ACTA ARITHMETICA
LXXX.1 (1997)

A stationary phase formula for exponential sums over $\mathbb{Z}/p^m\mathbb{Z}$ and applications to $\mathrm{GL}(3)$-Kloosterman sums

by

Romuald Dąbrowski and Benji Fisher (New York, N.Y.)

0. Introduction. One reason for studying the classical Kloosterman sums

$$S(a,b;c)=\sum_{xy\equiv 1\ (\mathrm{mod}\, c)} e^{2\pi i(ax+by)/c} \tag{0.1}$$

is that they can be used to express the Fourier coefficients of the Poincaré series for the group $\mathrm{GL}(2,\mathbb{Q})$. As Kloosterman [Kl] pointed out, estimates of the Kloosterman sums lead to bounds for the Fourier coefficients of modular forms (see also Selberg [Se]). To estimate $S(a,b;c)$ one easily reduces to the case $c=p^m$ and $p\nmid ab$, with p prime. Salié [Sa] explicitly calculated $S(a,b;p^m)$ when $m>1$ and Weil [W1] proved that $|S(a,b;p)|\le 2\sqrt{p}$ as a consequence of his proof of the Riemann hypothesis for curves.

Thanks to Deligne [D], we now have efficient techniques for estimating exponential sums such as $S(a,b;p)$. Paradoxically, the simpler case of exponential sums over $\mathbb{Z}/p^m\mathbb{Z}$ with $m>1$ is in some ways less well understood. Smith and Loxton [Sm1, Sm2, Lo-Sm1] generalized Salié's methods and Katz [K1] interpreted such results as a stationary phase formula. We take Katz's point of view, proving and generalizing his statement in Section 1. Our statement is very convenient for applications, and in many cases it gives sharper bounds than those of Smith and Loxton. We give several examples to illustrate the use of our theorem.

The theory of Poincaré series for $\mathrm{GL}(3,\mathbb{Q})$ was developed by Bump, Friedberg, and Goldfeld [B-F-G] and extended to $\mathrm{GL}(N,\mathbb{Q})$ independently by Friedberg [F] and Stevens [S]. The Fourier coefficients of these Poincaré

1991 *Mathematics Subject Classification*: Primary 11L05; Secondary 11F30, 11G25, 11L07, 11T23.

Key words and phrases: exponential sums, Fourier coefficients of automorphic forms.

The second author was partially supported by NSF grant DMS-92-04738.

series can be expressed in terms of certain exponential sums, which are therefore called $\mathrm{GL}(N,\mathbb{Q})$-Kloosterman sums. Following [S], we will denote these sums $\mathrm{Kl}(wt,\psi,\psi')$, where w is in the Weyl group of G, t is a diagonal matrix, and ψ, ψ' are characters of the group $U(\mathbb{Q})$ of unipotent upper triangular matrices, trivial on $U(\mathbb{Z})$. The $\mathrm{GL}(N,\mathbb{Q})$-Kloosterman sum is a product of *local* $\mathrm{GL}(N,\mathbb{Q}_p)$-Kloosterman sums $\mathrm{Kl}_p(wt,\psi,\psi')$; we will usually omit the subscript p.

Fix $N=3$ and let w_0 be the long element of the Weyl group. For $w\neq w_0$, sharp bounds for $\mathrm{Kl}(wt,\psi,\psi')$ are given in [B-F-G] and [L]. For $w=w_0$, the bound

$$\begin{gathered}|\mathrm{Kl}(w_0t,\psi,\psi')|\le C_{\psi,\psi'}(r+1)(s+1)p^{(r+s+\min\{r,s\})/2},\\ t=\begin{pmatrix}p^s&&\\&p^{r-s}&\\&&p^{-r}\end{pmatrix},\end{gathered}\tag{0.2}$$

for the local Kloosterman sum is given in [S, Theorem 5.1]. In Section 2 we find a fairly explicit expression for these long-element Kloosterman sums and in Section 3 we improve the bound (0.2).

A more detailed description of our results follows.

Stationary phase. Following Katz [K1], we describe our results in the language of schemes. While we have tried to present the material in a way that will be comprehensible even to those unfamiliar with this language, we fear that the language (and the level of generality) may obscure the fact that we have made one or two substantial improvements over previous results. We will therefore consider first the simplest case in which our improvements come into play.

Let f be a polynomial with coefficients in $\mathbb{Z}_p$ (or $\mathbb{Z}$ or $\mathbb{Z}/p^m\mathbb{Z}$) and consider the exponential sum

$$S_m(f)=\sum_{x=1}^{p^m}e^{2\pi i f(x)/p^m}=\sum_{x\in\mathbb{Z}/p^m\mathbb{Z}}e^{2\pi i f(x)/p^m}.$$

The basic idea, which goes back at least to Salié [Sa], is to use the Taylor expansion

$$f(x+p^{m-j}y)=f(x)+p^{m-j}f'(x)y+\tfrac{1}{2}p^{2(m-j)}f''(x)y^2+\dots$$

If $2(m-j)\ge m$ (and p is odd) it follows that

$$\begin{aligned}S_m(f)&=\frac{1}{p^j}\sum_{x\in\mathbb{Z}/p^m\mathbb{Z}}\ \sum_{y\in\mathbb{Z}/p^j\mathbb{Z}}e^{2\pi i f(x+p^{m-j}y)/p^m}\\&=\sum_{x\in\mathbb{Z}/p^m\mathbb{Z}}e^{2\pi i f(x)/p^m}\cdot\frac{1}{p^j}\sum_{y\in\mathbb{Z}/p^j\mathbb{Z}}e^{2\pi i f'(x)y/p^j}.\end{aligned}$$

The inner sum vanishes unless $f'(x) \equiv 0 \pmod{p^j}$, leading to

$$S_m(f) = \sum_{x \in \mathbb{Z}/p^m\mathbb{Z},\, f'(x) \equiv 0 \,(\mathrm{mod}\, p^j)} e^{2\pi i f(x)/p^m},$$

which we interpret as a sum over the *approximate critical points of* f.

In the simplest case, $f''(x)$ is a unit for every approximate critical point x of f and there are one-to-one correspondences $D(\mathbb{F}_p) \xrightarrow{\sim} D(\mathbb{Z}/p^j\mathbb{Z}) \xrightarrow{\sim} D(\mathbb{Z}_p)$, where we let $D(A) := \{x \in A : f'(x) = 0\}$ be the set of critical points in A. In particular, the number of critical points is at most the degree of f'. Following Katz, we focus on the p-adic critical points and rewrite the above equation as

$$S_m(f) = \sum_{x \in D(\mathbb{Z}_p)} \sum_{y \in \mathbb{Z}/p^{m-j}\mathbb{Z}} e^{2\pi i f(x+p^j y)/p^m}.$$

Taking one more term in the Taylor expansion, one identifies the inner sum as a power of p times a Gauss sum times $e^{2\pi i f(x)/p^m}$, the value of the exponential at the exact (p-adic) critical point.

Our main new idea is what to do when $f''(x)$ is not a unit for some approximate critical point x. Katz does not deal with this point; Smith [Sm2] and Loxton–Smith [Lo-Sm1] introduce some new ideas to estimate the number of approximate critical points; and for each such point they estimate the local term (a Gauss sum). We assume that j is sufficiently large, then apply Hensel's Lemma to lift the approximate critical points to exact ones. More precisely, assume that $f''(x) = p^h(\text{unit})$, where $x \in D(\mathbb{Z}/p^j\mathbb{Z})$ and $j \geq 2h + 1$. Then there is a unique exact critical point $x_0 \in D(\mathbb{Z}_p)$ such that $x \equiv x_0 \pmod{p^{j-h}}$. Now we group together all the terms coming from $x' \in D(\mathbb{Z}/p^j\mathbb{Z})$ that correspond to the same $x_0 \in D(\mathbb{Z}_p)$, to get one local term for x_0. This allows for further cancellation; since our local term is still a Gauss sum, we are able to realize this possibility. This is why we get better bounds, when j is sufficiently large, than those of Loxton–Smith. Our main result is unfortunately complicated since we need to allow the possibility of a different value of h for each approximate critical point (and it is certainly *not* sufficient to consider the value of $f''(x_0)$ for the exact critical points) but the examples show that this is rarely a problem.

When everything is worked through, we find that (with notation as above, and still assuming $p \neq 2$) our method works if $m \geq 3h + 2$. In order to get this same result when dealing with sums in several variables, we need a slight improvement (our Lemma 1.20) on the usual n-dimensional version of Hensel's Lemma (e.g., the one in Bourbaki [B, Chapter III, § 4.5, Theorem 2]): basically, looking at the Jacobian determinant is too sloppy. Although we only state this lemma for $\mathbb{Z}_p$, it clearly holds more generally ([1]).

[1] One of us has worked out a more general version in [Fi].

Finally, let us say where we still fall short of previous results. Loxton and Smith have reasonable results for all values of m, whereas our method works only for m sufficiently large. They also have results [Lo-Sm2] (only for one-variable sums, as considered above) when there is a multiple root of the derivative: that is, $f''(x_0) = 0$ for some exact critical point $x_0 \in D(\mathbb{Z}_p)$.

Let us now state our results in more generality.

Let V be a smooth, n-dimensional variety over $\mathbb{Z}_p$ and f a regular function on V. If the Hessian determinant of f is a unit at every critical point of $f \pmod p$ then, for all $m > 1$,

$$S = S_m(V, f) = \sum_{x \in V(\mathbb{Z}/p^m\mathbb{Z})} e^{2\pi i f(x)/p^m} \tag{0.3}$$

can be expressed as a sum, over the critical points x of f, of $p^{nm/2} e^{2\pi i f(x)/p^m}$ times a root of unity; this is the statement in [K1]. This is closely analogous to the classical stationary phase formula for estimating oscillatory integrals: we can think of S as p^{nm} times $\int_{V(\mathbb{Z}_p)} e^{2\pi i t f(x)}\,dx$ with $t = 1/p^m$; then $m > 1$ means that t is large (p-adically).

We have generalized Katz's statement by weakening the hypothesis that the Hessian determinant of f be a unit: we assume only that it is non-zero at every (approximate) critical point of f. (In fancy language, Katz assumes that the locus of critical points of f is étale over $\mathbb{Z}_p$; we assume that it is étale over $\mathbb{Q}_p$.) In [Lo-Sm1] the hypotheses are similar to ours, but only the case of affine space ($V = \mathbb{A}^n$) is considered. If the Hessian determinant is not a unit then [Lo-Sm1] gives bounds on $|S|$ for all $m > 1$; our result applies only for m sufficiently large, but then it leads to sharper bounds. Our main result is

THEOREM 0.1. *Let S be defined by* (0.3). *For sufficiently large m,*

$$S = p^{nm/2} \sum_{x \in D(\mathbb{Z}_p)} e^{2\pi i f(x)/p^m} G_m(H_x),$$

where D is the scheme of critical points of f, H_x is the Hessian matrix of f at x, and $G_m(H_x)$ is the normalized Gauss sum defined in Definition 1.2.

The hypotheses are stated precisely in Theorem 1.8.

By way of example, we apply our results to Gauss sums, Kloosterman sums (recovering the results of [Sa]), and the n-variable Kloosterman sums considered in [Sm1] and (later) in [L] (for $n = 3$) and [F].

GL(3)-*Kloosterman sums.* In Section 2 we evaluate the GL(3)-Kloosterman sum $\mathrm{Kl}(w_0 t, \psi, \psi')$, for the long element

$$w_0 = \begin{pmatrix} & & 1 \\ & -1 & \\ 1 & & \end{pmatrix}$$

of the Weyl group,

$$t = \begin{pmatrix} p^s & & \\ & p^{r-s} & \\ & & p^{-r} \end{pmatrix}, \qquad \psi\begin{pmatrix} 1 & x & z \\ & 1 & y \\ & & 1 \end{pmatrix} = e^{2\pi i(\nu_1 x + \nu_2 y)},$$

and

$$\psi'\begin{pmatrix} 1 & x & z \\ & 1 & y \\ & & 1 \end{pmatrix} = e^{2\pi i(\nu_1' x + \nu_2' y)}.$$

For simplicity, we will assume here that ν_1, ν_2, ν_1', and ν_2' are units in $\mathbb{Z}_p$. Our first result, Theorem 2.4, is a slightly more explicit formula than what is given in [S]. Our formula involves classical Kloosterman sums, as in (0.1), and sums of products of Kloosterman sums, similar to (0.4) below.

The Kloosterman sum $\mathrm{Kl}(w_0 t, \psi, \psi')$ is defined as the sum of $\psi(u)\psi'(u')$ over pairs $u \in U(\mathbb{Z}_p)\backslash U(\mathbb{Q}_p)$, $u' \in U(\mathbb{Q}_p)/U(\mathbb{Z}_p)$ such that

$$uw_0tu' \in X(w_0t) = U(\mathbb{Z}_p)\backslash U(\mathbb{Q}_p)w_0tU(\mathbb{Q}_p) \cap \mathrm{GL}(3,\mathbb{Z}_p)/U(\mathbb{Z}_p).$$

In order to calculate $\mathrm{Kl}(w_0t, \psi, \psi')$, one first breaks up $X(w_0t)$ into smooth strata. We use the same stratification as Stevens, but we associate each stratum with one of the cells of the Iwahori decomposition of $\mathrm{GL}(3,\mathbb{Z}_p)$. We hope that this approach will be helpful in the case of other reductive groups.

The rest of Section 2 is an elaborate bookkeeping exercise (one that would be greatly simplified if we assumed in Section 2, as we do here, that ν_1, ν_2, ν_1', and ν_2' are units). We express our results in terms of the sum of products

$$P(\gamma;\mathbb{Z}/p^r\mathbb{Z}) = \sum_{\substack{x\in(\mathbb{Z}/p^r\mathbb{Z})^\times \\ p\nmid(ax+b)(cx+d)}} S(1,x;p^r)S(1,\gamma(x);p^s), \tag{0.4}$$

where $\gamma(x) = (ax+b)/(cx+d)$ is a linear fractional transformation with $a,b,c,d \in \mathbb{Z}_p$ and $v_p(ad-bc) = s-r$. A simplified version of Theorem 2.11 is

Theorem 0.2. *Assume that $s \geq r \geq 2$ and let*

$$\gamma_m = \begin{pmatrix} \nu_2\nu_1' & 0 \\ p^{r-2m} & -p^{s-r}\nu_1\nu_2' \end{pmatrix}.$$

If $r = s$ then

$$\mathrm{Kl}(w_0t,\psi,\psi') = p^r\left[\frac{1}{p} + 1 + \sum_{1\leq m\leq r/2} p^{-m}P(\gamma_m;\mathbb{Z}/p^m\mathbb{Z})\right].$$

If $r < s$ then the Kloosterman sum vanishes if r is odd; *if r is even then*

$$\mathrm{Kl}(w_0t,\psi,\psi') = p^{r/2}P(\gamma_{r/2};\mathbb{Z}/p^{r/2}\mathbb{Z}).$$

In Section 3 we analyze the sums $P(\gamma;\mathbb{Z}/p^m\mathbb{Z})$. Using the stationary phase results of Section 1 and the l-adic techniques developed by Deligne and Katz [D, K2, K3], we estimate these sums in most cases. (We do not deal with $\gamma=\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)$ if $v_p(b/3)=v_p(c)<m$.) Our results show that almost all of the terms of the sum in Theorem 0.2 telescope or vanish. A simplified version of Theorem 3.7 is

THEOREM 0.3. *Assume that $p>3$ and $s\ge r\ge 2$; let γ_m be as in Theorem 0.2 and let $\varepsilon=v_p(\nu_1\nu_2'+\nu_2\nu_1')$. If $r=s$ then*

$$\mathrm{Kl}(w_0t,\psi,\psi')=p^r\Big[\sum_{m\in M}p^{-m}P(\gamma_m;\mathbb{Z}/p^m\mathbb{Z})+T\Big],$$

where

$$M=\left\{\frac{r+1}{3},\frac{r-\varepsilon}{2},\frac{r}{2}\right\}\cap\mathbb{Z}\cap\left[\frac{r+1}{3},\frac{r}{2}\right],\qquad T=\begin{cases}0 & \text{if } r>3\varepsilon+2,\\ p^{\lfloor r/3\rfloor} & \text{if } r\le 3\varepsilon+2,\end{cases}$$

and $\lfloor x\rfloor$ denotes the greatest integer in x. As $r\to\infty$, $|\mathrm{Kl}(w_0t,\psi,\psi')|=O(p^{5r/4})$ unless $\nu_1\nu_2'+\nu_2\nu_1'=0$ (or $\varepsilon=\infty$), in which case $|\mathrm{Kl}(w_0t,\psi,\psi')|=O(p^{4r/3})$. If $r<s$ then $|\mathrm{Kl}(w_0t,\psi,\psi')|\le 6p^{(3r+2s)/4}$ if r is even; if r is odd then the Kloosterman sum vanishes.

As promised, this represents an improvement over (0.2). There is little room left for cancellation, so our bounds should be sharp (with the exception of the constant $O(1)$ in Theorem 3.7 when $p=2$ or 3).

Open problems. Our work suggests the following problems; the third seems fairly manageable.

1. Globalize the explicit formulae for the $\mathrm{GL}(3,\mathbb{Q}_p)$-Kloosterman sums to obtain formulae for $\mathrm{GL}(3,\mathbb{Q})$-Kloosterman sums. Stevens notes in [S] that improved estimates for $\mathrm{Kl}(w_0t,\psi,\psi')$ will not yield a larger region of convergence of the Kloosterman zeta function. It is possible, however, that our fairly explicit formulae will be useful in the study of the zeta function.

2. Describe a smooth stratification of Kloosterman sets in the case of $\mathrm{GL}(N,\mathbb{Q}_p)$, $N>3$ (more generally, in the case of an arbitrary algebraic reductive group over a local field). We hope that a refinement of our method of breaking up the Kloosterman sets according to the Iwahori decomposition will yield such a stratification ([2]).

3. Extend Deligne's theory of exponential sums over $\mathbb{F}_p$ to handle sums over $\mathbb{Z}/p^m\mathbb{Z}$ by using Witt vectors to replace n-dimensional varieties over $\mathbb{Z}/p^m\mathbb{Z}$ with nm-dimensional varieties over $\mathbb{F}_p$. Prove a stationary phase theorem in this context. This should lead to a uniform method for estimating the sums $P(\gamma;\mathbb{Z}/p^m\mathbb{Z})$; in this paper, we use different methods, depending on γ and m.

[2] Some work along these lines has already been completed: see [D-R].

4. Use the ideas described above to remove the hypothesis that the scheme of critical points be generically étale, leading to a generalization of the work of Smith, Loxton, and Vaughan [Lo-Sm2, Lo-V] on one-variable sums (the case $V = \mathbb{A}^1$).

1. Stationary phase method for p-adic integrals. In this section, we discuss a p-adic analogue of the classical stationary phase method (see, e.g., [H, Section 7.7]) for finding asymptotics of integrals of the form $\int \phi(x)e^{2\pi i t f(x)}dx$ as $t \to \infty$. This analogue turns out to be very handy for estimating exponential sums over $\mathbb{Z}/p^m\mathbb{Z}$ when $m > 1$. (When $m = 1$, one uses Deligne's theory [D].) We have tried to present this material in a way that will be easy to use and we give several explicit examples.

Notation 1.1. We will use the following notation throughout this section: p is a prime, v_p is the valuation on the field $\mathbb{Q}_p$ of p-adic numbers, V is a smooth scheme of dimension $n \geq 1$ over $\mathbb{Z}_p$, $f : V \to \mathbb{A}^1 = \mathbb{A}^1_{\mathbb{Z}_p}$ is a $\mathbb{Z}_p$-morphism, and $D \subseteq V$ is the scheme of critical points of f. (Since we are familiar with it, we use the language of schemes. It should not be hard to translate into other languages—see the Explicitation subsection.) Let $H_x = H_{x,f}$ denote the Hessian matrix of f at x (cf. the Explicitation subsection) and let $H_x(z)$ denote the quadratic form $H_x(z) = {}^tzH_xz$. We let m be an integer greater than 1 and let

$$S = S_{m,V,f} = \sum_{x\in V(\mathbb{Z}/p^m\mathbb{Z})} e^{2\pi i f(x)/p^m}, \tag{1.1}$$

so that $\int_{V(\mathbb{Z}_p)} e^{2\pi i f(x)/p^m}\,dx$ can be interpreted as $p^{-nm}S$.

