this by stating certain special cases when the above reduction is sufficient, most importantly the case when $w = (ij)$. There are a number of other special cases when the reduction for $S_n^{(w)}(a)$ can be treated in a simple manner; for example, when $w(n) = 1$ or $n - 1$.

Again we assume that $a = \alpha I + \bar{a}$ is of the form as in (9), $\bar{a} = \sum_{j=1}^{n-1} \varepsilon_j e_{j,j+1}$, and $\lambda \vdash n$ is the partition corresponding to a. Define

$$K_\lambda^{(w)}(\alpha) = \sum_{x \in C_w} \psi(ax + x^{-1}).$$

Theorem 5.8. *Let $i < j$ and $(ij) \in W$ and $\alpha \neq 0$. Then*

$$K_\lambda^{(ij)}(\alpha) = q^{n(n-1)/2} K^{n-2} \cdot \begin{cases} (q-1)q & \text{if } j = i+1 \text{ and } \varepsilon_i = 0, \\ -q & \text{if } j = i+1 \text{ and } \varepsilon_i = 1, \\ (q-1)q^{j-i-d} & \text{if } j > i+1 \text{ and } \varepsilon_i = \varepsilon_{j-1} = 0, \end{cases}$$

where $d = |\{k \in \mathbb{N} \mid i < k < j-1 \text{ and } \varepsilon_k \neq 0\}|$.

If $j > i+1$ and either ε_i or $\varepsilon_{j-1} \neq 0$ then $K_\lambda^{(ij)}(\alpha) = 0$.

This allows us to compute the Bruhat cell polynomials for $n \leq 3$; see Tables 1 and 2.

λ	$w = \;$ (12)	I
[1, 1]	$q^2(q-1)$	qK^2
[2]	$-q^2$	qK^2

Table 1. The $n = 2$ case.

λ	$w = \;$ (13)	(12)	(23)	I
[1, 1, 1]	$q^5(q-1)K$	$q^4(q-1)K$	$q^4(q-1)K$	q^3K^3
[2, 1]	0	$q^4(q-1)K$	$-q^4K$	q^3K^3
[1, 2]	0	$-q^4K$	$q^4(q-1)K$	q^3K^3
[3]	0	$-q^4K$	$-q^4K$	q^3K^3

Table 2. The $n = 3$ case.

With a little more work one can calculate all the Bruhat cell polynomials for $n = 4$. We summarize the result in Tables 3 and 4.

The polynomials are not merely a permutation for different rearrangements of a partition; see, for example, the case $\lambda = [1, 2, 1]$ and $w = (14)(23)$.

We finish this section by giving the cell polynomials for the full block ($\lambda = [n]$) case for general n. Let $w = (i_1, j_1)(i_2, j_2) \cdots (i_r, j_r) \in W$ such that $i_k < j_k$ for any k. Then

$$K_{[n]}^w(\alpha) = \begin{cases} (-1)^r K^{n-2r} q^{n(n-1)/2+r} & \text{if } j_k - i_k = 1 \text{ for any } k, \\ 0 & \text{otherwise.} \end{cases}$$

λ	$w = \;$ (14)(23)	(13)(24)	(12)(34)	(14)	(13)
[1, 1, 1, 1]	$q^{10}(q-1)^2$	$q^9(q-1)^2$	$q^8(q-1)^2$	$q^9(q-1)K^2$	$q^8(q-1)K^2$
[2, 1, 1]	0	0	$-q^8(q-1)$	0	0
[1, 2, 1]	$-q^9(q-1)$	0	$q^8(q-1)^2$	$q^8(q-1)K^2$	0
[1, 1, 2]	0	0	$-q^8(q-1)$	0	$q^8(q-1)K^2$
[3, 1]	0	0	$-q^8(q-1)$	0	0
[2, 2]	0	$q^9(q-1)$	q^8	0	0
[1, 3]	0	0	$-q^8(q-1)$	0	0
[4]	0	0	q^8	0	0

Table 3. The $n = 4$ case (continued below).

λ	$w = \ (24)$	(12)	(23)	(34)	I
$[1,1,1,1]$	$q^8(q-1)K^2$	$q^7(q-1)K^2$	$q^7(q-1)K^2$	$q^7(q-1)K^2$	q^6K^4
$[2,1,1]$	$q^8(q-1)K^2$	$-q^7K^2$	$q^7(q-1)K^2$	$q^7(q-1)K^2$	q^6K^4
$[1,2,1]$	0	$q^7(q-1)K^2$	$-q^7K^2$	$q^7(q-1)K^2$	q^6K^4
$[1,1,2]$	0	$q^7(q-1)K^2$	$q^7(q-1)K^2$	$-q^7K^2$	q^6K^4
$[3,1]$	0	$-q^7K^2$	$-q^7K^2$	$q^7(q-1)K^2$	q^6K^4
$[2,2]$	0	$-q^7K^2$	$q^7(q-1)K^2$	$-q^7K^2$	q^6K^4
$[1,3]$	0	$q^7(q-1)K^2$	$-q^7K^2$	$-q^7K^2$	q^6K^4
$[4]$	0	$-q^7K^2$	$-q^7K^2$	$-q^7K^2$	q^6K^4

Table 4. The $n = 4$ case (continued).