Statements. Before stating any version of the stationary phase formula, we will discuss the Gauss sums that occur. For the usefulness of our normalization, see both Proposition 1.3 below and (for the case $n = 1$) Example 1.13. We will use the Gauss sum $G_h(A; v)$ only when $v = 0$ or $h = 1$.

Definition 1.2. Let A be a symmetric, $n \times n$ matrix with entries in $\mathbb{Z}_p$ and let $v \in \mathbb{Z}_p^n$. For $h \geq 1$, we define the normalized, n-dimensional *Gauss sum* associated with A and v to be

$$\begin{aligned} G_h(A; v) &= p^{-nh/2} \sum_{x\in(\mathbb{Z}/p^h\mathbb{Z})^n} e^{\pi i({}^txAx)/p^h} e^{2\pi i\, v\cdot x/p^h}; \\ G_h(A) &= G_h(A; 0); \end{aligned} \tag{1.2}$$

with the convention that $\pi i({}^txAx)/p^h$ means $2\pi i({}^tx \cdot \frac{1}{2}A \cdot x)/p^h$ if p is odd; and if $p = 2$ then it means $2\pi i({}^txAx)/p^{h+1}$—note that, since A is symmetric, ${}^txAx$ makes sense as an element of $\mathbb{Z}/2^{h+1}\mathbb{Z}$ if $x \in \mathbb{Z}/2^h\mathbb{Z}$.

Proposition 1.3. *Let A be as above.*

(a) *Assume that* $\det A \neq 0$ *and that h is large enough that $A' = p^h A^{-1}$ has entries in $\mathbb{Z}_p$. If $p = 2$, also assume that the diagonal entries of A' are even. Then $G_h(A)$ is $p^{v_p(\det A)/2}$ times a fourth root of unity* (*unless $p = 2$, in which case it may be an eighth root of unity*). *This root of unity depends only on whether h is even or odd*; *that is*, $G_h(A) = G_{h+2}(A)$.

(b) *If $v = Au$ then $G_h(A; v) = e^{-\pi i\, {}^t uAu/p^h} G_h(A)$. If p is odd then $G_h(A, v) \neq 0$ if and only if $v = Au$ for some u.*

(c) *Let r denote the rank of A, thought of as a linear transformation on $\mathbb{F}_p^n$. Either $G_1(A, v) = 0$ or it is $p^{(n-r)/2}$ times a root of unity as in* (a).

Proof. (a) Think of $(\mathbb{Z}/p^h\mathbb{Z})^n$ as $\mathbb{Z}_p^n/p^h\mathbb{Z}_p^n$ and note that, for any $y \in A'\mathbb{Z}_p^n$, we have ${}^t(x+y)A(x+y) \equiv {}^t xAx \pmod{2p^h}$. Since $A'\mathbb{Z}_p^n \supseteq A'A\mathbb{Z}_p^n = p^h\mathbb{Z}_p^n$ with index $p^{v_p(\det A)}$, we find

$$p^{nh/2} G_h(A) = \sum_{x \in \mathbb{Z}_p^n/p^h\mathbb{Z}_p^n} e^{\pi i({}^t xAx)/p^h} = p^{v_p(\det A)} \sum_{x \in \mathbb{Z}_p^n/A'\mathbb{Z}_p^n} e^{\pi i {}^t x(p^{-h}A)x}.$$

Note that $\mathbb{Z}_p^n$ and $A'\mathbb{Z}_p^n$ are duals with respect to the inner product on $\mathbb{Q}_p^n$ defined by $\langle x, y\rangle = {}^t x(p^{-h}A)y$, that $\langle x, x\rangle$ is even for all $x \in A'\mathbb{Z}_p^n$, and that $|\mathbb{Z}_p^n/A'\mathbb{Z}_p^n| = p^{v_p(\det A')} = p^{nh - v_p(\det A)}$. Thus the last sum above is $p^{v_p(\det A')/2}$ times Weil's invariant $\gamma_p(p^{-h}A)$ of the form $\langle\,,\rangle$. (Cf. [W2] or [M-H, Appendix 4], for example.) Checking the powers of p, one finds that

$$G_h(A) = p^{v_p(\det A)/2}\gamma_p(p^{-h}A). \tag{1.3}$$

Since γ_p gives a homomorphism from the Witt group $W(\mathbb{Q}_p)$ to $\mathbb{C}^\times$ and $|W(\mathbb{Q}_p)| = 4$ (except that $|W(\mathbb{Q}_2)| = 8$), $\gamma_p(p^{-h}A)$ is a root of unity, as stated. For any $P \in \mathrm{GL}_n(\mathbb{Q}_p)$, $\gamma_p(A) = \gamma_p({}^t PAP)$; thus $\gamma_p(A) = \gamma_p(p^2 A)$ (taking $P = pI_n$) and so $\gamma_p(p^{-h}A)$ depends only on the parity of h.

(b) Replacing x by $x + y$ in (1.2), one obtains

$$G_h(A; v) = e^{\pi i {}^t yAy/p^h} e^{2\pi i\, v\cdot y/p^h} G_h(A; v + Ay).$$

If $v = Au$ then, taking $y = -u$, we get the desired formula. Now assume that p is odd. If $G_h(A; v) \neq 0$ and $Ay = 0 \in (\mathbb{Z}/p^h\mathbb{Z})^n$ then $v \cdot y = 0 \in (\mathbb{Z}/p^h\mathbb{Z})^n$. Now the trick is to diagonalize A as a linear transformation: by the theory of elementary divisors, we can find invertible matrices P and Q with entries in $\mathbb{Z}_p$ such that $P^{-1}AQ$ is diagonal. If $P^{-1}AQy = 0$ then $AQy = 0$, so ${}^t vQy = v \cdot Qy = 0$; it follows that ${}^t vQ = {}^t u_0 P^{-1}AQ$, whence $v = Au$ with $u = {}^t P^{-1}u_0$.

(c) We will defer this proof until after Remark 1.14. (There is no circularity: Example 1.13 and Remark 1.14 rely on Theorem 1.8(b), which uses only part (a) of this proposition.) Unfortunately, we cannot simply quote [Lo-Sm1] because their result is incorrect when $p = 2$. ■

Note that the condition that $p^h A^{-1}$ have entries in $\mathbb{Z}_p$ is equivalent to each of the following: there is an A' with entries in $\mathbb{Z}_p$ such that $A'A = p^h I_n$; $A\mathbb{Z}_p^n \supseteq p^h\mathbb{Z}_p^n$; p^h kills $\mathbb{Z}_p^n/A\mathbb{Z}_p^n$. These conditions all make sense if we replace $\mathbb{Z}_p$ with $\mathbb{Z}/p^N\mathbb{Z}$ with $N > h$. Furthermore, $h > v_p(\det A)$ is good enough, by Cramer's rule. For a direct calculation of the Gauss sum $G_h(A)$, see Remark 1.14.

In [K1, p. 110], Katz states the following theorem, without proof:

THEOREM 1.4. *Suppose that f is a "Morse function": that is, assume that the scheme D of critical points of f in V is finite étale over $\mathbb{Z}_p$. Then $S = 0$ if $D(\mathbb{Z}_p)$ is empty. In general,*

$$S = p^{nm/2} \sum_{x\in D(\mathbb{Z}_p)} e^{2\pi i f(x)/p^m} G_m(H_x) \tag{1.4}$$

and

$$G_m(H_x) = \begin{cases} 1 & \text{if } m \text{ is even},\\ G_1(H_x) = \dfrac{1}{p^{n/2}} \displaystyle\sum_{z\in(\mathbb{F}_p)^n} e^{\pi i H_x(z)/p} & \text{if } m \text{ is odd}. \end{cases} \tag{1.5}$$

Remarks 1.5. (1) Since $G_m(H_x)$ is a root of 1, $|S| \le |D(\mathbb{Z}_p)|p^{nm/2}$.

(2) If D is not étale then it is still closed in V. We can apply the first part of the theorem to $V' = V - D$ and conclude that the sum over $V(\mathbb{Z}/p^m\mathbb{Z})$ is the same as the sum over $V(\mathbb{Z}/p^m\mathbb{Z}) \setminus V'(\mathbb{Z}/p^m\mathbb{Z})$. This is not the same as the sum over $D(\mathbb{Z}/p^m\mathbb{Z})$. Consider, for example, $V = \mathbb{G}_m = \operatorname{Spec}\mathbb{Z}_p[t, t^{-1}]$ and $f(t) = t + t^{-1}$. Then $D = \operatorname{Spec}\mathbb{Z}_p[t, t^{-1}]/(1 - t^{-2})$ (which is not étale if $p = 2$) and $V' = \operatorname{Spec}\mathbb{Z}_p[t, 1/t(t^2 - 1)]$. Therefore $V(\mathbb{Z}/p^m\mathbb{Z}) = (\mathbb{Z}/p^m\mathbb{Z})^\times$ and $V'(\mathbb{Z}/p^m\mathbb{Z}) = \{x \in \mathbb{Z}/p^m\mathbb{Z} : x \not\equiv 0, 1, -1 \pmod p\}$; and $D(\mathbb{Z}/p^m\mathbb{Z}) = \{1, -1\}$ if p is odd, $D(\mathbb{Z}/p^m\mathbb{Z}) = \{1, 1 + p^{m-1}, -1, -1 + p^{m-1}\}$ if $p = 2$.

(3) Katz interprets S as p^{nm} times the integral $\int_{V(\mathbb{Z}_p)} e^{2\pi i f(x)/p^m}\,dx$, but our interest is in the sum itself. Katz states the theorem for a slightly more general integrand: $e^{2\pi i t f(x)}$, with $t \in \mathbb{Q}_p$ and $v_p(t) = -m$. Since we are not interested in the variation with t, we absorb it into the function f: $tf(x) = (p^m t)f(x)/p^m$ and, since $p^m t \in \mathbb{Z}_p^\times$, the new function $(p^m t)f$ is still defined over $\mathbb{Z}_p$.

(4) We will prove the finer version of stationary phase given below. (The case $j = 1$ of Theorem 1.8(a) follows from Katz's version, cf. (2) above, as does the case $h = 0$, $k = 1$ of Theorem 1.8(b).) Note that Corollary 1.10 can be interpreted as a stationary phase formula for $\int_{V(\mathbb{Z}_p)} \phi(x)e^{2\pi i f(x)/p^m}\,dx$, where $\phi : V(\mathbb{Z}_p) \to \mathbb{C}$ is any locally constant function. (Presumably, the

stationary phase formula still holds for $\int_{V(\mathbb{Q}_p)} \phi(x)e^{2\pi i f(x)/p^m}\,dx$, where $\phi : V(\mathbb{Q}_p) \to \mathbb{C}$ is locally constant and compactly supported.)

Definition 1.6. Let $x \in D(\mathbb{Z}/p^k\mathbb{Z})$. We will say that x is an *h-étale critical point* of f if $h < k$ and H_x divides $p^h I_n$; that is, if there is a matrix H' (with entries in $\mathbb{Z}/p^k\mathbb{Z}$ or in $\mathbb{Z}_p$) such that $H'H_x = p^h I_n$. If $p = 2$ and one can take H' with even diagonal entries then we will say that x is *strictly h-étale.*

Remarks 1.7. (1) If x is a 0-étale critical point of f then the Hessian matrix is invertible at x, and so x is an étale point of the scheme D of critical points of f. (Cf. the Explicitation subsection, below.) Thus h-étale is a weakening of étale.

(2) By Cramer's rule, H_x divides $p^v I_n$ with $v = v_p(\det H_x)$.

(3) Assume that $x \in D(\mathbb{Z}/p^k\mathbb{Z})$ is h-étale. Then x is also h'-étale if $h \le h' < k$; and if $k < k'$ and $y \in D(\mathbb{Z}/p^{k'}\mathbb{Z})$ reduces to x then y is h-étale. If $p = 2$ and $h + 1 < k$ then x is strictly $(h + 1)$-étale.

Theorem 1.8. *Let m and j be positive integers, with $j \le m$. Let S be as in* (1.1) *and, for $\overline{x} \in V(\mathbb{Z}/p^j\mathbb{Z})$, let $S_{\overline{x}}$ represents the sum over all $x \in V(\mathbb{Z}/p^m\mathbb{Z})$ that reduce to $\overline{x}$, so that $S = \sum_{\overline{x}} S_{\overline{x}}$.*

(a) *If $2j \le m$ then $S_{\overline{x}} = 0$ unless $\overline{x} \in D(\mathbb{Z}/p^j\mathbb{Z})$. Now let $m = 2j$ or $2j + 1$ and let $x \in V(\mathbb{Z}/p^m\mathbb{Z})$ map to $\overline{x} \in D(\mathbb{Z}/p^j\mathbb{Z})$. If $m = 2j$ then*

$$S_{\overline{x}} = p^{nm/2} e^{2\pi i f(x)/p^m}.$$

If $m = 2j + 1$ then

$$S_{\overline{x}} = p^{nm/2} e^{2\pi i f(x)/p^m} G_1(H_x, p^{-j}\operatorname{grad} f(x)).$$

In particular, if we let s denote the maximum value of $n - \operatorname{rank}_{\mathbb{F}_p} H_{\overline{x}}$ for $\overline{x} \in D(\mathbb{Z}/p^j\mathbb{Z})$ then $|S| \le |D(\mathbb{Z}/p^j\mathbb{Z})| p^{(nm+s)/2}$.

(b) *Assume there are positive integers h and k, with $h < k$, such that every $\overline{x} \in D(\mathbb{Z}/p^k\mathbb{Z})$ is an h-étale critical point of f; if $p = 2$ and $m - h$ is even then also assume that all such $\overline{x}$ are strictly h-étale. If $m \ge 3h + 2$ and $m \ge 2k$ then*

$$S = \sum_{x \in D(\mathbb{Z}_p)} S_{\overline{x}}, \qquad S_{\overline{x}} = p^{nm/2} e^{2\pi i f(x)/p^m} G_m(H_x), \tag{1.6}$$

where $\overline{x}$ denotes the image of $x \in D(\mathbb{Z}_p)$ in $D(\mathbb{Z}/p^j\mathbb{Z})$, with $j = \lfloor (m-h)/2 \rfloor$. In particular, if we let s denote the maximum value of $v_p(\det H_x)$ for $x \in D(\mathbb{Z}_p)$ (so that $s \le nh$) then $|S| \le |D(\mathbb{Z}_p)| p^{(nm+s)/2}$.

Remarks 1.9. (1) Examples 1.15 and 1.16 show that the bounds on m are sometimes necessary, at least when $h = 1$.

(2) It seems to us that most of the power of the stationary phase method is in Theorem 1.8(a) (which follows from the fact that the sum of a nontrivial character over a finite group vanishes). For example, it leads to the bound $|S| \leq |D(\mathbb{Z}/p^j\mathbb{Z})| p^{n(m-j)}$. If D is étale then this is the "right" bound when $m = 2j$ and it is close when $m = 2j+1$.

(3) The main disadvantage of Theorem 1.8(a) is that it is hard to estimate the number of points in $D(\mathbb{Z}/p^j\mathbb{Z})$; this is done (in the case $V = \mathbb{A}^n$) in [Lo-Sm1]. There may also be cancellation among the terms $S_{\overline{x}}$ for $\overline{x} \in D(\mathbb{Z}/p^j\mathbb{Z})$; this is why part (b) leads to sharper bounds. In some cases, such as Example 1.17, there is enough control over the critical points to get reasonably good bounds from part (a).

Corollary 1.10. *Keep the notations of Theorem* 1.8(b) *and let* $\phi : V(\mathbb{Z}/p^j\mathbb{Z}) \to \mathbb{C}$ *be any function*; *also let* ϕ *denote the compositions* $V(\mathbb{Z}/p^m\mathbb{Z}) \to V(\mathbb{Z}/p^j\mathbb{Z}) \to \mathbb{C}$ *and* $D(\mathbb{Z}_p) \to V(\mathbb{Z}/p^j\mathbb{Z}) \to \mathbb{C}$ (*by abuse of notation*). *Let*

$$S(\phi) := \sum_{x \in V(\mathbb{Z}/p^m\mathbb{Z})} \phi(x) e^{2\pi i f(x)/p^m}.$$

Then

$$S(\phi) = p^{nm/2} \sum_{x \in D(\mathbb{Z}_p)} \phi(x) e^{2\pi i f(x)/p^m} G_m(H_x).$$

Explicitation. First, let us reassure those who are unfamiliar with the language of schemes that the notation $V(\mathbb{Z}_p)$, where $V \subseteq \mathbb{A}^n$ is the scheme defined by equations $f_i = 0$, denotes the set of solutions $x \in \mathbb{Z}_p^n$ to $f_i(x) = 0$. Similarly for $V(\mathbb{Z}/p^m\mathbb{Z})$ (or $V(R)$, where R is any $\mathbb{Z}_p$-algebra).

So far, we have been vague about the definition of D, simply referring to it as "the scheme of critical points of f". (Katz refers to D as "the subscheme … of V defined by the vanishing of grad(f)". We avoid this description because of Example 1.12.) Now we will be more precise.

First, recall the Jacobian criterion for smoothness. (Some standard references are [M, Section III.10], [D-G, Section I.4.4], and [SGA].) A scheme V over $\mathbb{Z}_p$ is smooth if, locally, $V = \operatorname{Spec} A$ with $A = \mathbb{Z}_p[t_1, \ldots, t_N]/(g_1, \ldots, g_r)$ and the $r \times r$ minors of $\partial(g_1, \ldots, g_r)/\partial(t_1, \ldots, t_N)$ generate the unit ideal in A. Equivalently, $N = n + r$ (where n is the dimension of V over $\mathbb{Z}_p$) and the Jacobian matrix has rank r at every point of V. In particular, $V/\mathbb{Z}_p$ is étale (i.e., smooth of dimension 0) if and only if it is locally of the form $V = \operatorname{Spec} A$, where $A = \mathbb{Z}_p[t_1, \ldots, t_N]/(g_1, \ldots, g_N)$ and the Jacobian matrix $\partial(g_1, \ldots, g_N)/\partial(t_1, \ldots, t_N)$ is invertible.

The simplest case is when V is affine space $\mathbb{A}^n = \mathbb{A}^n_{\mathbb{Z}_p}$ (or an open affine subscheme of $\mathbb{A}^n$). Then the $\mathbb{Z}_p$-morphism $f : V \to \mathbb{A}^1$ is simply a polynomial (or a rational function with denominator a unit on V). The gradient of f is the n-tuple $\operatorname{grad} f = (\partial f/\partial t_1, \dots, \partial f/\partial t_n)$ of polynomials (or rational functions) and $D \subseteq V$ is the closed subscheme defined by $\partial f/\partial t_1, \dots, \partial f/\partial t_n$. The Hessian matrix of f is $H = (\partial^2 f/\partial t_i \partial t_j)$, which is also the Jacobian matrix of $\operatorname{grad} f$. For a $\mathbb{Z}_p$-valued point $x \in V(\mathbb{Z}_p)$, the Hessian of f at x is $H_x = (\partial^2 f/\partial t_i \partial t_j|_x)$, a matrix with entries in $\mathbb{Z}_p$. By the Jacobian criterion, D is étale at x if and only if $x \in D$ and H_x is invertible. (Equivalently, $\det H_x \in \mathbb{Z}_p^\times$; or H is invertible as a matrix with entries in the local ring $\mathcal{O}_{V,x}$.)

In practice, V is often an affine variety; in general, this is true locally. So suppose that $V = \operatorname{Spec} A$ with $A = \mathbb{Z}_p[t_1, \dots, t_N]/(g_1, \dots, g_r)$; we can use the Jacobian criterion to check that V is smooth. The $\mathbb{Z}_p$-morphism $f : V \to \mathbb{A}^1$ can be thought of as a polynomial in $t_1, \dots, t_N$. We want D to be the scheme of "singular points of the level sets of f", so we define D by the condition that $f, g_1, \dots, g_r$ do *not* define a smooth scheme:

$$D := \operatorname{Spec} A/I,$$

where I is the ideal generated by the $(r+1) \times (r+1)$ minors of the Hessian $\partial(f, g_1, \dots, g_r)/\partial(t_1, \dots, t_N)$. That is, $\operatorname{grad} f$ should be a linear combination of $\operatorname{grad} g_1, \operatorname{grad} g_2, \dots, \operatorname{grad} g_r$ at every point of D. According to the method of Lagrange multipliers, D can be interpreted as the scheme of "critical points of f".

More intrinsically, the Jacobian criterion implies that $\Omega^1_{A/\mathbb{Z}_p}$, the module of differentials, is free (possibly after replacing V by a smaller neighborhood). If we choose a basis $\omega_1, \dots, \omega_n$ and let $\partial_1, \dots, \partial_n$ be the corresponding derivations $\partial_i : A \to A$ then we can let $D = \operatorname{Spec} B$, where $B = A/(\partial_1 f, \dots, \partial_n f)$; a different choice of basis for $\Omega^1_{A/\mathbb{Z}_p}$ leads to the same ideal in A. The Hessian matrix $H = (\partial_i \partial_j f)$ should be thought of as having entries in B; as such, a different choice of basis for $\Omega^1_{A/\mathbb{Z}_p}$ transforms H into $PH\,{}^tP$ with some invertible matrix P.

Assume that $x \in D(\mathbb{Z}/p^m\mathbb{Z})$ (or $x \in D(\mathbb{Z}_p)$). The Hessian matrix H_x of f at x is the Jacobian matrix of $\operatorname{grad} f$ at x, which presents $\Omega^1_{D/\mathbb{Z}_p,x}$, the stalk at x of the sheaf of differentials of D over $\mathbb{Z}_p$. Thus D is étale at x if and only if H_x is invertible. More generally, $v_p(\det H_x)$ is the length of $\Omega^1_{D/\mathbb{Z}_p,x}$ as a $\mathbb{Z}_p$-module and x is an h-étale critical point of f (Definition 1.6) if and only if $p^h \Omega^1_{D/\mathbb{Z}_p,x} = 0$. (If $p = 2$ then being a strictly h-étale critical point of is not an intrinsic property of $x \in D(\mathbb{Z}/p^m\mathbb{Z})$.) Note that if D is generically étale, so that $\mathbb{Q}_p \otimes \Omega^1_{D/\mathbb{Z}_p} = 0$, then $p^h \Omega^1_{D/\mathbb{Z}_p} = 0$ for some h, so that every $x \in D(\mathbb{Z}/p^m\mathbb{Z})$ with $m > h$ is an h-étale critical point of f.

Remark 1.11. Let $\widetilde{V} = \operatorname{Spec} \widetilde{A}$ with $\widetilde{A} = \mathbb{Z}_p[t_1, \dots, t_N]/(\widetilde{g}_1, \dots, \widetilde{g}_r)$ and $\widetilde{f} : \widetilde{V} \to \mathbb{A}^1$, where $\widetilde{f} \equiv f$ and $\widetilde{g}_i \equiv g_i \pmod{p}$. Then V is smooth if and only if $\widetilde{V}$ is and f is a "Morse function" if and only if $\widetilde{f}$ is. Furthermore, the number of critical points will be the same for f and $\widetilde{f}$; if we are only interested in estimating the sums then we may replace (V, f) with $(\widetilde{V}, \widetilde{f})$. (In the stationary phase formula, the Gauss sums will be the same for f and $\widetilde{f}$ but the exponentials will, in general, be different.)

Example 1.12. Let $V = \mathbb{G}_m = \operatorname{Spec} \mathbb{Z}_p[x, 1/x]$ and $f(x) = ax$, with $a \in \mathbb{Z}_p$. If a is a unit then $D = \emptyset$ and the sum vanishes; of course, Theorem 1.4 is just a grand generalization of the fact that the sum of a nontrivial character over a finite group vanishes. However, if a is not a unit then $D = \operatorname{Spec}(\mathbb{Z}_p/a\mathbb{Z}_p)[x, 1/x]$, which is not étale over $\mathbb{Z}_p$, so Theorem 1.4 *does not apply.* One way to phrase this caution is that if we refer to D as the scheme defined by "the vanishing of grad f", we mean "the vanishing (mod p) of grad f".

Of course, the sum is $p^m - p^{m-1}$ if $v_p(a) \geq m$, $-p^{m-1}$ if $v_p(a) = m - 1$, and it vanishes when $v_p(a) < m - 1$ (which often trips up those of us who are accustomed to the case $m = 1$). Since $D(\mathbb{Z}/p^j\mathbb{Z}) = \emptyset$ when $j > v_p(a)$, Theorem 1.8(a) gives the weaker result that the sum vanishes when $m \geq 2v_p(a) + 2$.

Example 1.13. We can recover part of the evaluation of one-dimensional Gauss sums (as in [Da, Section 2], for example), although we rely on the case $m = 1$ for odd p. Let

$$g_m(a) := \sum_{x \in \mathbb{Z}/p^m\mathbb{Z}} e^{2\pi i a x^2/p^m} = p^{m/2} G_m(2a) \tag{1.7}$$

for $a \in \mathbb{Z}_p^\times$ and $m \geq 1$. (Of course, we could let $g_m(p^j a) = p^j g_{m-j}(a)$ if $j < m$.) We have $V = \mathbb{A}^1 = \operatorname{Spec} \mathbb{Z}_p[x]$, $f(x) = ax^2$, $f'(x) = 2ax$, and $H_x = f''(x) = 2a$ so $D = \operatorname{Spec} \mathbb{Z}_p[x]/(2x)$ and $G_m(H_x) = G_m(2a) = p^{-m/2} g_m(a)$.

First consider the case $p > 2$, so that D is étale and we can apply Theorem 1.4. We find that $D(\mathbb{Z}_p) = \{0\}$ and so (using the known value of $g_1(a)$)

$$g_m(a) = \begin{cases} p^{m/2} & \text{if } 2 \mid m, \\ p^{(m-1)/2} g_1(a) = p^{m/2} \left(\dfrac{a}{p}\right) i^{(p-1)^2/4} & \text{if } 2 \nmid m > 1. \end{cases} \tag{1.8}$$

Now consider the case $p = 2$. We can take $h = 1$ or 2 (so that $m - h$ is odd) and $k = h + 1$ in Theorem 1.8(b). Again, $D(\mathbb{Z}_p) = \{0\}$, and so $g_m(a) = p^{(m-2)/2} g_2(a)$ if $m = 2j \geq 6$, $g_m(a) = p^{(m-3)/2} g_3(a)$ if $m = 2j + 1 \geq 9$. One easily calculates $g_m(a)$ by hand for $m = 1$, 2, and 3; for $m = 4$, 5, and 7 one can either calculate directly or check that the stationary phase argument still works. In terms of $\zeta_8 = e^{2\pi i/8} = (1 + i)/\sqrt{2}$ and $\varepsilon(a) = (-1)^{(a-1)/2} = \left(\frac{-1}{a}\right)$

(Jacobi symbol) one can state the result as follows:

$$g_1(a) = 0; \qquad g_m(a) = 2^{(m+1)/2}\begin{cases} \zeta_8^a & \text{if } 2 \nmid m > 1, \\ \zeta_8^{\varepsilon(a)} & \text{if } 2 \mid m. \end{cases} \tag{1.9}$$

These results can be stated more concisely in terms of the normalized Gauss sums $G_m(a) = p^{-m/2} g_m(a/2)$ (or $G_m(a) = p^{-(m+2)/2} g_{m+1}(a)$ if $p = 2$). Using the Jacobi symbol—$(a/p^m) = (a/p)^m$ if p is odd and $(2^m/a) = (-1)^{m(a^2-1)/8}$ if a is odd—one finds

$$G_m(a) = \begin{cases} \left(\dfrac{a}{p^m}\right)\zeta_8^{1-p^m} & \text{if } p \text{ is odd}, \\ \left(\dfrac{2^m}{a}\right)\zeta_8^a & \text{if } p = 2. \end{cases} \tag{1.10}$$

Remark 1.14. Let A be a symmetric, $n \times n$ matrix with entries in $\mathbb{Z}_p$ and consider the Gauss sums $G_m(A)$. If $A = \left(\begin{smallmatrix} A_1 & \\ & A_2 \end{smallmatrix}\right)$ is block-diagonal then it is easy to see that $G_m(A) = G_m(A_1)G_m(A_2)$. It follows from Example 1.13 that $G_m(A)$ is $p^{v_p(\det A)/2}$ times a root of unity (which depends on the parity of m) if A can be diagonalized, say

$$^tPAP = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{pmatrix},$$

and $m \geq v_p(a_i)$ for all i (cf. Proposition 1.3(a)). We claim that any symmetric matrix can be diagonalized, except that if $p = 2$ then we have to allow 2×2 blocks. First, factoring out a (scalar) power of p, we may assume that some entry of A is a unit. If p is odd then the polarization identity, $^txAy = \langle x, y\rangle = \frac{1}{2}(\langle x+y, x+y\rangle - \langle x, x\rangle - \langle y, y\rangle)$, shows that we may assume that the unit entry lies on the diagonal. If $p = 2$ then it is possible that all the diagonal entries are even and it is easy to see that this property will also hold for any similar matrix ^tPAP. In this case, we may assume that $a_{1,2} = a_{2,1}$ is a unit. In all cases, we may assume that A has the block form

$$A = \begin{pmatrix} P & B \\ ^tB & D \end{pmatrix},$$

where P is an invertible 1×1 or 2×2 block. Thus A is similar to

$$\begin{pmatrix} 1 & 0 \\ -{}^tBP^{-1} & 1 \end{pmatrix}\begin{pmatrix} P & B \\ ^tB & D \end{pmatrix}\begin{pmatrix} 1 & -P^{-1}B \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} P & 0 \\ 0 & D - {}^tBP^{-1}B \end{pmatrix}$$

and we are reduced to diagonalizing the smaller matrix $D - {}^tBP^{-1}B$.

To complete this proof of Proposition 1.3(a), it suffices to analyze the 2×2 blocks $A = \left(\begin{smallmatrix} a & b \\ b & d \end{smallmatrix}\right)$ with entries a, $d \in 2\mathbb{Z}_2$ and $b \in \mathbb{Z}_2^\times$. Multiplying the first row and column by b^{-1}, we may assume that $b = 1$; and then we may replace a with $(a + 2x + dx^2)/(1 + dx)^2$ or d with $(d + 2x + ax^2)/(1 + ax)^2$, for any $x \in \mathbb{Z}_2$. It is not hard to see that if either a or d is a multiple of 4 then we may assume that $a \equiv d \equiv 0 \pmod 8$; otherwise, we may assume that $a \equiv d \equiv 2 \pmod 8$. Then Hensel's Lemma (the one-dimensional case of Lemma 1.20) shows that we may assume that $a = d = 0$ or $a = d = 2$. It is easy to see that $G_m\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) = 1$ for all m. To show that $G_m\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right) = (-1)^m$, one can either calculate directly for $m = 1$ and $m = 2$ and then use stationary phase (Theorem 1.4) or one can note that

$$\begin{pmatrix} 2 & 1 & \\ 1 & 2 & 1 \\ & 1 & 0 \end{pmatrix} \text{ is similar to } \begin{pmatrix} 2 & 1 & \\ 1 & 2 & \\ & & -2/3 \end{pmatrix} \text{ and to } \begin{pmatrix} 2 & & \\ & 0 & 1 \\ & 1 & 0 \end{pmatrix},$$

so that

$$G_m\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} G_m(-2/3) = G_m(2) G_m\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

(The one-dimensional sums vanish for $m = 1$, so this case has to be checked separately.) Note that this example illustrates that the decomposition into 1×1 and 2×2 blocks is not unique. (One way to see that $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ are not similar is to note that their determinants differ by a factor of -3, which is not a square in $\mathbb{Z}_2$.)

Proof of Proposition 1.3(c). Using the fact that $G_h(A; v) = G_h({}^tPAP; {}^tPv)$ and arguing as above, we reduce to the case that A is a 1×1 matrix or (if $p = 2$) one of the standard 2×2 matrices. If p is odd then we reduce to the case $v = 0$ by part (b), and this is dealt with in Example 1.13, above. If $p = 2$ then, keeping part (b) in mind, one reduces the problem to a short calculation. ■

Example 1.15. We can evaluate the Kloosterman sum ($a \in \mathbb{Z}_p^\times$)

$$K(a; \mathbb{Z}/p^m\mathbb{Z}) := \sum_{x \in (\mathbb{Z}/p^m\mathbb{Z})^\times} e^{2\pi i(x + a/x)/p^m} = \sum_{\substack{x, y \in \mathbb{Z}/p^m\mathbb{Z} \\ xy = a}} e^{2\pi i(x+y)/p^m}$$

when $m > 1$, recovering Salié's formulae [Sa]. (When $m = 1$ we have the Hasse–Weil bound [W1]: $|K(a; \mathbb{F}_p)| \le 2\sqrt{p}$.)

The first method is to let $V = \mathbb{G}_m = \operatorname{Spec} \mathbb{Z}_p[x, 1/x]$ and $f(x) = x + a/x$. The dimension of V is $n = 1$ and D is defined by $1 - a/x^2 = 0$, or $x^2 = a$. If $x \in D(\mathbb{Z}_p)$ then $f(x) = 2x$ and, since $f''(x) = 2a/x^3$, the Hessian is simply the 1×1 matrix $H_x = (2a/x^3) = (2/x)$, so $G_m(H_x) = G_m(2/x) = G_m(2x)$. Theorem 1.8(b) gives

$$K(a; \mathbb{Z}/p^m\mathbb{Z}) = p^{m/2} \sum_{\substack{x \in \mathbb{Z}_p \\ x^2 = a}} e^{4\pi i x/p^m} G_m(2x) \tag{1.11}$$

for all $m \ge 2$ if p is odd; and for $m = 6$ and $m \ge 8$ if $p = 2$.

Suppose $p > 2$. Then $G_m(2x) = 1$ if m is even and, if m is odd,

$$G_m(2x) = p^{-1/2}g_1(x) = \left(\frac{x}{p}\right)\begin{cases} 1 & \text{if } p \equiv 1 \pmod 4, \\ i & \text{if } p \equiv 3 \pmod 4. \end{cases}$$

Therefore $K(a;\mathbb{Z}/p^m\mathbb{Z}) = 0$ if a is not a square; and if $a = x^2$ then

$$\begin{aligned} (1.12)\qquad & K(a;\mathbb{Z}/p^m\mathbb{Z}) \\ &= p^{m/2}\begin{cases} 2\cos(4\pi x/p^m) & \text{if } 2 \mid m, \\ 2\left(\frac{x}{p}\right)\cos(4\pi x/p^m) & \text{if } 2\nmid m,\ p \equiv 1 \pmod 4, \\ -2\left(\frac{x}{p}\right)\sin(4\pi x/p^m) & \text{if } 2\nmid m,\ p \equiv 3 \pmod 4. \end{cases} \end{aligned}$$

Now suppose $p = 2$. We have $D(\mathbb{Z}/p^j\mathbb{Z}) = \emptyset$ if $j \geq 3$ and $a \not\equiv 1 \pmod 8$; by Theorem 1.8(a), the Kloosterman sum vanishes if $m \geq 6$ and $a \not\equiv 1 \pmod 8$. If $a \equiv 1 \pmod 8$ then let $\alpha \in \mathbb{Z}_p$ be a square root of a; if m is even then it is convenient to fix the sign by choosing $\alpha \equiv 1 \pmod 4$. Evaluating (1.11), one finds

$$\begin{aligned} (1.13)\qquad & K(a;\mathbb{Z}/2^m\mathbb{Z}) \\ &= 2^{(m+1)/2}\begin{cases} 0 & \text{if } a \not\equiv 1 \pmod 8,\ m \geq 6; \\ 2\cos\left(\frac{4\pi\alpha}{2^m} + \frac{\pi}{4}\right) & \text{if } a = \alpha^2,\ \alpha \equiv 1 \pmod 4,\ \ 2 \mid m \geq 6; \\ 2\cos\left(\frac{4\pi\alpha}{2^m} + \frac{\pi\alpha}{4}\right) & \text{if } a = \alpha^2,\ 2\nmid m \geq 9. \end{cases} \end{aligned}$$

Calculating $K(a;\mathbb{Z}/2^m\mathbb{Z})$ for $m \leq 5$ and $m = 7$ is an easy exercise, but it is instructive: it shows that the bounds on m in Theorem 1.8(b) are sometimes needed and that for small m the absolute value of $K(a;\mathbb{Z}/2^m\mathbb{Z})$, and the values of a for which it is non-zero, do not follow the general pattern. (Besides, Salié does not seem to cover all the cases.) We therefore record the results. The formula

$$(1.14)\qquad K(a;\mathbb{Z}/2^m\mathbb{Z}) = 2^{m-1}\cos\left(\frac{2\pi(a+1)}{2^m}\right)$$

holds for $m = 1$ and 2; $m = 3$ and 4, $a \equiv 3 \pmod 4$; and $m = 5$, $a \equiv 5 \pmod 8$. For $m = 7$,

$$(1.15)\qquad K(a;\mathbb{Z}/2^m\mathbb{Z}) = 2^{m-2}\cos\left(\frac{2\pi(\alpha + a/\alpha)}{2^{m-1}}\right),$$

where $\alpha = 3$ if $a \equiv 1 \pmod{16}$ and $\alpha = 1$ if $a \equiv 9 \pmod{16}$. In all other cases, $K(a;\mathbb{Z}/2^m\mathbb{Z}) = 0$. Note that $|K(a;\mathbb{Z}/2^m\mathbb{Z})|$ is 2^{m-1} for $m \leq 4$, when it is non-zero.

The second method is to let $V \subseteq \mathbb{A}^2$ be defined by $xy = a$ and let $f(x,y) = x + y$. The Jacobian matrix associated with V is $(\,y \quad x\,)$; since y

and x generate the unit ideal in $\mathbb{Z}_p[x,y]/(xy-a)$, V is smooth. The gradient of f is $(1 \quad 1)$, and so D is defined by $xy-a=0$ and $0=\det\left(\begin{smallmatrix}1 & 1\\ y & x\end{smallmatrix}\right)=x-y$. The associated Jacobian matrix is $\left(\begin{smallmatrix}y & x\\ 1 & -1\end{smallmatrix}\right)$, which has determinant $-y-x$. In the coordinate ring $\mathbb{Z}_p[x,y]/(xy-a,x-y)$ of D, $-y-x=-2x$ and so $v_p(\det H_{(x,y)})=v_p(2)$. Calculating the Hessian exactly is much like using the first method.

Example 1.16. Consider the 3-variable Kloosterman sum ($a \in \mathbb{Z}_p^{\times}$)

$$K_3(a;\mathbb{Z}/p^m\mathbb{Z}) := \sum_{x,y\in(\mathbb{Z}/p^m\mathbb{Z})^{\times}} e^{2\pi i(x+y+a/xy)/p^m} = \sum_{\substack{x,y,z\in\mathbb{Z}/p^m\mathbb{Z}\\ xyz=a}} e^{2\pi i(x+y+z)/p^m}.$$

When $m=1$, Deligne [D] generalizes the Hasse–Weil bound: $|K_3(a;\mathbb{F}_p)|\le 3p$. Larsen estimates these sums for $m>1$ in [L], but his bound is not sharp for $p=3$.

Following the first method in Example 1.15, let $V=\operatorname{Spec}\mathbb{Z}_p[x,y,1/xy]$, or $V=\mathbb{G}_m\times\mathbb{G}_m$, and $f(x,y)=x+y+a/xy$. Then $D=\operatorname{Spec}\mathbb{Z}_p[x]/(x^3-a)$ and the Hessian at a point $x\in D(\mathbb{Z}_p)$ is $H_x=x^{-1}\left(\begin{smallmatrix}2 & 1\\ 1 & 2\end{smallmatrix}\right)$.

Assume first that $p\neq 3$, so that Theorem 1.4 applies. We find

$$K_3(a;\mathbb{Z}/p^m\mathbb{Z})=p^m\sum_{\substack{x\in\mathbb{Z}_p\\ x^3=a}} e^{2\pi i(3x)/p^m}G_m(H_x).$$

If $p>3$ then one can diagonalize the Hessian (as a bilinear form) and (1.10) implies that

$$G_m(H_x)=G_m\left(\frac{2}{x}\right)G_m\left(\frac{3}{2x}\right)=\left(\frac{3}{p^m}\right)\left(\frac{-1}{p^m}\right)=\left(\frac{p^m}{3}\right).$$

If $p=2$ then one calculates the Gauss sum as in Remark 1.14:

$$G_m(H_x)=(-1)^m=\left(\frac{2^m}{3}\right).$$

Thus

$$K_3(a;\mathbb{Z}/p^m\mathbb{Z})=\left(\frac{p^m}{3}\right)p^m\sum_{\substack{x\in\mathbb{Z}_p\\ x^3=a}} e^{2\pi i(3x)/p^m} \qquad (p\neq 3, m>1). \tag{1.16}$$

In particular, $|K_3(a;\mathbb{Z}/p^m\mathbb{Z})|\le 3p^m$.