Since this is merely for illustration we only give a sketch of the argument. For those $w \in W$ such that $K_n^w(\alpha) = 0$ one can find $(i, j) \in I$ such that $\mathrm{tr}(v^{-1}avwdu)$ is nonconstant and linear in $v_{i,j}$.

5D. *Regular semisimple matrices.* We have seen that for an $n \times n$ matrix a, whose characteristic polynomial P_a have no multiple roots, the cohomology associated to the Kloosterman sum $K_n(a)$ is pure. Assume now that this characteristic polynomial P_a is irreducible over $\mathbb{F}_q$. Let $\alpha \in \mathbb{F}_{q^n}$ be an eigenvalue of a, $P_a(\alpha) = 0$. The argument in Section 3E shows that over $\mathbb{F}_{q^n}$

$$H^\bullet = H_c^\bullet(\mathrm{GL}_n, x \mapsto \mathrm{tr}(ax + x^{-1})) = \left(\bigotimes_{i=1}^{n(n-1)/2} H_c^\bullet(\mathbb{A}^1, 0) \right) \otimes \left(\bigotimes_{j=1}^{n} H_c^\bullet(\mathbb{A}^1 \setminus \mathbb{A}^0, f_\alpha) \right),$$

where for $x \in (\mathbb{A}^1 \setminus \mathbb{A}^0)(\mathbb{F}_{q^n})$, $f_\alpha(x) = \alpha x + x^{-1}$, that corresponds to the scalar Kloosterman sum $K_1(\alpha, \mathbb{F}_{q^n}) = \lambda_1 + \lambda_2$. Here λ_1 and $\lambda_2 = \bar{\lambda}_1$ are the $\mathbb{F}_{q^n}$-Frobenius eigenvalues on $H_c^\bullet(\mathbb{A}^1 \setminus \mathbb{A}^0, f_\alpha)$.

Then, clearly, over $\mathbb{F}_{q^n}$ the Frobenius eigenvalues on the 2^n-dimensional space $\bigotimes_{j=1}^n H_c^1(\mathbb{A}^1 \setminus \mathbb{A}^0, f_\alpha)$ are of the form $\left(\prod_{i \in I} \lambda_1 \right) \left(\prod_{i \notin I} \lambda_2 \right) = \lambda_1^{|I|} \lambda_2^{n-|I|}$, where $I \subset \{1, \dots, n\}$ — therefore each of $\lambda_1^{|I|} \lambda_2^{n-|I|}$ has multiplicity $\binom{n}{j}$. If we fix some n-th roots of the λ_i, say $\eta_i^n = \lambda_i$, then we have that the Frobenius eigenvalues on H^{n^2} are of the form $\zeta_I q^{n(n-1)/2} \eta_1^{|I|} \eta_2^{n-|I|}$ where again $I \subset \{1, \dots, n\}$, and the ζ_I are n-th roots of unity, $\zeta_I^n = 1$ for all I. It is natural to make the following conjecture.[2]

Conjecture 5.9. If p is large enough, and the characteristic polynomial P_a is irreducible over $\mathbb{F}_q$ then

$$K_n(a, \mathbb{F}_q) = (-1)^{n+1} q^{n(n-1)/2} K_1(\alpha, \mathbb{F}_{q^n}).$$

The conjecture would follow if, for $I = \varnothing$ and $\{1, \dots, n\}$, the $\mathbb{F}_q$-Frobenius eigenvalues were $(-1)^{n+1} q^{n(n-1)/2} \lambda_1$, $(-1)^{n+1} q^{n(n-1)/2} \lambda_2$ and the others canceled after summing. For example, when

---

[2]While this paper was in print, Elad Zelingher [2023] announced a proof of this conjecture.

$n = 3$, this can happen if the eigenvalues μ_i for $1 \leq i \leq 8$ are the eight summands in the expansion of the product $(\eta_1 + \eta_2)(\omega\eta_1 + \omega^2\eta_2)(\omega^2\eta_1 + \omega\eta_2)q^3$ for $\omega^3 = 1$; this leads to

$$K_n(a, \mathbb{F}_q) = \sum_{i=1}^{8} \mu_i = q^3(\lambda_1 + \lambda_2) = (-1)^4 q^3 K_1(\alpha, \mathbb{F}_{q^3}),$$

exactly as desired.

The conjecture is partly based on the observation that if we let $K = \mathbb{F}_q[a] \subset M_n$, then K is a field naturally isomorphic to $\mathbb{F}_{q^n}$. K acts on M_n by left multiplication and it is easy to describe the K-algebra that arises for any n. We will use this below to handle the case $n = 2$, but such elementary methods get cumbersome and are unlikely to give a proof, or even offer any insight already for $n = 3$.

We now give a few numerical examples for $M_n(\mathbb{F}_{p^n})$ for small n and p checked with computer algebra systems pari/gp and Sage.

(i) Let $n = 3$, $p = 5$ and $\alpha \in \mathbb{F}_{125}$ be one of the roots of $x^3 + x^2 + 1$. Then $K_1(\alpha, \mathbb{F}_{125}) = (3 + \sqrt{5})/2$. On the other hand if

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & -1 \end{bmatrix}$$

then $K_3(A, \mathbb{F}_5) = 327.2542$ which agrees with $125 K_1(\alpha, \mathbb{F}_{125})$.

(ii) Let $n = 4$, $p = 3$ and $\alpha \in \mathbb{F}_{81}$ be a root of $x^4 + 2x^3 + 2$, and

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

Then $K_4(A, \mathbb{F}_3) = 11664$ which agrees with $-729 K_1(\alpha, \mathbb{F}_{81})$.

(iii) Let $n = 3$, $p \equiv 1$ (3) and

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \mu & 0 & 0 \end{bmatrix},$$

where $\mu \in \mathbb{F}_p^* \setminus (\mathbb{F}_p^*)^3$. If $\alpha \in \mathbb{F}_{p^3}$ is such that $\alpha^3 = \mu$ then one can check that

$$K(\alpha) = \sum e(3\mu c + (3a^2 - 3\mu cb)/\Delta(a, b, c)),$$

where $e(x) = e^{2\pi i x}$, $\Delta(a, b, c) = a^3 - 3\mu cba + (\mu b^3 + \mu^2 c^3)$ and where the sum is over $(a, b, c) \in \mathbb{F}_p^3 \setminus \{(0, 0, 0)\}$.

A direct calculation using Bruhat decomposition shows that the conjecture in this case is equivalent to

$$K(\alpha) = K_{13}(A) + (1 + (\mu/p))q,$$

where (μ/p) is the Legendre symbol and where

$$K_{13}(A) = \sum_{t_1,t_2,t_3 \in \mathbb{F}_p^*} e(\mu t_1^2 t_3^2 + t_1^2 t_3 + \mu t_3 + 1/t_2 - 1/(\mu t_1 t_3^2)).$$

Up to about $p \leq 200$, this can be checked fast even on a personal computer. For example, the order of 2 mod 199 is 99, and we have that

$$K\left(\sqrt[3]{2}, \mathbb{F}_{199^3}\right) = 3869.8269,$$

while for

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 2 & 0 & 0 \end{pmatrix}$$

we have

$$K_{13}(A) = 4267.8269$$

with a difference of 398, which shows that $K_3(A, \mathbb{F}_{199}) = K\left(\sqrt[3]{2}, \mathbb{F}_{199^3}\right)$. On the other hand 3 mod 199 is a primitive root, and so for

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 3 & 0 & 0 \end{pmatrix}$$

we have

$$K\left(\sqrt[3]{3}, \mathbb{F}_{199^3}\right) = K_{13}(A) = -2875.1994.$$

We also checked all $p = 3k + 1 \leq 200$, for which 2 is not a cube mod p and found that $K\left(\sqrt[3]{2}, \mathbb{F}_{p^3}\right) = K_3(A, \mathbb{F}_p)$ holds for all of them.

Proposition 5.10. *Assume that q is odd and let $\alpha \in \mathbb{F}_{q^2} \setminus \mathbb{F}_q$ and $a = \left(\begin{smallmatrix} 0 & 1 \\ -N(\alpha) & -\mathrm{Tr}(\alpha) \end{smallmatrix}\right)$, where N and Tr are the norm and trace of the field extension $\mathbb{F}_{q^2}/\mathbb{F}_q$. Then*

$$K_2(a, M_2(\mathbb{F}_q)) = -q K_1(\alpha, \mathbb{F}_{q^2}^*).$$

Remark 12. The cohomology complex $H_c^\bullet$ corresponding to the sum $K_2(a, M_2(\mathbb{F}_q))$ satisfies

$$\dim H_c^i = \begin{cases} 0 & \text{if } i \neq 4, \\ 4 & \text{if } i = 4, \end{cases}$$