Now let $p=3$. We can apply Theorem 1.8(b) with $h=1$ and $k=2$. Since $D(\mathbb{Z}_3)$ is the set of cube roots of a in $\mathbb{Z}_3$, it is empty unless $a\equiv\pm1 \pmod 9$; if a does have a cube root, it is unique and we denote it by x. If

$m \geq 5$, we find

$$K_3(a;\mathbb{Z}/3^m\mathbb{Z}) = 3^m e^{2\pi i(3x)/3^m} G_m \begin{pmatrix} 2/x & 1/x \\ 1/x & 2/x \end{pmatrix}.$$

As before, we can diagonalize the Hessian and express the Gauss sum as a product of one-variable Gauss sums: we get

$$G_m\left(\frac{2}{x}\right)G_m\left(\frac{3}{2x}\right) = 3^{1/2}G_m\left(\frac{2}{x}\right)G_{m-1}\left(\frac{1}{2x}\right) = \left(\frac{x}{3}\right)i3^{1/2},$$

since one of m, $m-1$ is even and the other is odd. Therefore

$$(1.17)\qquad K_3(a;\mathbb{Z}/3^m\mathbb{Z}) = 3^{m+1/2}e^{2\pi i(3x)/3^m}\left(\frac{x}{3}\right)i \qquad (a = x^3,\ m \geq 5).$$

Calculating the sums for small values of m, one finds that $K_3(a;\mathbb{Z}/3^m\mathbb{Z}) = 0$ if $m \geq 3$ unless $a \equiv \pm 1 \pmod 9$. Furthermore,

$$(1.18)\qquad K_3(a;\mathbb{Z}/3^m\mathbb{Z}) = \begin{cases} 3^m e^{2\pi i(2a+1/a)/3^m} & \text{if } m = 2, \\ 3^{m+1/2}e^{2\pi i(2a+1/a)/3^m}\left(\frac{x}{3}\right)i & \text{if } m = 3,\ a = x^3, \\ 3^{m+1/2}e^{2\pi i(3x)/3^m}e^{2\pi i(-a)/3}\left(\frac{x}{3}\right)i & \text{if } m = 4,\ a = x^3. \end{cases}$$

In all cases, the absolute value is bounded by $3^{m+1/2} = 3^m\sqrt{3}$, an *improvement* of $\sqrt{3}$ over the case $p \neq 3$ and a factor of 3 better than the bound in [L]. The increase in size of the local terms has been more than offset by the fact that $D(\mathbb{Z}_3)$ has at most one element.

EXAMPLE 1.17. Similarly, we consider the n-variable Kloosterman sum $(a \in \mathbb{Z}_p^\times)$

$$K_n(a;\mathbb{Z}/p^m\mathbb{Z}) := \sum_{\substack{x_1,\ldots,x_n \in \mathbb{Z}/p^m\mathbb{Z} \\ x_1\ldots x_n = a}} e^{2\pi i(x_1+\ldots+x_n)/p^m},$$

recovering the results of [Sm1]. Again, [D] gives $|K_n(a;\mathbb{F}_p)| \leq np^{(n-1)/2}$ for $m = 1$. Just as in Example 1.16, we find that $|K_n(a;\mathbb{Z}/p^m\mathbb{Z})| \leq np^{(n-1)m/2}$ if $p\nmid n$ and, for $h = v_p(n) > 0$ and $m \geq 3h+2$ ($m \geq 3h+6$ if $p = 2$), $|K_n(a;\mathbb{Z}/p^m\mathbb{Z})| \leq (p^{h/2}|D(\mathbb{Z}_p)|)p^{(n-1)m/2}$. Since $D(\mathbb{Z}_p)$ is the set of nth roots of a and p^hth roots are unique, when they exist, in $\mathbb{Z}_p^\times$ (unless $p = 2$, in which case there are 0 or 2 p^hth roots), we find that $|D(\mathbb{Z}_p)| \leq n/p^{h-v_p(2)}$. This leads to $|K_n(a;\mathbb{Z}/p^m\mathbb{Z})| \leq p^{v_p(2)-h/2}np^{(n-1)m/2}$.

If $1 < m < 3h+2$ ($1 < m < 3h+5$ if $p = 2$) then we use Theorem 1.8(a). It is easy to see that the Hessian matrix has rank $n-2$ on $\mathbb{F}_p^{n-1}$, so $|K_n(a;\mathbb{Z}/p^m\mathbb{Z})|$ is bounded by $|D(\mathbb{Z}/p^j\mathbb{Z})|p^{1/2}p^{(n-1)m/2}$. Let $m = 2j$ or $2j+1$. Since $D(\mathbb{Z}/p^j\mathbb{Z})$ is the set of nth roots of a in $\mathbb{Z}/p^j\mathbb{Z}$ and

$(\mathbb{Z}/p^j\mathbb{Z})^\times$ is cyclic (or $\{\pm 1\}$ times a cyclic group if $p = 2$) one finds that $|D(\mathbb{Z}/p^j\mathbb{Z})| = 0$ or $p^{v_p(2)+\min\{h,j-1-v_p(2)\}}(n, p-1)$ (where $(n, p-1)$ denotes the greatest common divisor). This leads to the bound $|K_n(a;\mathbb{Z}/p^m\mathbb{Z})| \le p^{v_p(2)+\min\{h,j-1-v_p(2)\}}(n, p-1)p^{(n-1)m/2}$.

Proofs. There are three steps in the proof of Theorem 1.8. First, we show that the fibers of the reduction map

$$\varrho : V(\mathbb{Z}/p^m\mathbb{Z}) \to V(\mathbb{Z}/p^j\mathbb{Z}) \tag{1.19}$$

are all isomorphic to $(p^j\mathbb{Z}/p^m\mathbb{Z})^n$. This allows us to reduce to the case $V = \mathbb{A}^n$; as a bonus, we recover the standard fact (a generalization of Hensel's Lemma) that $D(\mathbb{Z}_p) \xrightarrow{\sim} D(\mathbb{Z}/p^j\mathbb{Z})$ if D is étale. (This is the only step that involves the language of schemes. We suppose that there are other languages that also suffice to express the idea that if V is smooth and n-dimensional then every point $(\operatorname{mod} p^j)$ of V corresponds to $p^{n(m-j)}$ points $(\operatorname{mod} p^m)$ of V.) The second step is to show that $S_{\overline{x}} = 0$ if $\overline{x} \in V(\mathbb{Z}/p^j\mathbb{Z}) \setminus D(\mathbb{Z}/p^j\mathbb{Z})$ and the third step is to evaluate $S_{\overline{x}}$ when $x \in D(\mathbb{Z}_p)$.

LEMMA 1.18. *Let $f : V \to \mathbb{A}^1$, $\overline{x} \in V(\mathbb{Z}/p^j\mathbb{Z})$, and $S_{\overline{x}}$ be as in Theorem* 1.8. *Let $x_0 \in V$ be the closed point corresponding to $\overline{x}$ and fix an isomorphism of $\widehat{\mathcal{O}}_{V,x_0}$ with $\mathbb{Z}_p[[t_1, \ldots, t_n]]$* ([D-G, Smoothness Theorem, p. 137] *or* [EGA IV, Proposition 17.5.3]); *do this in such a way that $\overline{x}(t_i) = 0 \in \mathbb{Z}/p^j\mathbb{Z}$ for all i. Let $\widetilde{f} \in \Gamma(V, \mathcal{O}_V)$ correspond to f and* (*by abuse of notation*) *also let $\widetilde{f}$ denote its image in $\mathbb{Z}_p[[t_1, \ldots, t_n]]$. If $m \ge j$ then*

$$S_{\overline{x}} = \sum_{z \in (p^j\mathbb{Z}/p^m\mathbb{Z})^n} e^{2\pi i \tilde{f}(z)/p^m}. \tag{1.20}$$

That is, $S_{\overline{x}} = \widetilde{S}_{\overline{0}}$, where $\widetilde{S}_{\overline{0}}$ is the sum corresponding to $\widetilde{f} : \mathbb{A}^n \to \mathbb{A}^1$ (*more precisely*, *the map corresponding to any polynomial congruent to $\widetilde{f}$ modulo $(t_1, \ldots, t_n)^m$*) *and $\overline{0} \in \mathbb{A}^n(\mathbb{Z}/p^j\mathbb{Z})$. Furthermore, $\overline{x}$ is a critical point of f if and only if $\overline{0}$ is a critical point of $\widetilde{f}$*; *if so*, *the Hessian of f at $\overline{x}$ is the same as the Hessian of $\widetilde{f}$ at $\overline{0}$.*

Proof. A $\mathbb{Z}/p^j\mathbb{Z}$-valued point $\overline{x}$ of V can be thought of as a closed point $x_0 \in V$ and a local homomorphism $\overline{x} : \mathcal{O}_{V,x_0} \to \mathbb{Z}/p^j\mathbb{Z}$. This extends naturally to a map (also denoted $\overline{x}$, by abuse of notation) $\widehat{\mathcal{O}}_{V,x_0} \to \mathbb{Z}/p^j\mathbb{Z}$. Thus the expression $\overline{x}(t_i)$ in the statement of the lemma makes sense.

Similarly, any $x \in V(\mathbb{Z}/p^m\mathbb{Z})$ that reduces to $\overline{x}$ can be thought of as a local homomorphism $x : \mathcal{O}_{V,x_0} \to \mathbb{Z}/p^m\mathbb{Z}$ such that $\overline{x}$ is the composition of x with the reduction map $\mathbb{Z}/p^m\mathbb{Z} \to \mathbb{Z}/p^j\mathbb{Z}$. In other words, the fiber of ϱ over $\overline{x}$ is the set of lifts to $\mathbb{Z}/p^m\mathbb{Z}$ of $\overline{x} : \mathcal{O}_{V,x_0} \to \mathbb{Z}/p^j\mathbb{Z}$ or of $\overline{x} : \widehat{\mathcal{O}}_{V,x_0} \to \mathbb{Z}/p^j\mathbb{Z}$. Identifying $\widehat{\mathcal{O}}_{V,x_0}$ with $\mathbb{Z}_p[[t_1, \ldots, t_n]]$, a lift x is parameterized by the n-tuple

$\widetilde{x} = (x(t_1), \ldots, x(t_n))$ in $(p^j\mathbb{Z}/p^m\mathbb{Z})^n$. Since $f(x) = \widetilde{f}(\widetilde{x}) \in \mathbb{Z}/p^m\mathbb{Z}$, (1.20) follows.

As in the Explicitation subsection, choose a basis $\partial_1, \ldots, \partial_n$ of the $\mathcal{O}_{V,x_0}$-module $\mathrm{Der}_{\mathbb{Z}_p}(\mathcal{O}_{V,x_0}, \mathcal{O}_{V,x_0})$ of $\mathbb{Z}_p$-linear derivations $\partial : \mathcal{O}_{V,x_0} \to \mathcal{O}_{V,x_0}$. Thus $\overline{x}$ is a critical point of f if and only if $\overline{x}(\partial_i \widetilde{f}) = 0$ for all i. Then $\partial_1, \ldots, \partial_n$ is also a basis of $\mathrm{Der}_{\mathbb{Z}_p}(\widehat{\mathcal{O}}_{V,x_0}, \widehat{\mathcal{O}}_{V,x_0})$ (as an $\widehat{\mathcal{O}}_{V,x_0}$-module). Since $\partial/\partial t_1, \ldots, \partial/\partial t_n$ is another such basis, $\overline{x}$ is a critical point of f if and only if $\partial \widetilde{f}/\partial t_i|_{\overline{0}} = 0 \in \mathbb{Z}/p^j\mathbb{Z}$ for all i. Similarly, if $\overline{x}$ is a critical point then the matrices $H = (\overline{x}(\partial_i \partial_j \widetilde{f}))$ and $\widetilde{H} = ((\partial^2 \widetilde{f}/\partial t_i \partial t_j)|_{\overline{0}})$ are related by $\widetilde{H} = {}^tPHP$, where P is the change-of-basis matrix. ■

Corollary 1.19 (Hensel's Lemma). *Let D be an étale scheme over $\mathbb{Z}_p$. The reduction map $D(\mathbb{Z}_p) \to D(\mathbb{Z}/p^j\mathbb{Z})$ is a bijection for all $j \geq 1$.*

Proof. Since "étale" means "smooth, of relative dimension 0", the argument in the lemma applies (taking $n = 0$ and "$m = \infty$", so to speak). ■

Proof of Theorem 1.8(a). We may assume, thanks to Lemma 1.18, that $V = \mathbb{A}^n$ and $\overline{x} \in \mathbb{A}^n(\mathbb{Z}/p^j\mathbb{Z}) = (\mathbb{Z}/p^j\mathbb{Z})^n$. As x runs over the fiber $\varrho^{-1}(\overline{x})$ and y runs over $(\mathbb{Z}/p^j\mathbb{Z})^n$, $x + p^{m-j}y$ runs through the fiber, taking on every value p^{nj} times. (If we restricted x to a set of coset representatives, we could count every element of the fiber only once.) Furthermore, since we are assuming $m \geq 2j$,

$$f(x + p^{m-j}y) = f(x) + p^{m-j}\,\mathrm{grad}\, f(x) \cdot y \in \mathbb{Z}/p^m\mathbb{Z}.$$

Therefore,

$$\begin{aligned} S_{\overline{x}} &= \frac{1}{p^{nj}} \sum_{x \in \varrho^{-1}(\overline{x})} \sum_{y \in (\mathbb{Z}/p^j\mathbb{Z})^n} e^{2\pi i f(x + p^{m-j}y)/p^m} \\ &= \sum_{x \in \varrho^{-1}(\overline{x})} e^{2\pi i f(x)/p^m} \cdot \frac{1}{p^{nj}} \sum_{y \in (\mathbb{Z}/p^j\mathbb{Z})^n} e^{2\pi i\, \mathrm{grad}\, f(x) \cdot y/p^j}. \end{aligned}$$

The inner sum vanishes unless $\mathrm{grad}\, f(x) \equiv 0 \pmod{p^j}$, i.e., unless $\overline{x} \in D(\mathbb{Z}/p^j\mathbb{Z})$. (Note that for small values of j the fiber $\varrho^{-1}(\overline{x})$ is large and we break it up into many small pieces; the sum over each piece vanishes unless $\overline{x} \in D(\mathbb{Z}/p^j\mathbb{Z})$.)

Now suppose that $m = 2j$ or $2j + 1$ and let $x \in \mathbb{A}^n(\mathbb{Z}/p^m\mathbb{Z})$ map to $\overline{x} \in D(\mathbb{Z}/p^j\mathbb{Z})$. The fiber $\varrho^{-1}(\overline{x})$ is parameterized by $x + p^j y$ with $y \in \mathbb{A}^n(\mathbb{Z}/p^{m-j}\mathbb{Z})$. If $m = 2j$ then $f(x + p^j y) = f(x) + p^j\, \mathrm{grad}\, f(x) \cdot y \in \mathbb{Z}/p^m\mathbb{Z}$; if $m = 2j + 1$ then there is an extra term $\frac{1}{2}p^{2j}H_x(y)$. In either case, the formula for $S_{\overline{x}}$ follows easily. The estimate on $|S|$ follows from Proposition 1.3(c). ■

We now turn to the proof of Theorem 1.8(b). In order to relax Katz's hypothesis that D be étale, we start with another version of Hensel's Lemma.

This is a simpler, but more explicit, version of [G, §3 Lemma 2]. We could give a similar proof, which would work over any Henselian discrete valuation ring, but we prefer to give one along the lines of the usual proof of Hensel's Lemma.

LEMMA 1.20 (Hensel's Lemma revisited). *Let* $f_1,\ldots,f_n\in\mathbb{Z}_p[x_1,\ldots,x_n]$ *and* $a\in\mathbb{Z}_p^n$. *Let* $J=\partial(f_1,\ldots,f_n)/\partial(x_1,\ldots,x_n)$ *be the Jacobian matrix*; *assume that* $\det J(a)\neq 0$ *and let* h *be large enough that* $p^hJ(a)^{-1}$ *has entries in* $\mathbb{Z}_p$. *If* $j\geq h+1$ *and the column vector* $F(a)=(f_1(a),\ldots,f_n(a))$ *lies in* $p^jJ(a)\mathbb{Z}_p^n$, *say* $F(a)=p^jJ(a)b$, *then there is a unique* $\alpha\in\mathbb{Z}_p^n$ *such that* $F(\alpha)=0$ *and* $\alpha\equiv a \pmod{p^j}$. *In fact*, α *is the unique root of* F *such that* $\alpha\equiv a \pmod{p^{h+1}}$.

Note that $F(a)\in p^jJ(a)\mathbb{Z}_p^n$ is implied by $F(a)\equiv 0 \pmod{p^{h+j}}$. By Cramer's Rule, the lemma applies with $h=v_p(\det J(a))$.

Proof. By the polynomial version of Taylor's theorem,

$$F(a+p^jx)\equiv F(a)+p^jJ(a)x \pmod{p^{2j}}.$$

Letting $a_1=a-p^jb$, we find $F(a_1)\equiv 0 \pmod{p^{2j}}$, so $F(a_1)=p^{2j-h}J(a)b_1$. Since $j_1=2j-h>j$, we can iterate this process. Taking $\alpha=\lim_{n\to\infty}a_n$, we find $F(\alpha)=0$ and $\alpha\equiv a \pmod{p^j}$.

For the uniqueness statement, suppose $F(\alpha)=0=F(\alpha+p^jx)$ with $j\geq h+1$ and $x\in\mathbb{Z}_p^n\setminus p\mathbb{Z}_p^n$. Then $p^jJ(a)x\equiv 0 \pmod{p^{2j}}$, which implies $x\equiv 0 \pmod{p^{j-h}}$, a contradiction. ■

Proof of Theorem 1.8(b). Since $m\geq 3h+2$, we find that $j\geq h+1$ and $3j\geq m$. Let $\overline{x}\in V(\mathbb{Z}/p^j\mathbb{Z})$. First, note that if $j<k$ and no $\overline{y}\in D(\mathbb{Z}/p^k\mathbb{Z})$ reduces to $\overline{x}$ then $S_{\overline{x}}=0$, by part (a). We may assume, therefore, that $H_{\overline{x}}$ divides p^hI_n. (The awkward condition on h and k in the statement of Theorem 1.8(b) is intended to insure that this argument applies.)

Let $x\in V(\mathbb{Z}_p)$ be a representative of $\overline{x}$. By Lemma 1.18, we may assume that $V=\mathbb{A}^n$. Since $3j\geq m$, $f(x+p^jy)\equiv f(x)+p^j\operatorname{grad}f(x)\cdot y+\frac{1}{2}p^{2j}H_x(y) \pmod{p^m}$ and so

$$S_{\overline{x}}=e^{2\pi if(x)/p^m}\sum_{y\in(\mathbb{Z}/p^{m-j}\mathbb{Z})^n}e^{2\pi i(\operatorname{grad}f(x)\cdot y+\frac{1}{2}p^jH_x(y))/p^{m-j}}.$$

As in the proof of Proposition 1.3(a), think of $(\mathbb{Z}/p^{m-j}\mathbb{Z})^n$ as $\mathbb{Z}_p^n/p^{m-j}\mathbb{Z}_p^n$ and note that $H_x'\mathbb{Z}_p^n\supseteq p^{m-j}\mathbb{Z}_p^n$, where H_x' is symmetric and $H_x'H_x=p^{m-2j}I_n$. If $y\in\mathbb{Z}_p^n$ and $z\in H_x'\mathbb{Z}_p^n$ then $\frac{1}{2}p^jH_x(y+z)\equiv\frac{1}{2}p^jH_x(y)$

(mod p^{m-j}). Therefore

$$S_{\overline{x}} = e^{2\pi i f(x)/p^m} \sum_{y\in\mathbb{Z}_p^n/H_x'\mathbb{Z}_p^n} e^{2\pi i(\operatorname{grad} f(x)\cdot y+\frac{1}{2}p^j H_x(y))/p^{m-j}} \times \sum_{z\in H_x'\mathbb{Z}_p^n/p^{m-j}\mathbb{Z}_p^n} e^{2\pi i \operatorname{grad} f(x)\cdot z/p^{m-j}},$$

where $\mathbb{Z}_p^n/H_x'\mathbb{Z}_p^n$ means a set of coset representatives. The inner sum vanishes unless $p^{m-j} \mid \operatorname{grad} f(x)\cdot z$ for all $z \in H_x'\mathbb{Z}_p^n$, which implies $\operatorname{grad} f(x) \in p^j H_x\mathbb{Z}_p^n$.