and the Frobenius eigenvalues μ_i on H_c^4 satisfy $\mu_1^2 = q^2 \lambda_1^2$, $\mu_2^2 = q^4$, $\mu_3^2 = q^4$, and $\mu_4^2 = q^2 \lambda_2^2$ where $\lambda_2 = \bar{\lambda}_1$ are the eigenvalues corresponding to $K_1(\alpha, \mathbb{F}_{q^2}^*)$. Apart from permutations the proposition determines the sign of the square roots, we have $\mu_1 = -q\lambda_1$, $\mu_2 = q^2$, $\mu_3 = -q^2$ and $\mu_4 = -q\lambda_2$. Thus $K_2(a, M_2(\mathbb{F}_{q^m})) = \sum_{i=1}^4 \mu_i^m$. Here again we have cancellation: $\mu_2^m + \mu_3^m = 0 \iff 2 \nmid m$.

Proof of Proposition 5.10. As above let $K = \mathbb{F}_q[a]$ be the subring of M_2 generated by a. K is isomorphic to $\mathbb{F}_{q^2}$ by the assumption on α. Also, the vector space $\mathbb{F}_q^2$ as an $\mathbb{F}_q[a]$-module is isomorphic to the $\mathbb{F}_q$ vector space $\mathbb{F}_{q^2}$, with a acting via multiplication by α. For the moment denote this action by L_α, $L_\alpha : \beta \mapsto \alpha\beta$.

Let $\mathbb{F}_{q^2}\langle\tau\rangle$ be the noncommutative ring of twisted polynomials, $\sum_i \xi_i \tau^i$, subject to $\tau\xi = F(\xi)\tau$, where $F(\xi) = \xi^q$ is the Frobenius automorphism of $\mathbb{F}_{q^2}/\mathbb{F}_q$.

There is an obvious map from $\mathcal{M}_2 = \mathbb{F}_{q^2}\langle\tau\rangle/(\tau^2 - 1)$ to $M_2(\mathbb{F}_q)$, sending $\xi_1 + \xi_2\tau$ to the $\mathbb{F}_q$-linear transformation $L_{\xi_0} + L_{\xi_1} F$. It is not difficult to see that this linear map is injective, and so by dimension count, an isomorphism. This identifies M_2 with $\mathcal{M}_2$, and it is easy to check that under this identification $\psi(L_{\xi_0} + L_{\xi_1} F) = \varphi_2(\xi_0)$, where $\varphi_2 = \varphi \circ \mathrm{Tr}_{\mathbb{F}_{q^2}/\mathbb{F}_q}$. It follows that

$$\sum_{x\in M_2(\mathbb{F}_q)^*} \psi(ax + x^{-1}) = \sum_{\xi_0+\xi_1\tau\in\mathcal{M}_2^*} \varphi_2(\alpha\xi_0 + (\xi_0 + \xi_1\tau)^{-1}).$$

An easy calculations shows that $(1 + \xi\tau) \in \mathcal{M}_2^*$ exactly when $N(\xi) \neq 1$, and then $(1 + \xi\tau)^{-1} = \frac{1}{1-N(\xi)}(1 - \xi\tau)$. One also has that $(\xi\tau)^{-1} = F(\xi^{-1})\tau$ and so

$$\mathcal{M}_2^* = \{\xi_1\tau \mid \xi_1 \in \mathbb{F}_{q^2}^*\} \cup \{\xi_0(1 + \xi_1\tau) \mid \xi_0 \in \mathbb{F}_{q^2}^*, \; \xi_1 \in \mathbb{F}_{q^2}, \; N(\xi_1) \neq 1\}.$$

Therefore

$$\sum_{x\in\mathrm{GL}_2(\mathbb{F}_q)} \psi(ax + x^{-1}) = q^2 - 1 + \sum_{\substack{\xi_0\in\mathbb{F}_{q^2}^* \\ \xi_1\in\mathbb{F}_{q^2} \\ N(\xi_1)\neq 1}} \varphi_2\big(\alpha\xi_0 + (1 - N(\xi_1))^{-1}\xi_0^{-1}\big).$$

Now the norm map N is a surjective homomorphism from $\mathbb{F}_{q^2}^* \to \mathbb{F}_q^*$, with a kernel of size $q + 1$ and so for $\gamma \in \mathbb{F}_q$

$$\big|\{\xi \in \mathbb{F}_{q^2} \mid (1 - N(\xi))^{-1} = \gamma\}\big| = \begin{cases} 0 & \text{if } \gamma = 0, \\ 1 & \text{if } \gamma = 1, \\ q+1 & \text{if } \gamma \neq 0, 1. \end{cases}$$

This gives

$$\sum_{x\in\mathrm{GL}_2(\mathbb{F}_q)} \psi(ax + x^{-1}) = q^2 - 1 + (q+1) \sum_{\substack{\xi_0\in\mathbb{F}_{q^2}^* \\ \gamma\in\mathbb{F}_q^*}} \varphi_2(\alpha\xi_0 + \gamma\xi_0^{-1}) - q \sum_{\xi_0\in\mathbb{F}_{q^2}^*} \varphi_2(\alpha\xi_0 + \xi_0^{-1}).$$

Finally

$$\sum_{\substack{\xi_0\in\mathbb{F}_{q^2}^* \\ \gamma\in\mathbb{F}_q^*}} \varphi_2(\alpha\xi_0 + \gamma\xi_0^{-1}) = \sum_{\substack{\xi_0\in\mathbb{F}_{q^2}^* \\ \gamma\in\mathbb{F}_q^*}} \varphi_2(\alpha\xi_0)\phi(\gamma\,\mathrm{Tr}\,\xi_0^{-1}) = -\sum_{\substack{\xi_0\in\mathbb{F}_{q^2}^* \\ \mathrm{Tr}\,\xi_0^{-1}=0}} \varphi_2(\alpha\xi_0) = -(q-1). \qquad\square$$

5E. *The purity locus.* We have seen that for a regular semisimple element $a \in M_n(\mathbb{F}_q)$ the Kloosterman sum $K_n(a)$ is pure. The tables above already suggest that for a matrix a with more than one Jordan block for an eigenvalue, that sum cannot be pure. This can be seen without reference to cohomology. To see this assume that a has a single eigenvalue α. Recall that

$$K_n(a) = P(A, G, K) = \sum_{2f\leq n} c_f(q, q - 1)K^{n-2f},$$

where c_f are polynomials. We give the A and G weight 1, and K weight $\frac{1}{2}$, so the polynomial P has a weighted degree, which determines the order of magnitude (in q) of its value. Now $f = 1$ corresponds to simple transpositions, and by Theorem 5.8 one sees that these sums are too large in magnitude to be pure if not all ε_i are 0.

It is an intriguing question what happens for $K_{[n]}$ when a has only one Jordan block. The recursion formula gives

$$K_{[n]} = q^{n-1} K K_{[n-1]} - q^{2n-2} K_{[n-2]},$$

where $K = K_1(\alpha)$. Let $k_n = q^{-n(n-1)/2} K_{[n]}$, so that we have

$$k_n = K k_{n-1} - q k_{n-2}.$$

It follows that there exist c_1, c_2 such that

$$k_n = c_1 \lambda_1^n + c_2 \lambda_2^n,$$

where λ_1, λ_2 are the roots of $X^2 - KX + q$. These are exactly the eigenvalues of Frobenius acting on the cohomology of the Kloosterman sheaf. Using that $K = \lambda_1 + \lambda_2$, and that $K_{[2]} = -q^2 + K_1^2 q$ we get that $k_1 = \lambda_1 + \lambda_2$, and that $k_2 = k_1^2 - q = \lambda_1^2 + \lambda_1 \lambda_2 + \lambda_2^2$. Therefore $c_1 = \frac{\lambda_1}{\lambda_1 - \lambda_2}$, $c_2 = -\frac{\lambda_2}{\lambda_1 - \lambda_2}$ and

$$k_n = \frac{\lambda_1^{n+1} - \lambda_2^{n+1}}{\lambda_1 - \lambda_2} = \sum_{j=0}^{n} \lambda_1^j \lambda_2^{n-j}. \tag{57}$$

This evaluation has an interesting interpretation. Let $X^2 - K_1 X + q = (X - \lambda_1)(X - \lambda_2)$, with $\lambda_{1,2} = q^{1/2} e^{\pm i\theta}$, so that $K_1 = 2q^{1/2} \cos\theta$. We have that

$$k_n = \frac{\lambda_1^{n+1} - \lambda_2^{n+1}}{\lambda_1 - \lambda_2} = q^{n/2} \frac{e^{(n+1)\theta} - e^{-(n+1)\theta}}{e^\theta - e^{-\theta}} = \frac{\sin(n+1)\theta}{\sin\theta}.$$

Therefore

$$K_{[n]} = q^{n(n-1)/2} U_n(\cos\theta), \tag{58}$$

where U_n is the Chebyshev polynomial of the second kind. The Sato–Tate distribution of the angles of $K_1(\alpha)$ over the valuations of a global field is then equivalent to nontrivial cancellation in the sums

$$\sum_{N(v) \leq x} K_n(a, \mathbb{F}_v) / N(v)^{n(n-1)/2},$$

where $a = \alpha I + \sum_{i=1}^{n-1} e_{i,i+1}$.

Getting back to the question of purity these sums are pure from a numerical point of view, but this in itself does not rule out a cohomology with a nilpotent Frobenius action.

For example, in the case $n = 2$ it is easy to see that the cohomologies corresponding to the Bruhat cells are as follows.