Assume now that $S_{\overline{x}} \neq 0$, so that $\operatorname{grad} f(x) \in p^j H_x\mathbb{Z}_p^n$. By Lemma 1.20, we may assume $\operatorname{grad} f(x) = 0$, i.e., $x \in D(\mathbb{Z}_p)$. Therefore

$$S_{\overline{x}} = e^{2\pi i f(x)/p^m} \sum_{y\in(\mathbb{Z}/p^{m-j}\mathbb{Z})^n} e^{\pi i H_x(y)/p^{m-2j}}$$

and this sum equals $p^{nj}\sqrt{p}^{\,n(m-2j)}G_{m-2j}(H_x)$. Finally, $G_{m-2j}(H_x) = G_m(H_x)$ by Proposition 1.3. ■

2. The GL(3)-Kloosterman sum for the long element of the Weyl group. In this section, we compute the local GL(3)-Kloosterman sums attached to the long element of the Weyl group. Our first result, Theorem 2.4, is implicit in [S] (cf. Remark 2.5(2)). The rest of the section is devoted to expressing our results in terms of sums of products of classical Kloosterman sums $S(\mu,\nu;p^m) = S(1,\mu\nu;p^m)$ where $p\nmid\mu\nu$. Our result, rather messy for small values of r and s (notation as in Theorem 2.4), is given in Theorem 2.11. In the case where $r = s$ is large, further analysis of the sums of products in Section 3 leads to yet another expression for the GL(3)-Kloosterman sum in Theorem 3.7(a).

Our notation mostly follows [S].

Notation 2.1. Let p be a fixed prime and let $\mathbb{Q}_p$ and $\mathbb{Z}_p$ denote, respectively, the field of p-adic numbers and the ring of p-adic integers, and let v_p be the valuation on $\mathbb{Q}_p$. Let

$$R_m = \mathbb{Z}_p/p^m\mathbb{Z}_p = \mathbb{Z}/p^m\mathbb{Z} \quad (m \geq 1).$$

We will usually write $\mathbb{F}_p$ instead of R_1. We will let

$$e(x) = e^{2\pi i x}, \qquad e_m(x) = e(x/p^m).$$

We will use a variation on the Kronecker delta: if $\mathcal{P}$ is some condition, let

$$\delta_{\mathcal{P}} = \begin{cases} 1 & \text{if } \mathcal{P} \text{ holds,} \\ 0 & \text{otherwise.} \end{cases}$$

For example, $\delta_{m=1}$ means the same thing as the traditional Kronecker delta $\delta_{m,1}$.

Let $G = \mathrm{GL}(3, \mathbb{Q}_p)$. Let $U \subseteq G$ denote the set of all unipotent matrices

$$u(x_1, x_2, x_3) = \begin{pmatrix} 1 & x_1 & x_3 \\ 0 & 1 & x_2 \\ 0 & 0 & 1 \end{pmatrix}$$

with $x_1, x_2, x_3 \in \mathbb{Q}_p$. Let T denote the diagonal subgroup of G and $W = N_G(T)/T$ denote the Weyl group of G relative to T. We will identify W with the symmetric group S_3. Let

$$w_0 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \leftrightarrow (13)$$

be the *long element* of W.

Let ϱ denote the reduction modulo p homomorphism from $G(\mathbb{Z}_p)$ onto $G(\mathbb{F}_p)$. Let $\overline{B}$ denote the group of upper triangular matrices in $G(\mathbb{F}_p)$. Then $B = \varrho^{-1}(\overline{B})$ is the standard *Iwahori subgroup* of G. For any $\tau \in W$ let

$$B(\tau) = B\tau B \qquad \text{(the } \textit{Iwahori cell} \text{ corresponding to } \tau\text{)}.$$

Then

$$G(\mathbb{Z}_p) = \bigsqcup_{\tau \in W} B(\tau) \qquad \text{(the } \textit{Iwahori decomposition}\text{)}$$

since we have the Bruhat decomposition $G(\mathbb{F}_p) = \bigsqcup_{\tau \in W} \overline{B}\tau\overline{B}$.

Lemma 2.2. *Let*

$$A = (a_{ij}) \in G(\mathbb{Z}_p), \qquad A_{13} = \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix},$$

and let $B(\tau)$ be the Iwahori cell containing A. Then

- $\tau = w_0$ *if and only if* a_{31}, $A_{13} \in \mathbb{Z}_p^\times$;
- $\tau = \begin{pmatrix} & 1 & \\ & & 1 \\ 1 & & \end{pmatrix} \leftrightarrow (132)$ *if and only if* $a_{31} \in \mathbb{Z}_p^\times$ *and* $A_{13} \in p\mathbb{Z}_p$;
- $\tau = \begin{pmatrix} & & 1 \\ 1 & & \\ & 1 & \end{pmatrix} \leftrightarrow (123)$ *if and only if* $a_{31} \in p\mathbb{Z}_p$ *and* a_{21}, $a_{32} \in \mathbb{Z}_p^\times$;
- $\tau = \begin{pmatrix} & 1 & \\ 1 & & \\ & & -1 \end{pmatrix} \leftrightarrow (12)$ *if and only if* a_{31}, $a_{32} \in p\mathbb{Z}_p$ *and* $a_{21} \in \mathbb{Z}_p^\times$;
- $\tau = \begin{pmatrix} -1 & & \\ & & 1 \\ & 1 & \end{pmatrix} \leftrightarrow (23)$ *if and only if* a_{31}, $a_{21} \in p\mathbb{Z}_p$ *and* $a_{32} \in \mathbb{Z}_p^\times$;
- $\tau = e = \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} \leftrightarrow (1)$ *if and only if* a_{31}, a_{21}, $a_{32} \in p\mathbb{Z}_p$.

Proof. The "only if" direction is easily checked. Since the conditions are mutually exclusive and the Iwahori cells partition $G(\mathbb{Z}_p)$, the "if" direction follows. ∎

NOTATION 2.3. For any $t \in T$ and $\tau \in W$ let

$$C(w_0t) = (Uw_0tU) \cap G(\mathbb{Z}_p), \qquad X(w_0t) = U(\mathbb{Z}_p)\backslash C(w_0t)/U(\mathbb{Z}_p),$$
$$C_\tau(w_0t) = (Uw_0tU) \cap B(\tau), \qquad X_\tau(w_0t) = U(\mathbb{Z}_p)\backslash C_\tau(w_0t)/U(\mathbb{Z}_p).$$

An elementary calculation (cf. (2.5)) shows that, in order for $C(w_0t) \neq \emptyset$, we must have $t \in \operatorname{diag}(p^s, p^{r-s}, p^{-r})T(\mathbb{Z}_p)$ for some non-negative integers r and s.

The continuous characters on U, trivial on $U(\mathbb{Z}_p)$, are of the form

$$\psi_{\nu_1,\nu_2}(u(x_1,x_2,x_3)) = e^{2\pi i(\nu_1x_1+\nu_2x_2)},$$

where ν_1 and ν_2 are p-adic integers. We say that ψ_{ν_1,ν_2} is *regular* if and only if ν_1 and ν_2 are non-zero.

Let $t \in T$ and fix characters $\psi = \psi_{\nu_1,\nu_2}$ and $\psi' = \psi_{\nu_1',\nu_2'}$ of U, trivial on $U(\mathbb{Z}_p)$. The corresponding long-element Kloosterman sum is defined by

$$\mathrm{Kl}(w_0t,\psi,\psi') = \sum \psi(u)\psi'(u'),$$

where uw_0tu' runs over a set of representatives of $X(w_0t)$. For any $\tau \in W$ let $\mathrm{Kl}_\tau(w_0t,\psi,\psi')$ denote the corresponding sum, where uw_0tu' runs over a set of representatives of $X_\tau(w_0t)$. Clearly

$$\mathrm{Kl}(w_0t,\psi,\psi') = \sum_{\tau\in W} \mathrm{Kl}_\tau(w_0t,\psi,\psi'),$$

corresponding to the Iwahori decomposition of $G(\mathbb{Z}_p)$ (or the Bruhat decomposition of $G(\mathbb{F}_p)$).

We will evaluate $\mathrm{Kl}(w_0t,\psi,\psi')$ by computing $\mathrm{Kl}_\tau(w_0t,\psi,\psi')$ for each τ.

Symmetries. (Cf. [S], Theorem 3.2.) First, we observe that the above sums have the following symmetries. Let ι and ω be, respectively, the automorphism and anti-automorphism of G given by

$$\iota(g) = w_0{}^tg^{-1}w_0 \quad \text{and} \quad \omega(g) = w_0{}^tgw_0.$$

Note that ι and ω are of order 2 and they preserve the subgroups $G(\mathbb{Z}_p)$, U, $U(\mathbb{Z}_p)$, B, T, and $N_G(T)$. Therefore ι and ω induce transformations of W (also denoted by ι and ω). One checks that

$$\operatorname{diag}(t_1,t_2,t_3) = \iota(\operatorname{diag}(1/t_3,1/t_2,1/t_1)) = \omega(\operatorname{diag}(t_3,t_2,t_1)); \tag{2.1}$$

$$\psi_{\nu_1,\nu_2} = \psi_{\nu_2,\nu_1} \circ \iota = \psi_{-\nu_2,-\nu_1} \circ \omega; \tag{2.2}$$

$$\begin{aligned} \mathrm{Kl}_\tau(w_0t,\psi,\psi') &= \mathrm{Kl}_{\iota(\tau)}(w_0\iota(t),\psi\circ\iota,\psi'\circ\iota) \\ &= \mathrm{Kl}_{\omega(\tau)}(w_0t,\psi'\circ\omega,\psi\circ\omega); \end{aligned} \tag{2.3}$$

$$\mathrm{Kl}_\tau(w_0t,\psi,\psi') = \mathrm{Kl}_\tau(w_0t\varepsilon,\psi,\psi'_\varepsilon) = \mathrm{Kl}_\tau(w_0t\iota(\varepsilon),\psi_\varepsilon,\psi'), \tag{2.4}$$

where $\varepsilon \in T(\mathbb{Z}_p)$ and $\psi_\varepsilon(u) = \psi(\varepsilon u\varepsilon^{-1})$. Formulae (2.1) through (2.4) reduce the problem of calculating $\mathrm{Kl}_\tau(w_0t,\psi,\psi')$ to the case $t = \operatorname{diag}(p^s,p^{r-s},p^{-r})$

with $r \le s$; and we will be able to combine the cases $\tau = (12)$ and $\tau = (23)$.

Theorem 2.4. *Let*

$$t = \begin{pmatrix} p^s & & \\ & p^{r-s} & \\ & & p^{-r} \end{pmatrix}, \quad \psi = \psi_{\nu_1,\nu_2}, \quad \psi' = \psi_{\nu_1',\nu_2'}, \quad \tau \in W.$$

The partial Kloosterman sums $\mathrm{Kl}_\tau = \mathrm{Kl}_\tau(w_0 t, \psi, \psi')$ *and the sizes of the Kloosterman sets* $X_\tau = X_\tau(w_0 t)$ *are given by the following formulae*:

$$\mathrm{Kl}_{w_0} = |X_{w_0}| = \delta_{r=s=0};$$

$$\mathrm{Kl}_{(123)} = \delta_{s>r=0} S(\nu_2, \nu_1'; p^s), \quad |X_{(123)}| = \delta_{s>r=0} p^s(1-1/p);$$

$$\mathrm{Kl}_{(132)} = \delta_{r>s=0} S(\nu_1, \nu_2'; p^r), \quad |X_{(132)}| = \delta_{r>s=0} p^r(1-1/p);$$

$$\mathrm{Kl}_{(12)} = \delta_{r,s>0} S(p^s\nu_1, \nu_2'; p^r) S(\nu_2, p^r\nu_1'; p^s),$$

$$|X_{(12)}| = \delta_{r,s>0} p^{r+s}(1-1/p)^2;$$

$$\mathrm{Kl}_{(23)} = \delta_{r,s>0} S(\nu_1, p^s\nu_2'; p^r) S(p^r\nu_2, \nu_1'; p^s),$$

$$|X_{(23)}| = \delta_{r,s>0} p^{r+s}(1-1/p)^2;$$

$$\mathrm{Kl}_e = \delta_{r>0} \sum_{\substack{1\le\alpha,\beta\le r\\ \alpha+\beta\ge r}} p^{-(\alpha+\beta)}(1-\delta_{\alpha=r}/p)^{-1}(1-\delta_{\beta=r}/p)^{-1}$$

$$\times \sum_{\substack{A\in R_r^\times\\ p\nmid p^{\alpha+\beta-r}A-p^{s-r}}} S(p^\alpha\nu_1 A, p^\beta\nu_2'; p^r) S\left(p^\beta\nu_2, p^\alpha\nu_1' \frac{A}{p^{\alpha+\beta-r}A - p^{s-r}}; p^s\right),$$

$$|X_e| = (r-1)p^{r+s}(1-1/p)^3 + \delta_{r=s} p^{2s-1}(1-1/p).$$

In the formula for Kl_e, *we assume* $r \le s$. *If* $r > s$ *then switch* $r \leftrightarrow s$, $\nu_1 \leftrightarrow \nu_2$, $\nu_1' \leftrightarrow \nu_2'$.

Remarks 2.5. (1) The formulae for $|X_\tau(w_0t)|$ follow from those for $\mathrm{Kl}_\tau(w_0t, \psi, \psi')$ by taking $\nu_1 = \nu_2 = \nu_1' = \nu_2' = 0$. Note that the outer sum in Kl_e is empty if $r = 0$ and the inner sum is empty if $r < s$ and $\alpha + \beta > r$. We will give a more explicit version of the above formula for $\mathrm{Kl}(w_0t, \psi, \psi')$ when r, $s > 0$ in Theorem 2.11.

(2) Stevens gives equivalent results in [S, (5.10) and (5.11)]. With $n = w_0t$, Stevens's $S_{a,b}(n, \psi, \psi')$ is the term $\alpha = s - a$, $\beta = r - b$ in the sum for Kl_e if $a < s$ and $b < r$; if $a = s$ and $b < r$ (so $b = \max\{0, r-s\}$) then $S_{a,b}(n, \psi, \psi') = \mathrm{Kl}_{(23)}$; if $a < s$ and $b = r$ (so $a = \max\{0, s-r\}$) then $S_{a,b}(n, \psi, \psi') = \mathrm{Kl}_{(12)}$; and if $a = s$ and $b = r$ then $S_{a,b}(n, \psi, \psi') = \mathrm{Kl}_{(123)}$, $\mathrm{Kl}_{(132)}$, or Kl_{w_0}, depending on whether $r = 0$ or $s = 0$. To derive our

formulae from Stevens's, one must carefully count the orbits of the $T(\mathbb{Z}_p)$-action. We prefer to avoid the $T(\mathbb{Z}_p)$-action entirely; besides, we believe that looking at the Iwahori cells will be useful when considering $\mathrm{GL}(N)$ with $N > 3$.

The next lemma will allow us to express the Kloosterman set $X_\tau(w_0t)$ as a quotient of an algebraic subset $Y_\tau(w_0t)$ of $R_m^6 = (\mathbb{Z}/p^m\mathbb{Z})^6$, with $m = \max\{r, s\}$.

LEMMA 2.6. *Let $C \subseteq U$. Assume that $U(\mathbb{Z}_p)C = C$ and that there are integers i_1, i_2, i_3, m with $0 \le i_1, i_2, i_3 \le m$ and $i_3 \ge i_1 + i_2$ such that $p^{i_1}x_1$, $p^{i_2}x_2$, $p^{i_3}x_3 \in \mathbb{Z}_p$ whenever $u = u(x_1, x_2, x_3) \in C$. Let $\Phi : C \to R_m^3$ be defined by $\Phi(u) = (p^{i_1}x_1, p^{i_2}x_2, p^{i_3}x_3) \pmod{p^m}$. Then $\Phi(C) \cong H\backslash C$, where $H \lhd U(\mathbb{Z}_p)$ is defined by*

$$H = \{u(x_1, x_2, x_3) : p^{i_1}x_1, p^{i_2}x_2, p^{i_3}x_3 \in p^m\mathbb{Z}_p\}.$$

Furthermore, the fibers of $H\backslash C \to U(\mathbb{Z}_p)\backslash C$ each have $(U(\mathbb{Z}_p){:}\, H)$ elements and $(U(\mathbb{Z}_p){:}\, H) = |\Phi(U(\mathbb{Z}_p))| = p^{3m-(i_1+i_2+i_3)}$.

Proof. Left to the reader. ∎

Proof of Theorem 2.4. Let $A = (a_{ij}) = uw_0tu' \in G(\mathbb{Z}_p)$, with $u = u(x_1, x_2, x_3)$, $u' = u(x_1', x_2', x_3')$, and $t = \mathrm{diag}(p^s, p^{r-s}, p^{-r})$. Then

$$A = \begin{pmatrix} p^sx_3 & p^sx_3x_1' - p^{r-s}x_1 & p^sx_3x_3' - p^{r-s}x_1x_2' + p^{-r} \\ p^sx_2 & p^sx_2x_1' - p^{r-s} & p^sx_2x_3' - p^{r-s}x_2' \\ p^s & p^sx_1' & p^sx_3' \end{pmatrix}. \tag{2.5}$$

Let $B(\tau)$ be the Iwahori cell containing A, so that $A \in C_\tau(w_0t)$. The numbers that, according to Lemma 2.2, determine τ are $a_{31} = p^s$, $a_{21} = p^sx_2$, $a_{32} = p^sx_1'$, and $A_{13} = p^r$. It is not hard to see that $A \in G(\mathbb{Z}_p)$ is equivalent to $r, s \ge 0$,

$$p^rx_1,\ p^sx_2,\ p^sx_3,\ p^sx_1',\ p^rx_2',\ p^sx_3' \in \mathbb{Z}_p, \tag{2.6}$$

and

$$\begin{aligned} p^sx_3 \cdot p^sx_1' &\equiv p^rx_1 \pmod{p^s}, \\ p^sx_2 \cdot p^sx_1' &\equiv p^r \pmod{p^s}, \\ p^r \cdot p^sx_3 \cdot p^sx_3' - p^rx_1 \cdot p^rx_2' + p^s &\equiv 0 \pmod{p^{r+s}}, \\ p^sx_2 \cdot p^sx_3' &\equiv p^rx_2' \pmod{p^s}. \end{aligned} \tag{2.7}$$

Case $\tau = w_0$. By Lemma 2.2 and (2.5), $r = s = 0$. Then (2.6) yields u, $u' \in U(\mathbb{Z}_p)$. Thus $X_\tau(w_0t)$ consists of only one double coset and $\mathrm{Kl}_\tau(w_0t, \psi, \psi') = 1$.

Case $\tau = (123)$. By Lemma 2.2 and (2.5), $s > r = 0$. Let

$$\begin{aligned} &\Phi : C_\tau(w_0t) \to Y_\tau(w_0t) \subseteq R_s^6, \\ &u(x)w_0tu(x') \mapsto (\overline{y}, \overline{y}') = (x_1, p^sx_2, p^sx_3, p^sx_1', x_2', p^sx_3') \pmod{p^s}. \end{aligned}$$

One can check that $Y_\tau(w_0t)$, the image of Φ, is given by

$$Y_\tau(w_0t) = \{(\overline{y},\overline{y}') \in R_s^6 : \overline{y}_2\overline{y}'_1 = 1,\ \overline{y}_3\overline{y}'_1 = \overline{y}_1,\ \overline{y}_2\overline{y}'_3 = \overline{y}'_2\}.$$

By Lemma 2.6, $Y_\tau(w_0t)$ is an N-to-1 cover of $X_\tau(w_0t)$, with $N = p^{2s}$. Thus

$$\begin{aligned}\mathrm{Kl}_\tau(w_0t,\psi,\psi') &= \frac{1}{p^{2s}} \sum_{(\overline{y},\overline{y}')\in Y_\tau(w_0t)} e\biggl(\nu_1\overline{y}_1 + \frac{\nu_2\overline{y}_2}{p^s} + \frac{\nu'_1\overline{y}'_1}{p^s} + \nu'_2\overline{y}'_2\biggr)\\ &= \sum_{\overline{y}_2\overline{y}'_1=1} e_s(\nu_2\overline{y}_2 + \nu'_1\overline{y}'_1) = S(\nu_2,\nu'_1;p^s).\end{aligned}$$

Case $\tau = (132)$. We reduce this to the previous case, using (2.1)–(2.3).