On $C_I = B$ the trace of $\alpha x + x^{-1}$ can be written as a product of two Kloosterman sums over the diagonal elements; thus we have

$$H_{C_I}^\bullet = H^\bullet(C_I, x \mapsto \mathrm{tr}(ax + x^{-1})) = H_c^\bullet(\mathbb{A}^1 - \mathbb{A}^0, f_\alpha) \otimes H_c^\bullet(\mathbb{A}^1 - \mathbb{A}^0, f_\alpha) \otimes H_c^\bullet(\mathbb{A}^1, 0).$$

That implies $\dim H_{C_I}^i = 0$ unless $i = 4$ and $\dim H_{C_I}^4 = 4$.

On the nontrivial cell $C_w = UwB$ we have seen that the sum (and the cohomology) cancels on the subvariety $\alpha \det b \neq 1$ and on the rest we have

$$H_{C_w}^\bullet = H^\bullet(C_w, x \mapsto \mathrm{tr}(ax + x^{-1})) = H_c^\bullet(\mathbb{A}^1 - \mathbb{A}^0, \mathrm{id}) \otimes H_c^\bullet(\mathbb{A}^1, 0) \otimes H_c^\bullet(\mathbb{A}^1, 0).$$

That implies $\dim H_{C_w}^i = 0$ unless $i = 5$ and $\dim H_{C_w}^5 = 1$.

Thus the long exact sequence of the excision ($C_w = G \setminus C_I$) gives

$$0 \to H_G^4 \to H_{C_I}^4 \to H_{C_w}^5 \to H_G^5 \to 0.$$

Either $\dim H_G^4 = 4$ and $\dim H_G^5 = 1$ or $\dim H_G^4 = 3$ and $\dim H_G^5 = 0$ seems to be possible.

The same problem exists for higher-degree cases.

Acknowledgements

References

[Adolphson and Sperber 1989] A. Adolphson and S. Sperber, "Exponential sums and Newton polyhedra: cohomology and estimates", *Ann. of Math.* (2) **130**:2 (1989), 367–406. MR Zbl

[Borel 1991] A. Borel, *Linear algebraic groups*, 2nd ed., Grad. Texts in Math. **126**, Springer, 1991. MR Zbl

[Bürgstein and Hesselink 1987] H. Bürgstein and W. H. Hesselink, "Algorithmic orbit classification for some Borel group actions", *Compos. Math.* **61**:1 (1987), 3–41. MR Zbl

[Cameron 2017] P. J. Cameron, *Notes on counting: an introduction to enumerative combinatorics*, Austral. Math. Soc. Lect. Ser. **26**, Cambridge Univ. Press, 2017. MR Zbl

[Carlitz 1969] L. Carlitz, "Kloosterman sums and finite field extensions", *Acta Arith.* **16** (1969), 179–193. MR Zbl

[Chae and Kim 2003] H.-j. Chae and D. S. Kim, "*L* functions of some exponential sums of finite classical groups", *Math. Ann.* **326**:3 (2003), 479–487. MR Zbl

[Dabrowski 1993] R. Dabrowski, "Kloosterman sums for Chevalley groups", *Trans. Amer. Math. Soc.* **337**:2 (1993), 757–769. MR Zbl

[Dabrowski and Reeder 1998] R. Dąbrowski and M. Reeder, "Kloosterman sets in reductive groups", *J. Number Theory* **73**:2 (1998), 228–255. MR Zbl

[Deligne 1980] P. Deligne, "La conjecture de Weil, II", *Inst. Hautes Études Sci. Publ. Math.* **52** (1980), 137–252. MR Zbl

[Denef and Loeser 1991] J. Denef and F. Loeser, "Weights of exponential sums, intersection cohomology, and Newton polyhedra", *Invent. Math.* **106**:2 (1991), 275–294. MR Zbl

[Deshouillers and Iwaniec 1982] J.-M. Deshouillers and H. Iwaniec, "Kloosterman sums and Fourier coefficients of cusp forms", *Invent. Math.* **70**:2 (1982), 219–288. MR Zbl

[Einsiedler et al. 2016] M. Einsiedler, S. Mozes, N. Shah, and U. Shapira, "Equidistribution of primitive rational points on expanding horospheres", *Compos. Math.* **152**:4 (2016), 667–692. MR Zbl

[El-Baz et al. 2022] D. El-Baz, M. Lee, and A. Strömbergsson, "Effective equidistribution of primitive rational points on expanding horospheres", preprint, 2022. arXiv 2212.07408

[Erdélyi et al. 2024a] M. Erdélyi, W. Sawin, and Á. Tóth, "The purity locus of matrix Kloosterman sums", *Trans. Amer. Math. Soc.* **377**:6 (2024), 4117–4132. MR Zbl

[Erdélyi et al. 2024b] M. Erdélyi, Á. Tóth, and G. Zábrádi, "Matrix Kloosterman sums modulo prime powers", *Math. Z.* **306**:4 (2024), art. id. 68. MR Zbl

[Fouvry and Katz 2001] É. Fouvry and N. Katz, "A general stratification theorem for exponential sums, and applications", *J. Reine Angew. Math.* **540** (2001), 115–166. MR Zbl

[Fresán and Jossen 2020] J. Fresán and P. Jossen, "Exponential motives", preprint, 2020, available at https://tinyurl.com/expmot.

[Friedberg 1987] S. Friedberg, "Poincaré series for GL(n): Fourier expansion, Kloosterman sums, and algebreo-geometric estimates", *Math. Z.* **196**:2 (1987), 165–188. MR Zbl

[Fulman 2001] J. Fulman, "A new bound for Kloosterman sums", preprint, 2001. arXiv math/0105172

[Goldfeld and Sarnak 1983] D. Goldfeld and P. Sarnak, "Sums of Kloosterman sums", *Invent. Math.* **71**:2 (1983), 243–250. MR Zbl

[Grothendieck 1965] A. Grothendieck, "Formule de Lefschetz et rationalité des fonctions L", exposé 279 in *Séminaire Bourbaki*, 1964/1965, Benjamin, Amsterdam, 1965. Reprinted as pp. 41–55 in *Séminaire Bourbaki* **9**, Soc. Math. France, Paris, 1995. MR Zbl

[Hasse 1935] H. Hasse, "Theorie der relativ-zyklischen algebraischen Funktionenkörper, insbesondere bei endlichem Konstantenkörper", *J. Reine Angew. Math.* **172** (1935), 37–54. MR Zbl

[Heath-Brown 2000] D. R. Heath-Brown, "Arithmetic applications of Kloosterman sums", *Nieuw Arch. Wiskd.* (5) **1**:4 (2000), 380–384. MR Zbl

[Herz 1955] C. S. Herz, "Bessel functions of matrix argument", *Ann. of Math.* (2) **61** (1955), 474–523. MR Zbl

[Hille and Röhrle 1997] L. Hille and G. Röhrle, "On parabolic subgroups of classical groups with a finite number of orbits on the unipotent radial", *C. R. Acad. Sci. Paris Sér. I Math.* **325**:5 (1997), 465–470. MR Zbl

[Hodges 1956] J. H. Hodges, "Weighted partitions for general matrices over a finite field", *Duke Math. J.* **23** (1956), 545–552. MR Zbl

[James and Kerber 1981] G. James and A. Kerber, *The representation theory of the symmetric group*, Encycl. Math. Appl. **16**, Addison-Wesley, Reading, MA, 1981. MR Zbl

[Kashin 1990] V. V. Kashin, "Orbits of an adjoint and co-adjoint action of Borel subgroups of a semisimple algebraic group", pp. 141–158 in *Problems in group theory and homological algebra*, edited by A. L. Onishchik, Yaroslavl Gos. Univ., 1990. In Russian. MR Zbl

[Katz 1980] N. M. Katz, *Sommes exponentielles*, Astérisque **79**, Soc. Math. France, Paris, 1980. MR Zbl

[Katz 1988] N. M. Katz, *Gauss sums, Kloosterman sums, and monodromy groups*, Ann. of Math. Stud. **116**, Princeton Univ. Press, 1988. MR Zbl

[Katz and Laumon 1985] N. M. Katz and G. Laumon, "Transformation de Fourier et majoration de sommes exponentielles", *Inst. Hautes Études Sci. Publ. Math.* **62** (1985), 145–202. MR Zbl

[Kiehl and Weissauer 2001] R. Kiehl and R. Weissauer, *Weil conjectures, perverse sheaves and l'adic Fourier transform*, Ergebnisse der Math. (3) **42**, Springer, 2001. MR Zbl

[Kim 1998] D. S. Kim, "Gauss sums for symplectic groups over a finite field", *Monatsh. Math.* **126**:1 (1998), 55–71. MR Zbl

[Kloosterman 1927] H. D. Kloosterman, "On the representation of numbers in the form $ax^2 + by^2 + cz^2 + dt^2$", *Acta Math.* **49**:3-4 (1927), 407–464. MR Zbl

[Kowalski et al. 2017] E. Kowalski, P. Michel, and W. Sawin, "Bilinear forms with Kloosterman sums and applications", *Ann. of Math.* (2) **186**:2 (2017), 413–500. MR Zbl

[Landsberg 1893] G. Landsberg, "Ueber eine Anzahlbestimmung und eine damit zusammenhängende Reihe", *J. Reine Angew. Math.* **111** (1893), 87–88. MR Zbl