Case $\tau = (12)$. By Lemma 2.2 and (2.5), $r > 0$, $s > 0$, $p^sx_2 \in \mathbb{Z}_p^\times$, and $p^sx'_1 \in p\mathbb{Z}_p$; it follows that $p^sx'_3 \in \mathbb{Z}_p^\times$. By (2.1)–(2.3) we may assume $r \le s$. Then (2.7) shows that x_1, $p^{s-r}x'_1 \in \mathbb{Z}_p$. Thus we let

$$\begin{aligned}&\Phi : C_\tau(w_0t) \to Y_\tau(w_0t) \subseteq R_s^6,\\ &u(x)w_0tu(x') \mapsto (\overline{y},\overline{y}') = (x_1, p^sx_2, p^sx_3, p^{s-r}x'_1, p^rx'_2, p^sx'_3) \pmod{p^s}.\end{aligned}$$

One can check that $Y_\tau(w_0t)$, the image of Φ, is given by

$$\begin{aligned}Y_\tau(w_0t) = \{(\overline{y},\overline{y}') \in R_s^6 : \overline{y}_2,\overline{y}'_2 \in R_s^\times, p^r(\overline{y}_2\overline{y}'_1 - 1) = 0,&\\ \overline{y}_2\overline{y}'_3 = \overline{y}'_2, \overline{y}_3\overline{y}'_3 = \overline{y}_1\overline{y}'_2 - p^{s-r}\}.&\end{aligned}$$

By Lemma 2.6, $Y_\tau(w_0t)$ is an N-to-1 cover of $X_\tau(w_0t)$, with $N = p^{2s}$. If $\overline{y}_1$, $\overline{y}'_1 \in R_s$, $\overline{y}_2$, $\overline{y}'_2 \in R_s^\times$ are given with $\overline{y}_2\overline{y}'_1 \equiv 1 \pmod{p^{s-r}}$ then $\overline{y}_3$ and $\overline{y}'_3$ are determined and so

$$\begin{aligned}&\mathrm{Kl}_\tau(w_0t,\psi,\psi')\\ &\quad= \frac{1}{p^{2s}} \sum_{\overline{y}_1\in R_s} e(\nu_1\overline{y}_1) \sum_{\overline{y}'_2\in R_s^\times} e\biggl(\frac{\nu'_2\overline{y}'_2}{p^r}\biggr) \sum_{\substack{\overline{y}_2\in R_s^\times,\overline{y}'_1\in R_s\\ \overline{y}_2\overline{y}'_1\equiv 1 \,(\mathrm{mod}\, p^{s-r})}} e\biggl(\frac{\nu_2\overline{y}_2}{p^s} + \frac{\nu'_1\overline{y}'_1}{p^{s-r}}\biggr)\\ &\quad= p^{-2s}\cdot p^s \cdot p^{s-r}S(0,\nu'_2;p^r)\cdot p^rS(\nu_2,p^r\nu'_1;p^s).\end{aligned}$$

Case $\tau = (23)$. We reduce this to the previous case, using (2.1)–(2.3).

Case $\tau = e$. By Lemma 2.2 and (2.5), $r > 0$, $s > 0$, and p^sx_2, $p^sx'_1 \in p\mathbb{Z}_p$; it follows that p^sx_3, $p^sx'_3 \in \mathbb{Z}_p^\times$. Let $\alpha = \min\{v_p(p^sx'_1), s\}$, $\beta = \min\{v_p(p^sx_2), s\}$. By (2.1)–(2.3) we may assume $r \le s$ and $\alpha \le \beta$. Then (2.7) shows that $\alpha+\beta \ge r$, with equality if $r < s$; and $\min\{v_p(p^rx_1), s\} = \alpha$, $\min\{v_p(p^rx'_2), s\} = \beta$. We let $C_{\alpha,\beta}(w_0t) \subseteq C_e(w_0t)$ be the set of all matrices with these properties:

$$\begin{aligned}C_{\alpha,\beta}(w_0t) = \{(a_{ij}) \in C_e(w_0t) : \min\{v_p(a_{32}), s\} = \alpha, \min\{v_p(a_{21}), s\} = \beta\};&\\ C_e(w_0t) = \bigsqcup_{1\le\alpha,\beta\le s} C_{\alpha,\beta}(w_0t);&\end{aligned}$$

and we define

$$\begin{aligned}&\varPhi : C_{\alpha,\beta}(w_0t) \to Y_{\alpha,\beta}(w_0t) \subseteq R_s^6,\\ &u(x)w_0tu(x') \mapsto (\overline{y},\overline{y}')\\ &\qquad\qquad = (p^{r-\alpha}x_1, p^{s-\beta}x_2, p^sx_3, p^{s-\alpha}x_1', p^{r-\beta}x_2', p^sx_3') \pmod{p^s}.\end{aligned}$$

One can check that $Y_{\alpha,\beta}(w_0t)$, the image of $\varPhi$, is given by

$$\begin{aligned}\{(\overline{y},\overline{y}') \in R_s^6 : {}&\overline{y}_3,\overline{y}_3' \in R_s^\times, \alpha = s \text{ or } \overline{y}_1,\overline{y}_1' \in R_s^\times, \beta = s \text{ or } \overline{y}_2,\overline{y}_2' \in R_s^\times,\\ &p^\alpha(\overline{y}_3\overline{y}_1' - \overline{y}_1) = 0,\ p^\beta(\overline{y}_2\overline{y}_3' - \overline{y}_2') = 0,\ \overline{y}_3\overline{y}_3' - p^{\alpha+\beta-r}\overline{y}_1\overline{y}_2' + p^{s-r} = 0\}.\end{aligned}$$

Let $X_{\alpha,\beta}(w_0t) = U(\mathbb{Z}_p)\backslash C_{\alpha,\beta}(w_0t)/U(\mathbb{Z}_p)$, and let $\mathrm{Kl}_{\alpha,\beta}(w_0t,\psi,\psi')$ denote the sum over $X_{\alpha,\beta}(w_0t)$. By Lemma 2.6, $Y_{\alpha,\beta}(w_0t)$ is an N-to-1 cover of $X_{\alpha,\beta}(w_0t)$ with $N = p^{2(s-r+\alpha+\beta)}$. Given $(\overline{y},\overline{y}') \in Y_{\alpha,\beta}(w_0t)$, let $A = \overline{y}_1\overline{y}_2'$; then one can check that $p^{\alpha+\beta-r}A - p^{s-r} \in R_s^\times$ and $\overline{y}_2\overline{y}_1' = A(p^{\alpha+\beta-r}A - p^{s-r})^{-1}$. Conversely, given these relations there are $p^\alpha(1-1/p)^{\delta_{\alpha=s}}$ pairs $(\overline{y}_3,\overline{y}_3')$ that satisfy $(\overline{y},\overline{y}') \in Y_{\alpha,\beta}(w_0t)$.

First assume $\beta < s$. Then A, $\overline{y}_1$, and $\overline{y}_1'$ determine $\overline{y}_2'$ and p^β choices for $\overline{y}_2$, so

$$\begin{aligned}\mathrm{Kl}_{\alpha,\beta}(w_0t,\psi,\psi') = \frac{1}{N}p^\alpha \sum_{\substack{A\in R_s^\times\\ p\nmid p^{\alpha+\beta-r}A-p^{s-r}}} \sum_{\overline{y}_1\in R_s^\times} e\biggl(\frac{\nu_1\overline{y}_1}{p^{r-\alpha}} + \frac{\nu_2'A\overline{y}_1^{-1}}{p^{r-\beta}}\biggr)\\ \times \sum_{\overline{y}_1'\in R_s^\times} p^\beta e\biggl(\frac{\nu_1'\overline{y}_1'}{p^{s-\alpha}} + \frac{\nu_2'}{p^{s-\beta}}\cdot\frac{A}{p^{\alpha+\beta-r}A-p^{s-r}}\overline{y}_1'^{-1}\biggr).\end{aligned}$$

Now $N = p^{2(s-r+\alpha+\beta)}$, the sum over $\overline{y}_1$ gives $p^{s-r}S(p^\alpha\nu_1, p^\beta\nu_2'A; p^r)$, the sum over $\overline{y}_1'$ gives $p^\beta S\bigl(p^\alpha\nu_1', p^\beta\nu_2\frac{A}{p^{\alpha+\beta-r}A-p^{s-r}}; p^s\bigr)$, and the value of A only matters $\pmod{p^r}$, so we get

$$\begin{aligned}&\mathrm{Kl}_{\alpha,\beta}(w_0t,\psi,\psi')\\ &= p^{-(\alpha+\beta)} \sum_{\substack{A\in R_r^\times\\ p\nmid p^{\alpha+\beta-r}A-p^{s-r}}} S(p^\alpha\nu_1, p^\beta\nu_2'A; p^r)S\biggl(p^\alpha\nu_1', \frac{p^\beta\nu_2A}{p^{\alpha+\beta-r}A-p^{s-r}}; p^s\biggr).\end{aligned}$$

Now suppose $\beta = s$, which implies $r = s$. If $\alpha < s$ then the above terms are independent of A but we lose the restriction that A be a unit, so we add a factor of $(1-1/p)^{-1}$. If $\alpha = \beta = s$ then we also lose the restriction that $\overline{y}_1$, $\overline{y}_1'$ be units, so we add two more factors of $(1-1/p)^{-1}$, one of which is canceled by the requirement that $\overline{y}_3$ be a unit. In all cases, the correct factor is $(1-1/p)^{-(\delta_{\alpha=s}+\delta_{\beta=s})}$. ■

Next we will give a more explicit expression for $\mathrm{Kl}(w_0t,\psi,\psi')$. To do this we have to calculate certain sums of products of the classical Kloosterman sums. We will deal with this problem first, and then we will apply the results to $\mathrm{Kl}(w_0 t, \psi, \psi')$. We begin by recalling some basic properties of the classical Kloosterman sums:

Classical Kloosterman sums. Let ν_1, ν_2 be p-adic integers. Then the classical Kloosterman sums are given by

$$S(\nu_1, \nu_2; p^m) = \begin{cases} \displaystyle\sum_{\substack{x,y \in R_m \\ xy \equiv 1 \ (\mathrm{mod}\, p^m)}} e_m(\nu_1 x + \nu_2 y) & \text{if } m \geq 1, \\ 1 & \text{if } m = 0. \end{cases} \tag{2.8}$$

For any unit x,

$$S(\nu_1 x, \nu_2; p^m) = S(\nu_1, \nu_2 x; p^m) = S(\nu_2, \nu_1 x; p^m). \tag{2.9}$$

Suppose that ν_1 and ν_2 are units. Then for any non-negative integers N_1 and N_2,

$$\begin{aligned} &S(p^{N_1}\nu_1, p^{N_2}\nu_2; p^m) \\ &\quad = p^{N_1} \delta_{N_1 = N_2 \leq m-1} S(\nu_1, \nu_2; p^{m-N_1}) + p^m \Delta(N_1, N_2; m) \end{aligned} \tag{2.10}$$

where Δ is defined in Notation 2.7 below. Observe that, in this formula for the Kloosterman sum, only one of the two terms can be non-zero; the possible values of Δ are 0, $-1/p$, or $1 - 1/p$; and $S(\nu_1, \nu_2; p^{m-N_1}) = S(1, \nu_1\nu_2; p^{m-N_1})$.

Notation 2.7. Let

$$\begin{aligned} \Delta(N_1, N_2; m) &= \left(1 - \frac{1}{p}\right)\delta_{\min\{N_1,N_2\} \geq m} - \frac{1}{p}\delta_{m-1 = \min\{N_1,N_2\} < \max\{N_1,N_2\}} \\ &= \delta_{\min\{N_1,N_2\} \geq m} - \frac{1}{p}(\delta_{\min\{N_1,N_2\} \geq m-1} - \delta_{m-1 = N_1 = N_2}). \end{aligned}$$

Fix $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, where a, b, c, $d \in \mathbb{Z}_p$, and assume that neither of the rows of γ is divisible by p. Set $\delta = v_p(\det \gamma)$. Let $n \geq 1$ and $1 \leq m \leq n + \delta$ be integers. Then $\gamma(x) = (ax + b)/(cx + d)$ gives a well defined map from $\{x \in R_n : p \nmid cx + d\}$ to R_m. If c is a unit then it is convenient to extend γ to all of R_n, by setting

$$\gamma(x) = a/c \qquad \text{if } p \mid cx + d.$$

Finally, let $\varrho : R_n \to \mathbb{F}_p$ denote the reduction map and, for any $X \subseteq \mathbb{F}_p$, set

$$P_X(\gamma; R_n) = \sum_{x \in R_n \setminus \varrho^{-1}(X)} S(1, x; p^n) S(1, \gamma(x); p^m).$$

If $X = \{x \in R_n : p \mid x(ax + b)(cx + d)\}$ then we will write simply $P(\gamma; R_n)$.

LEMMA 2.8. *Assume that* $m \le n+\delta$ *and that* $\{x \in \mathbb{F}_p : cx+d=0\} \subseteq X \subseteq \mathbb{F}_p$. *Then*

$$\sum_{x\in R_n\setminus\varrho^{-1}(X)} S(1,\gamma(x);p^m) = p^n\delta_{m\le\delta}S(1,\gamma(0);p^m) - p^{n-1}\delta_{m\le\delta+1}\sum_{x\in X} S(1,\gamma(x);p^m).$$

Proof. First, note that for any $x \in R_n$ satisfying $p\nmid cx+d$, and for any positive integer $r \le n$,

$$\gamma : x+p^rR_n \to \gamma(x)+p^{\delta+r}R_m$$

and all the fibers have $\min\{p^{n-r}, p^{n+\delta-m}\}$ elements. Therefore

$$\sum_{y\in x+p^rR_n} e_m(\gamma(y)) = p^{n-r}\delta_{m\le\delta+r}e_m(\gamma(x)). \tag{2.11}$$

Next, our extension of γ to R_n implies that, for $m \le \delta+1$,

$$\sum_{x\in\mathbb{F}_p} e_m(\gamma(x)) = p\,\delta_{m\le\delta}e_m(\gamma(0)). \tag{2.12}$$

We can now compute

$$\begin{aligned}\sum_{x\in R_n\setminus\varrho^{-1}(X)} & S(1,\gamma(x);p^m)\\ &= \sum_{t\in R_m^\times} e_m(1/t) \sum_{x\in R_n\setminus\varrho^{-1}(X)} e_m(t\gamma(x))\\ &= \sum_{t\in R_m^\times} e_m(1/t)p^{n-1}\delta_{m\le\delta+1} \sum_{x\in\mathbb{F}_p\setminus X} e_m(t\gamma(x))\\ &= p^{n-1}\delta_{m\le\delta+1}\sum_{t\in R_m^\times} e_m(1/t)\Big[p\delta_{m\le\delta}e_m(t\gamma(0)) - \sum_{x\in X} e_m(t\gamma(x))\Big]\\ &= p^n\delta_{m\le\delta}S(1,\gamma(0);p^m) - p^{n-1}\delta_{m\le\delta+1}\sum_{x\in X} S(1,\gamma(x);p^m). \quad\blacksquare\end{aligned}$$

In particular, if $1 \le m \le n$ then (taking $\gamma(x)=x$)

$$\sum_{x\in R_n\setminus\varrho^{-1}(X)} S(1,x;p^m) = -p^{n-1}\delta_{m=1}\sum_{x\in X} S(1,x;p). \tag{2.13}$$

PROPOSITION 2.9. *Assume that* $\{x\in\mathbb{F}_p : x(ax+b)(cx+d)=0\} \subseteq X \subseteq \mathbb{F}_p$ *and let* $m=n+\delta$. *Fix non-negative integers* $N_1 \le N_2$, $M_1 \le M_2$ *and let*

$$S = \sum_{x\in R_n\setminus\varrho^{-1}(X)} S(p^{N_1}, p^{N_2}x; p^n)S(p^{M_1}, p^{M_2}\gamma(x); p^m).$$

Then

$$S = S_1 + S_2 + S_3 + S_4,$$

where

$$\begin{aligned} S_1 &= p^{2n+m}(1-|X|/p)\varDelta(N_1,N_2;n)\varDelta(M_1,M_2;m),\\ S_2 &= p^{2n+M_1}\delta_{n-1\le M_1=M_2\le m-1}\varDelta(N_1,N_2;n)\\ &\qquad\times\Big[\delta_{M_1\ge n}S(1,\gamma(0);p^{m-M_1}) - \frac{1}{p}\sum_{x\in X}S(1,\gamma(x);p^{m-M_1})\Big],\\ S_3 &= -p^{2n-2}\delta_{N_1=N_2=n-1<M_2}S(p^{M_1},p^{M_2}\gamma(0);p^m)\sum_{x\in X}S(1,x;p),\\ S_4 &= \delta_{N_1=N_2=M_1=M_2\le n-1}p^{3N_1}P_X(\gamma;R_{n-N_1}). \end{aligned}$$

Remark 2.10. Note that in the formula for S, at most one term $S_i \ne 0$ for given values of N_i and M_i. By (2.9) the assumption $N_1 \le N_2$ and $M_1 \le M_2$ is not essential.

Proof of Proposition 2.9. If $S \ne 0$ then, by (2.10), we have to be in one of the following two cases (otherwise $S(p^{N_1}, p^{N_2}x; p^n) = 0$):

Case 1: $N_1 \ge n$, or $N_1 = n-1 < N_2$. Then (2.10) gives

$$\begin{aligned} S &= p^n\varDelta(N_1,N_2;n)\sum_{x\in R_n\setminus\varrho^{-1}(X)}S(p^{M_1},p^{M_2}\gamma(x);p^m)\\ &= p^{2n+m}\varDelta(N_1,N_2;n)(1-|X|/p)\varDelta(M_1,M_2;m)\\ &\quad + p^{n+M_1}\varDelta(N_1,N_2;n)\delta_{M_1=M_2\le m-1}\sum_{x\in R_n\setminus\varrho^{-1}(X)}S(1,\gamma(x);p^{m-M_1}). \end{aligned}$$

Now Lemma 2.8 shows that $S = S_1 + S_2$.

Case 2: $N_1 = N_2 \le n-1$. First, suppose that $M_2 > N_1$, so that $S(p^{M_1}, p^{M_2}\gamma(x); p^m)$ depends only on $x \pmod{p^{n-N_1-1}}$. If $N_1 = n-1$ then

$$S = S(p^{M_1},p^{M_2}\gamma(0);p^m)\cdot p^{N_1}\sum_{x\in\mathbb{F}_p\setminus X}p^{N_1}S(1,x;p),$$

so (2.13) shows that $S = S_3$. If $N_1 < n-1$ then

$$\begin{aligned} S = &\sum_{x\in R_{n-N_1-1}\setminus\varrho^{-1}(X)}S(p^{M_1},p^{M_2}\gamma(x);p^m)\\ &\times\sum_{y\in R_{N_1+1}}p^{N_1}S(1,x+p^{n-N_1-1}y;p^{n-N_1}), \end{aligned}$$

where we choose a representative $x \in R_n$ for the inner sum. The sum over y vanishes by Lemma 2.8.

Finally, suppose $M_2 \le N_1$. According to (2.10), $S(p^{M_1}, p^{M_2}\gamma(x); p^m) = 0$ unless $M_1 = M_2$, in which case

$$\begin{aligned} S &= p^{N_1+M_1} \sum_{x \in R_n \setminus \varrho^{-1}(X)} S(1, x; p^{n-N_1}) S(1, \gamma(x); p^{m-M_1}) \\ &= p^{N_1+M_1} \sum_{x \in R_{n-N_1} \setminus \varrho^{-1}(X)} S(1, x; p^{n-N_1}) \\ &\quad \times \sum_{y \in R_{N_1}} S(1, \gamma(x + p^{n-N_1} y); p^{m-M_1}). \end{aligned}$$

By (2.11), one sees that the inner sum is $p^{N_1}\delta_{N_1 \le M_1} S(1, \gamma(x); p^{m-M_1})$, which implies $S = p^{3N_1}\delta_{M_1=N_1} P_X(\gamma; R_{n-N_1}) = S_4$. ■

Theorem 2.11. *Let*

$$t = \begin{pmatrix} p^s & & \\ & p^{r-s} & \\ & & p^{-r} \end{pmatrix},$$

with $s \ge r > 0$. (*If* $r > s$ *then apply* (2.1)–(2.3).) *Also let* $\psi = \psi_{\nu_1,\nu_2}$, $\psi' = \psi_{\nu_1',\nu_2'}$, $N_1 = v_p(\nu_1)$, $M_2 = v_p(\nu_2)$, $M_1 = v_p(\nu_1')$, $N_2 = v_p(\nu_2')$,

$$n_1 = \min\{N_1, M_1 - (s-r)\}, \qquad n_2 = \min\{N_2, M_2 - (s-r)\},$$
$$\gamma_m = \begin{pmatrix} \nu_1'\nu_2/p^{M_1+M_2} & 0 \\ p^{(r-n_1-n_2)-2m} & -p^{s-r}\nu_1\nu_2'/p^{N_1+N_2} \end{pmatrix},$$

and use Notation 2.7.

(a) *If* $s = r \ge n_1 + n_2 + 2$ *then* $\mathrm{Kl}(w_0 t, \psi, \psi')$ *is given by*

$$\begin{aligned} p^{r+n_1+n_2}\Bigg[\frac{1}{p} + 1 + \sum_{1 \le m \le (r-n_1-n_2)/2} p^{-m} P(\gamma_m; R_m)\Bigg]\delta_{N_1=M_1}\delta_{M_2=N_2} \\ - p^{r+n_1+n_2}\delta_{r=n_1+n_2+2}\Bigg[\frac{1}{p}(\delta_{N_1<M_1}\delta_{M_2<N_2} + \delta_{N_1>M_1}\delta_{M_2>N_2}) \\ + S(1, p^{-M_2-M_1}\nu_2\nu_1'; p)\bigg\{\bigg(1 - \frac{1}{p}\bigg)\delta_{N_1>M_1}\delta_{M_2<N_2} \\ - \frac{1}{p}(\delta_{N_1=M_1}\delta_{M_2<N_2} + \delta_{N_1>M_1}\delta_{M_2=N_2})\bigg\} \\ + S(1, p^{-N_1-N_2}\nu_1\nu_2'; p)\bigg\{\bigg(1 - \frac{1}{p}\bigg)\delta_{N_1<M_1}\delta_{M_2>N_2} \\ - \frac{1}{p}(\delta_{N_1=M_1}\delta_{M_2>N_2} + \delta_{N_1<M_1}\delta_{M_2=N_2})\bigg\}\Bigg]. \end{aligned}$$

(b) *If* $s = r \le n_1 + n_2 + 1$ *then* $\mathrm{Kl}(w_0t, \psi, \psi')$ *is given by*

$$p^{2r}\left(1-\frac{1}{p}\right)^3 \min\{r-1, n_1, n_2, n_1+n_2-(r-1)\}$$
$$+\, p^{2r}\left(1-\frac{1}{p}\right)\left(\delta_{r\le n_1} + \delta_{r\le n_2} - \frac{1}{p}\right)$$
$$+\, p^{2r-1}\left[\left(\delta_{r\ge n_1+1} - \frac{1}{p}\delta_{r\ge n_1+2}\right)\delta_{N_1=M_1} + \left(\delta_{r\ge n_2+1} - \frac{1}{p}\delta_{r\ge n_2+2}\right)\delta_{M_2=N_2}\right]$$
$$+\, p^{2r-2}\left(1-\frac{1}{p}\right)(\delta_{r\ge n_1+2} + \delta_{r\ge n_2+2}).$$

(c) *If* $s > r \ge N_1 + N_2 + 2$ *then* $\mathrm{Kl}(w_0t, \psi, \psi')$ *is given by*

$$p^{(r+3N_1+3N_2)/2}P(\gamma; R_{(r-N_1-N_2)/2})\delta_{N_1=M_1}\delta_{M_2=N_2}\delta_{r\equiv N_1+N_2\ (\mathrm{mod}\ 2)}$$
$$+\, p^{N_1+N_2}S(p^{N_2+1}\nu_1', p^{N_1+1}\nu_2; p^s)\delta_{N_1<M_1}\delta_{M_2>N_2}\delta_{r=N_1+N_2+2}.$$

(d) *If* $r < s$ *and* $r \le N_1 + N_2 + 1$ *then* $\mathrm{Kl}(w_0t, \psi, \psi')$ *is given by*

$$S(0, \nu_2'; p^r)S(\nu_2, p^r\nu_1'; p^s) + S(\nu_1, 0; p^r)S(p^r\nu_2, \nu_1'; p^s)$$
$$+\, p^r\Delta\left(\frac{M_1+M_2+r}{2}, r; r\right)\Delta\left(N_1+M_2, M_1+N_2; \frac{r+M_1+M_2}{2}\right)$$
$$\times\, S(p^{(r-M_1+M_2)/2}\nu_1', p^{(r+M_1-M_2)/2}\nu_2; p^s)$$
$$\times\, \delta_{r\ge|M_1-M_2|+2}\delta_{r\equiv M_1+M_2\ (\mathrm{mod}\ 2)}\delta_{2s\ge r+M_1+M_2+2}$$
$$+\, p^{r+s}\left(1-\frac{1}{p}\right)\left[\left(1-\frac{1}{p}\right)^2 \min\{r-1, n_1, n_2, n_1+n_2-(r-1)\}\right.$$
$$\left. -\,\frac{1}{p}\left(\delta_{N_1\ne M_1-(s-r)} - \frac{1}{p}\right)\delta_{r\ge n_1+2} - \frac{1}{p}\left(\delta_{M_2-(s-r)\ne N_2} - \frac{1}{p}\right)\delta_{r\ge n_2+2}\right]$$
$$\times\, \delta_{r\le n_1+n_2+1}\delta_{n_1,n_2\ge 0}$$
$$+\, p^{r+s-2}\left(1-\frac{1}{p}\right)[\delta_{M_1-N_1<s-r<M_2-N_2}\delta_{s=M_1+N_2+2}$$
$$+\, \delta_{M_1-N_1>s-r>M_2-N_2}\delta_{s=N_1+M_2+2}]\delta_{n_1,n_2\ge 0}.$$

Proof. In the notation of Theorem 2.4, $\mathrm{Kl}(w_0t, \psi, \psi') = \mathrm{Kl}_{(12)} + \mathrm{Kl}_{(23)} + \mathrm{Kl}_e$ since we are assuming r, $s \ne 0$. If we let $x = \nu_1\nu_2'A/p^{N_1+N_2}$ and $\gamma_m = \left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)$, with $m = r - (\alpha+\beta+n_1+n_2)/2$, then the inner sum in the expression for Kl_e becomes

$$S^{\alpha,\beta} = \sum_{\substack{x\in R_r^\times \\ p\nmid(ax+b)(cx+d)}} S(p^{\alpha+N_1}, p^{\beta+N_2}x; p^r)S(p^{\alpha+M_1}, p^{\beta+M_2}\gamma(x); p^s),$$

thanks to (2.9). Applying Proposition 2.9, with $X = \{x \in \mathbb{F}_p : x(ax+b) \times (cx+d) = 0\}$, we get $S^{\alpha,\beta} = S_1^{\alpha,\beta} + S_2^{\alpha,\beta} + S_3^{\alpha,\beta} + S_4^{\alpha,\beta}$; one must sum over α and β. For $i = 2$ and 3 there is at most one pair (α, β) for which $S_i^{\alpha,\beta} \neq 0$ (cf. Remark 2.5(1)) so these terms are easy to sum. The case $i = 4$ is not hard but the case $i = 1$ is a bit of a chore. When $r = s$ it is convenient to note that $S_1^{\alpha,\beta} = 0$ unless $0 \le r - \alpha \le n_1 + 1$ and $0 \le r - \beta \le n_2 + 1$; one can sum over $\overline{\alpha} = r - \alpha$ and $\overline{\beta} = r - \beta$ and add a factor of $\delta_{\overline{\alpha}+\overline{\beta} \le r}$ to recover the original bound, $\alpha + \beta \ge r$. (It helps to note that $S_1^{r,0} = \mathrm{Kl}_{(12)}$ and $S_1^{0,r} = \mathrm{Kl}_{(23)}$.) When $r < s$ we have $\alpha + \beta = r$ and it seems simplest to count the number of pairs (α, β) giving each of the four possible values of $S_1^{\alpha,\beta}$. ∎

3. Sums of products of Kloosterman sums. In this section we consider the exponential sums appearing in the expressions for the GL(3)-Kloosterman sums in Theorem 2.11. Using the stationary phase method of Section 1 and l-adic cohomology, we can estimate these sums. At the end of the section, we derive our final estimates for the GL(3)-Kloosterman sums.

Notation 3.1. Fix a prime p. As in Section 2, let $R_m = \mathbb{Z}/p^m\mathbb{Z} = \mathbb{Z}_p/p^m\mathbb{Z}_p$, let v_p denote the valuation on $\mathbb{Q}_p$, and let

$$e_m(x) = e^{2\pi i x/p^m}$$

whenever this makes sense: $x \in \mathbb{C}$ or $\mathbb{Z}_p$ or R_m. In this section, we will work exclusively with

$$K(\nu; R_m) = S(1, \nu; p^m) = \sum_{x \in R_m^\times} e_m\left(x + \frac{\nu}{x}\right) \tag{3.1}$$

where ν is a unit in $\mathbb{Z}_p$.

As in Section 2, if $\mathcal{P}$ is some condition then let

$$\delta_{\mathcal{P}} = \begin{cases} 1 & \text{if } \mathcal{P} \text{ holds,} \\ 0 & \text{otherwise.} \end{cases}$$

Fix a matrix $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, where $a, b, c, d \in \mathbb{Z}_p$. Assume that

$$p \nmid a \text{ or } p \nmid b \quad \text{and} \quad p \nmid c \text{ or } p \nmid d. \tag{3.2}$$

Let

$$ad - bc = \det \gamma = p^\delta u \qquad (u \in \mathbb{Z}_p^\times).$$

Fix m and define h by

$$h = \delta + v_p(2) + \min\{v_p(b), v_p(3c), m\}.$$

With this notation, $\gamma(x) = (ax+b)/(cx+d)$ gives a well-defined element of $R_{m+\delta}^\times$ if $x \in R_m$ and $p \nmid ax + b$, $cx + d$; and (0.4) becomes

$$P(\gamma; R_m) = \sum_{\substack{x\in R_m^\times \\ p\nmid ax+b,\, cx+d}} K(x; R_m)K(\gamma(x); R_{m+\delta}). \tag{3.3}$$

Outline. The goal of this section is to estimate the sums $P(\gamma; R_m)$. The sums that come up when evaluating GL(3)-Kloosterman sums all have $b = 0$, but we will only assume (in some cases) that $v_p(b) \neq v_p(3c)$.

We use different techniques in different cases. Proposition 3.3 deals with the case $c \equiv 0 \pmod{p^m}$: an elementary calculation expresses $P(\gamma; R_m)$ in terms of a classical Kloosterman sum. In the remaining cases, we assume $c \not\equiv 0 \pmod{p^m}$. Proposition 3.4 deals with the case $m = 1$, using the l-adic techniques developed by Deligne and Katz. Proposition 3.5 deals with the case $m > 1$, using Katz's principle of stationary phase, as described in Section 1. The following theorem summarizes our results, although Propositions 3.3–3.5 have more precise statements.

Theorem 3.2. *Use Notation* 3.1; *if* $m > 1$ *and* $m > v_p(c)$ *then assume that* $v_p(b) \neq v_p(3c)$. *Then*

$$|P(\gamma; R_m)| \le \sqrt{p}^{\,3m+h}(12\,\delta_{p>3} + 108\sqrt{3}\,\delta_{p=3} + 2^9\sqrt{2}\,\delta_{p=2}).$$

Furthermore, $P(\gamma; R_m) = 0$ *in the following situations* (*with some extra conditions if* $p = 2$ *or* 3):

(1) $v_p(c) > v_p(b) = 0$, $m + \delta > 1$, *and* $v_p(u^2 - ad^3) \neq 0$;

(2) $\delta > 0 = v_p(c)$ *and* a/c *not a square or* $\delta > 0 = v_p(b)$ *and* b/d *not a square*;

(3) $v = \min\{v_p(b), v_p(3c)\} > 0$ *and* $v_p(a - d) < \min\{v, m - 1\}$; *or* $0 < v < m - 1$ *and* $v_p(a - d) \neq v$.

Before considering $P(\gamma; R_m)$ we will recall the bounds on classical Kloosterman sums. Suppose that ν is a unit and that p is odd, $m > 1$, or $p = 2$, $m \ge 8$. According to Example 1.15 (or [Sa]),

$$K(\nu; R_m) = \sqrt{p}^{\,m} \sum_{\substack{\alpha^2=\nu \\ \alpha\in\mathbb{Z}_p^\times}} G_m(2\alpha)e_m(2\alpha), \tag{3.4}$$

where $G_m(2\alpha)$ is the normalized Gauss sum, as in Section 1. In particular, $K(\nu; p^m) = 0$ if ν is not a square; and

$$|K(\nu; p^m)| \le 2\sqrt{p}^{\,m+v_p(2)}. \tag{3.5}$$

According to Example 1.15, this bound holds for $p = 2$ and all $m \ge 2$; and the vanishing statement holds for $p = 2$ and $m \ge 6$. Finally, the Hasse–Weil estimate [W1] says that (3.5) holds, without the term $v_p(2)$, if $m = 1$.

Proposition 3.3. *Using Notation* 3.1, *assume that* $v_p(c) \ge m$. *Then* d *is a unit and*

$$P(\gamma; R_m) = p^m S\left(\frac{u - p^\delta d^2}{u}, \frac{b}{d}; p^{m+\delta}\right) + \delta_{m=1} \begin{cases} K(b/d; \mathbb{F}_p) + K(-b/a; \mathbb{F}_p) & \text{if } \delta = 0 = v_p(b), \\ -1 & \text{if } \delta = 0 < v_p(b), \\ K(b/d; R_{1+\delta}) & \text{if } \delta > 0 = v_p(b). \end{cases}$$

In particular,

$$|P(\gamma; R_m)| \le 4\sqrt{p}^{\,3m+h}.$$

Furthermore, the main term vanishes unless $v_p(b) = v_p(u - p^\delta d^2) = v_p(u^2 - ad^3)$ *or both* $v_p(b)$ *and* $v_p(u^2 - ad^3)$ *are at least* $m + \delta - 1$. *If* $\delta > 0$ (*and* $m + \delta \ge 6$ *if* $p = 2$) *then the main term vanishes unless* $v_p(b) = 0$ *and* b/d *is a square.*

Proof. The estimate and the vanishing follow from the formula for $P(\gamma; R_m)$, using (2.10) and (3.5). Since $c \equiv 0 \pmod{p^m}$, d is a unit by (3.2). Thus

$$\gamma(x) = \frac{ax + b}{cx + d} \equiv p^\delta \frac{u}{d^2} x + \frac{b}{d} \pmod{p^{m+\delta}},$$

so we may reduce to the case $c = 0$, $d = 1$, $\gamma(x) = ax + b$, so that $a = \det\gamma = p^\delta u$.

Putting the definition (3.1) of the Kloosterman sums into the definition (3.3) of $P(\gamma; R_m)$ and switching the order of summation, we find

$$\begin{aligned} (3.6)\quad P(\gamma; R_m) &= \sum_{\substack{s \in R_m^\times \\ t \in R_{m+\delta}^\times}} e_m(s^{-1}) e_{m+\delta}(t^{-1} + bt) \sum_{\substack{x \in R_m^\times \\ p \nmid ax+b}} e_m((s + ut)x) \\ &= \sum_{\substack{s \in R_m^\times \\ t \in R_{m+\delta}^\times}} e_m(s^{-1}) e_{m+\delta}(t^{-1} + bt) \sum_{x \in R_m} e_m((s + ut)x) \\ &\quad - \sum_{\substack{s \in R_m^\times \\ t \in R_{m+\delta}^\times}} e_m(s^{-1}) e_{m+\delta}(t^{-1} + bt) \sum_{\substack{x \in R_m \\ p \mid x(ax+b)}} e_m((s + ut)x). \end{aligned}$$

The inner sum in the first term in (3.6) gives p^m if $s + ut \equiv 0 \pmod{p^m}$ and vanishes otherwise. The first term in (3.6) is thus

$$p^m S\left(\frac{u - p^\delta}{u}, b; p^{m+\delta}\right),$$

the desired main term in $P(\gamma; R_m)$.

Now consider the second term in (3.6). If $p \mid x(ax+b)$ then $x = py$ or $py - b/a$ with $y \in R_{m-1}$; we only need the second possibility if $\delta = 0 = v_p(b)$ (i.e., both a and b are units, in which case $u = a$). Summing over y, we find

$$\sum_{\substack{x\in R_m\\ p\mid x(ax+b)}} e_m((s+ut)x)$$
$$= p^{m-1}\delta_{s+ut\equiv 0\ (\mathrm{mod}\ p^{m-1})}[1 + e_m((s+ut)(-b/a))\delta_{\delta=0=v_p(b)}].$$

Using this, the second term in (3.6) becomes

$$\begin{aligned}(3.7)\qquad -p^{m-1} &\sum_{\substack{s\in R_m^\times,t\in R_{m+\delta}^\times\\ s+ut\equiv 0\ (\mathrm{mod}\ p^{m-1})}} e_m\left(\frac{1}{s}\right)e_{m+\delta}\left(\frac{1}{t}+bt\right)\\ -p^{m-1}\delta_{\delta=0=v_p(b)} &\sum_{\substack{s,t\in R_m^\times\\ s+ut\equiv 0\ (\mathrm{mod}\ p^{m-1})}} e_m\left(\frac{1}{s}-\frac{b}{a}s\right)e_{m+\delta}\left(\frac{1}{t}\right)\end{aligned}$$

(where we have used $u = a$ if $\delta = 0 = v_p(b)$).

First suppose that $m = 1$, so that the condition $s + ut \equiv 0 \pmod{p^{m-1}}$ is automatic and the factor p^{m-1} is trivial. The sum over s in the first term of (3.7) and the sum over t in the second term each give -1; by (2.10), (3.7) gives the desired terms in $P(\gamma; R_m)$.

Now suppose that $m > 1$; we must show that (3.7) vanishes. In the first term of (3.7), let $s = -ut + p^{m-1}z$, with $z \in \mathbb{F}_p$; in the second term, let $t = -u^{-1}s + p^{m-1}z$. In both cases, the sum over z vanishes. ∎

PROPOSITION 3.4. *Using Notation* 3.1, *assume that* $m = 1$ *and* $p\nmid c$. *Then*

$$|P(\gamma;\mathbb{F}_p)| \le \sqrt{p}^{\,3+\delta}(12 - 2\,\delta_{\delta>0} - 4\,\delta_{p\mid d} - 4\,\delta_{\delta=0}\delta_{p\mid ab}).$$

Furthermore, *if* $\delta > 0$ ($\delta \ge 5$ *if* $p = 2$) *and* a/c *is not a square then* $P(\gamma;\mathbb{F}_p) = 0$.

We suspect that 12, the maximal value of the constant, is never best possible.

Proof. Choose an auxiliary prime $l \ne p$. Let $\mathcal{K}$ be the Kloosterman sheaf on $\mathbb{G}_m \otimes \mathbb{F}_p$, as in [K2] and [K3]. Thus $\mathcal{K}$ is a lisse $\overline{\mathbb{Q}}_l$-sheaf of rank 2 on $\mathbb{G}_m$ that is pure of weight 1, tame at 0, and totally wild with Swan conductor 1 at ∞; and for any $a \in \mathbb{F}_p^\times = \mathbb{G}_m(\mathbb{F}_p)$,

$$\mathrm{trace}(\mathrm{Frob}_a|\mathcal{K}) = -K(a;\mathbb{F}_p).$$

First consider the case $\delta = 0$. We have

$$\begin{aligned} P(\gamma;\mathbb{F}_p) &= \sum_{\substack{x\in\mathbb{F}_p^\times \\ p\nmid ax+b,\, cx+d}} K(x;\mathbb{F}_p)K(\gamma(x);\mathbb{F}_p) \\ &= \sum_{x\in U(\mathbb{F}_p)} \operatorname{trace}(\mathrm{Frob}_x|\mathcal{K}\otimes\gamma^*\mathcal{K}), \end{aligned}$$

where we let

$$U = \operatorname{Spec}\mathbb{F}_p[t, 1/t(at+b)(ct+d)]$$

(so that $U = \mathbb{G}_m \setminus \{-b/a, -d/c\}$ if a is a unit) and let γ denote the map $U \to \mathbb{G}_m$ induced by (the matrix) γ.

The sheaf

$$\mathcal{F} = \mathcal{K}\otimes\gamma^*\mathcal{K}$$

is lisse of rank 4 on U, pure of weight 2. We claim that $\mathcal{F}$ is geometrically irreducible. Indeed, $\mathcal{K}$ is totally wild at ∞, with Swan conductor 1, so it is irreducible as a representation of the wild inertia group P_∞. Similarly, $\gamma^*\mathcal{K}$ is irreducible as a representation of $P_{-d/c}$. Since $\mathcal{K}$ is tame at $-d/c$ it is trivial as a representation of $P_{-d/c}$; similarly, $\gamma^*\mathcal{K}$ is trivial as a representation of P_∞. Therefore $\mathcal{F}$ is irreducible as a representation of $P_\infty \times P_{-d/c}$ and, *a fortiori*, as a representation of $\pi_1(U\otimes\overline{\mathbb{F}}_p)$.