[Laumon 2000] G. Laumon, "Exponential sums and *l*-adic cohomology: a survey", *Israel J. Math.* **120** (2000), 225–257. MR Zbl

[Lee and Marklof 2018] M. Lee and J. Marklof, "Effective equidistribution of rational points on expanding horospheres", *Int. Math. Res. Not.* **2018**:21 (2018), 6581–6610. MR Zbl

[Linnik 1963] J. V. Linnik, "Additive problems and eigenvalues of the modular operators", pp. 270–284 in *Proceedings of the International Congress of Mathematicians* (Stockholm, 1962), edited by V. Stenström, Inst. Mittag-Leffler, Djursholm, Sweden, 1963. MR Zbl

[Luo et al. 1995] W. Luo, Z. Rudnick, and P. Sarnak, "On Selberg's eigenvalue conjecture", *Geom. Funct. Anal.* **5**:2 (1995), 387–401. MR Zbl

[Marklof 2010] J. Marklof, "The asymptotic distribution of Frobenius numbers", *Invent. Math.* **181**:1 (2010), 179–207. MR Zbl

[Milne 1980] J. S. Milne, *Étale cohomology*, Princeton Math. Ser. **33**, Princeton Univ. Press, 1980. MR Zbl

[Milne 2016] J. S. Milne, "The Riemann hypothesis over finite fields from Weil to the present day", pp. 487–565 in *The legacy of Bernhard Riemann after one hundred and fifty years*, *II*, edited by L. Ji et al., Adv. Lect. Math. **35**, Int. Press, Somerville, MA, 2016. MR Zbl

[Mordell 1963] L. J. Mordell, "On a special polynomial congruence and exponential sum", pp. 29–32 in *Calcutta Math. Soc. Golden Jubilee Commemoration*, *I* (Calcutta, 1958-1959), Calcutta Math. Soc., 1963. MR Zbl

[Morrison 2006] K. E. Morrison, "Integer sequences and matrices over finite fields", *J. Integer Seq.* **9**:2 (2006), art. id. 06.2.1. MR Zbl

[Petersson 1932] H. Petersson, "Über die Entwicklungskoeffizienten der automorphen Formen", *Acta Math.* **58**:1 (1932), 169–215. MR Zbl

[Poincaré 1911] H. Poincaré, "Fonctions modulaires et fonctions fuchsiennes", *Ann. Fac. Sci. Toulouse Sci. Math. Sci. Phys.* (3) **3** (1911), 125–149. MR Zbl

[Selberg 1965] A. Selberg, "On the estimation of Fourier coefficients of modular forms", pp. 1–15 in *Theory of numbers* (Pasadena, CA, 1963), edited by A. L. Whiteman, Proc. Sympos. Pure Math. **8**, Amer. Math. Soc., Providence, RI, 1965. MR Zbl

[SGA 4^1/$_2$ 1977] P. Deligne, "Applications de la formule des traces aux sommes trigonométriques", pp. 168–232 in *Cohomologie étale* (Séminaire de Géométrie Algébrique du Bois Marie), edited by P. Deligne, Lecture Notes in Math. **569**, Springer, 1977. MR Zbl

[Springer 1998] T. A. Springer, *Linear algebraic groups*, 2nd ed., Progr. Math. **9**, Birkhäuser, Boston, MA, 1998. MR Zbl

[Stanley 1986] R. P. Stanley, *Enumerative combinatorics, I*, Wadsworth & Brooks, Monterey, CA, 1986. MR Zbl

[Stevens 1987] G. Stevens, "Poincaré series on GL(r) and Kloostermann [sic] sums", *Math. Ann.* **277**:1 (1987), 25–51. MR Zbl

[Sylvester 1885] J. J. Sylvester, "Sur l'équation en matrices $px = xq$", *C. R. Acad. Sci. Paris* **99**:2 (1885), 67–71. Zbl JFM

[Weil 1948a] A. Weil, "On some exponential sums", *Proc. Nat. Acad. Sci. U.S.A.* **34** (1948), 204–207. MR Zbl

[Weil 1948b] A. Weil, *Sur les courbes algébriques et les variétés qui s'en déduisent*, Publ. Math. Inst. Univ. Strasbourg **7**, Hermann & Cie, Paris, 1948. MR Zbl

[Zelingher 2023] E. Zelingher, "On matrix Kloosterman sums and Hall–Littlewood polynomials", preprint, 2023. arXiv 2312.13121

merdelyi@math.bme.hu   *Department of Algebra and Geometry,*
                                         Budapest University of Technology and Economics, Budapest, Hungary

arpad.toth@ttk.elte.hu   *Department of Analysis, Eötvös Loránd University, Budapest, Hungary*

Algebra & Number Theory

msp.org/ant

Algebra & Number Theory

Volume 18   No. 12   2024