Since $\mathcal{F}$ is a lisse sheaf and U is affine, $H_c^0(U\otimes\overline{\mathbb{F}}_p,\mathcal{F}) = 0$. Since $\mathcal{F}$ is geometrically irreducible, $H_c^2(U\otimes\overline{\mathbb{F}}_p,\mathcal{F}) = 0$. Therefore, the Lefschetz Trace Formula gives simply

$$\begin{aligned} P(\gamma;\mathbb{F}_p) &= \sum_{i=0}^{2}(-1)^i \operatorname{trace}(\mathrm{Frob}_{\mathbb{F}_p}|H_c^i(U\otimes\overline{\mathbb{F}}_p,\mathcal{F})) \\ &= -\operatorname{trace}(\mathrm{Frob}_{\mathbb{F}_p}|H_c^1(\mathcal{F})). \end{aligned}$$

By Weil II (i.e., Deligne's second proof of the Weil Conjectures), $H_c^1(\mathcal{F})$ is mixed, of weights ≤ 3, which implies that

$$(3.8) \qquad |P(\gamma;\mathbb{F}_p)| \le h_c^1(\mathcal{F})\sqrt{p}^{\,3}.$$

Next, we use the Euler–Poincaré Formula to evaluate $h_c^1(\mathcal{F})$:

$$h_c^1(\mathcal{F}) = -\chi_c(U\otimes\overline{\mathbb{F}}_p,\mathcal{F}) = -\chi_c(U\otimes\overline{\mathbb{F}}_p)\operatorname{rank}(\mathcal{F}) + \sum_{x\in\mathbb{P}^1\setminus U}\mathrm{Swan}_x(\mathcal{F}).$$

We have $\operatorname{rank}(\mathcal{F}) = 4$; $-\chi_c(U\otimes\overline{\mathbb{F}}_p) = |(\mathbb{P}^1 - U)(\overline{\mathbb{F}}_p)| - 2 = |\{0,\infty,-b/a, -d/c\}| - 2$; $\mathcal{F}$ is tame at 0 and $-b/a$ and wild at ∞ and $-d/c$, with $\mathrm{Swan}_\infty(\mathcal{F}) = 2\,\mathrm{Swan}_\infty(\mathcal{K}) = 2$ and $\mathrm{Swan}_{-d/c}(\mathcal{F}) = 2\,\mathrm{Swan}_{-d/c}(\gamma^*\mathcal{K}) = 2$; so

$$h_c^1(\mathcal{F}) = 4(|\{0,\infty,-b/a,-d/c\}| - 2) + 2 + 2 = 12 - 4(\delta_{p|d} + \delta_{p|ab}).$$

Using these estimates of $h_c^1(\mathcal{F})$ in (3.8), we get the desired result.

Now consider the case $\delta > 0$. If $p = 2$ then there are no terms in the sum (3.3) defining $P(\gamma; \mathbb{F}_p)$ if d is odd and only one term if d is even. In the latter case, we have $P(\gamma; \mathbb{F}_2) = K(1; \mathbb{F}_2)K(\gamma(1); R_{1+\delta}) = K(\gamma(1); R_{1+\delta})$; by (3.5), this implies that $|P(\gamma; \mathbb{F}_2)| \leq 2\sqrt{2}^{2+\delta}$. According to Example 1.15, the Kloosterman sum vanishes if $1 + \delta \geq 6$ and a is not a square.

From now on, assume $p > 2$. We are assuming that c is a unit, and so a is also a unit, by (3.2). We will use Salié's formula for Kloosterman sums in the form (3.4) to evaluate $K(\gamma(x); R_{1+\delta})$. We have

$$\gamma(x) = \frac{ax+b}{cx+d} = \frac{a}{c} - p^\delta \frac{u}{c} \cdot \frac{1}{cx+d} = \frac{a}{c}\left(1 - p^\delta \frac{u}{a} \cdot \frac{1}{cx+d}\right).$$

In particular, $\gamma(x)$ is a square if and only if a/c is. Therefore $P(\gamma; \mathbb{F}_p)$ vanishes unless a/c is a square, so assume that it is. If $a/c = \alpha^2 \in \mathbb{Z}_p^\times$ then

$$\gamma(x) = \alpha^2\left(1 - p^\delta \frac{u}{a} \cdot \frac{1}{cx+d}\right) \equiv \left[\alpha\left(1 - \frac{1}{2}p^\delta \frac{u}{a} \cdot \frac{1}{cx+d}\right)\right]^2 \pmod{p^{1+\delta}}$$

and so

$$\begin{aligned} K(\gamma(x); R_{1+\delta}) &= \sqrt{p}^{1+\delta} \sum_{\alpha^2 = a/c} G_{1+\delta}(2\alpha) e_{1+\delta}\left(2\alpha\left(1 - \frac{1}{2}p^\delta \frac{u}{a} \cdot \frac{1}{cx+d}\right)\right) \\ &= \sqrt{p}^{1+\delta} \sum_{\alpha^2 = a/c} G_{1+\delta}(2\alpha) e_{1+\delta}(2\alpha) e_1\left(-\alpha \frac{u}{a} \cdot \frac{1}{cx+d}\right). \end{aligned}$$

Since $\delta > 0$, $ax + b$ is a unit if and only if $cx + d$ is, and so

$$(3.9) \quad P(\gamma; \mathbb{F}_p) = \sqrt{p}^{1+\delta} \sum_{\alpha^2 = a/c} G_{1+\delta}(2\alpha) e_{1+\delta}(2\alpha) \sum_{\substack{x \in \mathbb{F}_p^\times \\ p \nmid cx+d}} K(x; \mathbb{F}_p) e_1\left(-\alpha \frac{u}{a} \cdot \frac{1}{cx+d}\right).$$

The argument so far is similar to the one we will use when $m > 1$ and $\delta > 0$; there we will use stationary phase to evaluate the inner sum, but here we must use l-adic techniques. Let $\mathcal{L} = \mathcal{L}_{e_1}$ denote the standard rank-one lisse sheaf on $\mathbb{A}^1_{\mathbb{F}_p}$ such that, for all $a \in \mathbb{F}_p = \mathbb{A}^1(\mathbb{F}_p)$,

$$\operatorname{trace}(\operatorname{Frob}_a | \mathcal{L}) = e_1(a).$$

We can use the Lefschetz Trace Formula to evaluate the inner sum in (3.9): letting

$$U = \mathbb{G}_m \setminus \{-d/c\} \quad \text{and} \quad \mathcal{F} = \mathcal{K} \otimes \left[x \mapsto -\alpha \frac{u}{a} \cdot \frac{1}{cx+d}\right]^* \mathcal{L},$$

$$\sum_{\substack{x\in\mathbb{F}_p^\times\\ p\nmid cx+d}} K(x;\mathbb{F}_p)e_1\biggl(-\alpha\frac{u}{a}\cdot\frac{1}{cx+d}\biggr)=\sum_{i=0}^{2}\mathrm{trace}(\mathrm{Frob}_{\mathbb{F}_p}|H_c^i(U\otimes\overline{\mathbb{F}}_p,\mathcal{F})).$$

Since $\mathcal{F}$ is lisse and U is affine, $H_c^0(\mathcal{F})=0$. Since the pull-back of $\mathcal{L}$ is lisse at ∞, $\mathcal{F}\cong\mathcal{K}$ as P_∞-representations, so $\mathcal{F}$ is geometrically irreducible and $H_c^2(\mathcal{F})=0$. Since $\mathcal{L}$ is pure of weight 0, $\mathcal{F}$ is pure of weight 1 and so $H_c^1(\mathcal{F})$ is mixed of weights ≤ 2 by Weil II. Therefore the inner sum in (3.9) is bounded by $h_c^1(\mathcal{F})p$. We use the Euler–Poincaré Formula to evaluate $h_c^1(\mathcal{F})$:

$$h_c^1(\mathcal{F})=2(|\{0,\infty,-d/c\}|-2)+1+2=5-2\,\delta_{p|d}$$

since $\mathrm{Swan}_\infty(\mathcal{K})=\mathrm{Swan}_\infty(\mathcal{L})=1$ and they are tame elsewhere.

Since $|G_{1+\delta}(2\alpha)|=1$ and the inner sum in (3.9) is bounded by $h_c^1(\mathcal{F})p$,

$$|P(\gamma;\mathbb{F}_p)|\le\sqrt{p}^{\,1+\delta}\cdot2\cdot(5-2\,\delta_{p|d})p=(10-4\,\delta_{p|d})\sqrt{p}^{\,3+\delta}.\ \blacksquare$$

PROPOSITION 3.5. *Using Notation* 3.1, *assume that* $m>1$, $v_p(c)<m$, *and* $v_p(b)\ne v_p(3c)$. *Then*

$$|P(\gamma;R_m)|\le\sqrt{p}^{\,3m+h}(6\,\delta_{p>3}+108\sqrt{3}\,\delta_{p=3}+2^9\sqrt{2}\,\delta_{p=2}).\tag{3.10}$$

Furthermore, *in order for* $P(\gamma;R_m)\ne0$, *we have conditions in the following cases*:

(1) *Assume* $v_p(b)>v_p(3c)$ *and* $v_p(c)=0$; *if* $p=2$ *or* 3, *assume* $v_p(b)\ge 3$; *if* $p=3$, *assume* $m\ge5$; *if* $p=2$, *assume* $m\ge3\delta+8$ *or* $\delta\ge4$ *and* $m+\delta\ne7$. *Then au is a cube.*

(2) *Assume* $v_p(c)>v_p(b)=0$; *if* $p=2$, *assume* $v_p(c)\ge3$ *and either* $m\ge3\delta+8$ *or* $\delta\ge4$ *and* $m+\delta\ne7$. *Then* $v_p(u^2-ad^3)=0$ *and* $(u^2-ad^3)/bd$ *is a square.*

(3) *Assume* $\delta>0$; *if* $p=2$, *assume* $\delta\ge3$ *and* $m+\delta\ge6$. *If* $v_p(c)=0$ *then* a/c *is a square*; *if* $v_p(b)=0$ *then* b/d *is a square.*

(4) *Assume* $v=\min\{v_p(b),v_p(c)\}>0$. *Then* $v_p(a-d)\ge v$. *Assume also that* $m\ge v+2$; *if* $p=3$ *and* $v_p(b)>v_p(3c)$, *assume that* $m\ge v+5$; *if* $p=2$, *assume that* $m\ge v+8$. *Then* $v_p(a-d)=\min\{v_p(b),v_p(3c)\}$. *If* $v_p(b)>v_p(3c)$ (*and* $v_p(b)\ge v+3\ge6$ *if* $p=2$) *then* $(a-d)/3c$ *is a square*; *if* $v_p(3c)>v_p(b)$ (*and* $v_p(c)\ge v+3\ge6$ *if* $p=2$) *then* $a(a-d)/d$ *is a square.*

Proof. *First assume that* $m\ge3h+2$; if $p=2$ then assume $m\ge3h+5$. Putting the definition (3.1) of the Kloosterman sum into the definition (3.3) of $P(\gamma;R_m)$, we get

$$P(\gamma;R_m)=\frac{1}{p^{2\delta}}\sum_{\substack{x,s,t\in R_{m+\delta}^\times\\ p\nmid ax+b,cx+d}} e_{m+\delta}\biggl(p^\delta\biggl(s+\frac{x}{s}\biggr)+t+\frac{\gamma(x)}{t}\biggr).$$

We try to evaluate this sum by the method of stationary phase: let

$$f(x,s,t)=p^\delta\biggl(s+\frac{x}{s}\biggr)+t+\frac{\gamma(x)}{t};$$

$$\operatorname{grad} f=\biggl(\frac{p^\delta}{s}+\frac{p^\delta u}{(cx+d)^2t},p^\delta\biggl(1-\frac{x}{s^2}\biggr),1-\frac{\gamma(x)}{t^2}\biggr).$$

Letting H be the Hessian matrix, one calculates

$$\det H=p^{2\delta}\frac{-2ux\gamma(x)}{(cx+d)^3s^3t^4}\biggl(4c+\frac{p^\delta u}{ax+b}+\frac{(cx+d)^3}{ux}\cdot\frac{t}{s}\biggr).$$

At a $\mathbb{Z}_p$-valued critical point, we use $p^\delta u=ad-bc$ and $t/s=-u/(cx+d)^2$ and simplify to get

$$\det H=2p^{2\delta}(\mathrm{unit})(3c-b/(x\gamma(x))).$$

Since two rows of H are multiples of p^δ, its adjoint matrix is a multiple of p^δ and so we can use $h=v_p(2)+\delta+v_p(3c-b/(x\gamma(x)))$, $k=h+1$ in Theorem 1.8(b). Thus, for $m\ge 3h+2$ ($m\ge 3h+5$ if $p=2$), we get $|P(\gamma;R_m)|\le|D(\mathbb{Z}_p)|\sqrt{p}^{\,3m+h}$, where D is the scheme of critical points.

Now $D(\mathbb{Z}_p)=\{s\in\mathbb{Z}_p^\times : u^2s^2=(as^2+b)(cs^2+d)^3\}$. Suppose that the polynomial $g(x)=u^2x-(ax+b)(cx+d)^3$ vanishes; then $ac^3=bd^3=0$, so by assumption (3.2), either $a=d=0$ or $b=c=0$. In the former case, the polynomial does not vanish; we do not deal with the latter case here, since we are assuming $v_p(b)\ne v_p(3c)$, but it is covered by Proposition 3.3. Thus $g(x)$ does not vanish and $|D(\mathbb{Z}_p)|\le 8$.

We can say more about $D(\mathbb{Z}_p)$ in most cases. First, note that $D(\mathbb{Z}_p)=\widetilde{D}(\mathbb{Z}_p)$, where $\widetilde{D}\subseteq\mathbb{A}^1$ is defined by

$$\widetilde{D}=\operatorname{Spec}\mathbb{Z}_p[s,s^{-1}]/g(s^2)=\operatorname{Spec}\mathbb{Z}_p[s,s^{-1}]/(u^2s^2-(as^2+b)(cs^2+d)^3).$$

For any $s\in\widetilde{D}(\mathbb{Z}_p)$, we have $dg(s^2)/(ds)=2$ (unit) $(3c-b/(x\gamma(x)))$, so $v_p(dg(s^2)/(ds))=h-\delta$; let $\widetilde{h}:=h-\delta$. Thus $\widetilde{D}(\mathbb{Z}_p)$ is in one-to-one correspondence with the image of $\widetilde{D}(R_{2\tilde{h}+1})$ in $\widetilde{D}(R_{\tilde{h}+1})$, by Lemma 1.20. (If $\widetilde{h}=0$ then $\widetilde{D}(\mathbb{Z}_p)=\widetilde{D}(\mathbb{F}_p)$.) It follows that $|\widetilde{D}(\mathbb{Z}_p)|$ depends only on the coefficients a, b, c, d modulo $p^{2\tilde{h}+1}$. For example, if $v_p(b)\ge 2\widetilde{h}+1$, then we can replace b with 0; the condition $u^2x=(ax+b)(cx+d)^3$ becomes $u^2/a=(cx+d)^3$ and so $D(\mathbb{Z}_p)$ has at most 6 points, none at all unless u^2/a is a cube. Similarly, if $v_p(c)\ge 2\widetilde{h}+1$ then $u^2x=(ax+b)(cx+d)^3$ becomes $ax+b=(u^2/d^3)x$ and so $\widetilde{D}(\mathbb{Z}_p)$ has at most 2 points, none at all unless $v_p(u^2-ad^3)=v_p(b)$ and $(u^2-ad^3)/(bd)$ is a square.

If $\delta = 0$ then this proves (1) and (2); if $h = 0$ then it also proves (3.10). The rest of the proof consists of dealing with small values of m. Generally we will use a change of variables to get an étale scheme of critical points. For $p = 2$ or 3 and small values of m we will use crude estimates.

Now assume that $\delta > 0$. If $p = 2$ then assume $\delta \geq 3$ and $m + \delta \geq 8$. Note that ac or bd must be a unit, thanks to the assumptions (3.2) and $\delta > 0$.

As in the case $m = 1$, $\delta > 0$ we will use Salié's formula for Kloosterman sums, in the form (3.4):

$$K(\gamma(x); R_{m+\delta}) = \sqrt{p}^{\,m+\delta} \sum_{\substack{t \in R_{m+\delta}^{\times} \\ t^2 = \gamma(x)}} G_{m+\delta}(2t) e_{m+\delta}(2t) \qquad (p \neq 2);$$

if $p = 2$ then we should replace $R_{m+\delta}^{\times}$ with $R_{m+\delta-1}^{\times}$ (and still require $t^2 = \gamma(x)$ in $R_{m+\delta}$). Now

$$\gamma(x) = \frac{ax+b}{cx+d} = \frac{a}{c}\left(1 - p^{\delta}\frac{u}{a(cx+d)}\right) \quad \text{or} \quad \frac{b}{d}\left(1 + p^{\delta}\frac{ux}{b(cx+d)}\right),$$

depending on whether ac or bd is a unit. In particular, $\gamma(x) \equiv a/c$ or $b/d \pmod{p^{\delta}}$, so $\gamma(x)$ is a square if and only if a/c (respectively, b/d) is. If a/c (respectively, b/d) is not a square then $P(\gamma; R_m) = 0$; this much is true even if $p = 2$, $\delta \geq 3$, and $m + \delta \geq 6$.

Assume, therefore, that $a/c = \alpha^2$ (respectively, $b/d = \alpha^2$), with $\alpha \in \mathbb{Z}_p^{\times}$. Then $\sqrt{\gamma(x)} = \alpha(1 - p^{\delta}u/(a(cx+d)))^{1/2}$ (resp., $\alpha(1 + p^{\delta}ux/(b(cx+d)))^{1/2}$) makes sense as an element of $R_{m+\delta}[x, (cx+d)^{-1}]$. Furthermore, we have $\sqrt{\gamma(x)} \equiv \alpha \pmod{\frac{1}{2}p^{\delta}}$, so $2p^{-\delta}(\sqrt{\gamma(x)} - \alpha)$ makes sense as an element of $R_m[x, (cx+d)^{-1}]$.

If $p = 2$ then assume now that $\delta \geq 4$; then $G_{m+\delta}(2\sqrt{\gamma(x)}) = G_{m+\delta}(2\alpha)$ by Example 1.13. Therefore

$$K(\gamma(x); R_{m+\delta}) = \sqrt{p}^{\,m+\delta} \sum_{\alpha} G_{m+\delta}(2\alpha) e_{m+\delta}(2\alpha) e_m\left(2\frac{\sqrt{\gamma(x)} - \alpha}{p^{\delta}}\right).$$

(Here and below, the sum is over $\alpha \in \mathbb{Z}_p^{\times}$ such that $\alpha^2 = a/c$ or b/d.) We can use this and the definition (3.1) of $K(x; R_m)$ in the definition (3.3) of $P(\gamma; R_m)$:

$$P(\gamma; R_m) = \sqrt{p}^{\,m+\delta} \sum_{\alpha} G_{m+\delta}(2\alpha) e_{m+\delta}(2\alpha) \sum_{\substack{x, s \in R_m^{\times} \\ p \nmid ax+b, cx+d}} e_m\left(s + \frac{x}{s} + 2\frac{\sqrt{\gamma(x)} - \alpha}{p^{\delta}}\right).$$

We apply stationary phase with

$$f(x,s) = s + \frac{x}{s} + 2\frac{\sqrt{\gamma(x)} - \alpha}{p^\delta},$$
$$\operatorname{grad} f = \left(\frac{1}{s} + u\gamma(x)^{-1/2}(cx+d)^{-2}, 1 - \frac{x}{s^2}\right).$$

We calculate

$$\begin{aligned}\det H &= \frac{-ux}{s^3\sqrt{\gamma(x)}(cx+d)^3}\left[4c + \frac{p^\delta u}{\gamma(x)(cx+d)} + \frac{\sqrt{\gamma(x)}(cx+d)^3}{uxs}\right]\\ &= (\text{unit})(3c - b/(x\gamma(x))),\end{aligned}$$

where the second expression is valid at a critical point.

If $p \neq 3$ or if $p = 3 \nmid b$ then $\det H$ is a unit and Theorem 1.4 applies for all $m \geq 2$; if $p = 3$ and $v_p(b) > v_p(3c) = 1$ (so that $p \nmid c$) then Theorem 1.8(b) applies as soon as $m \geq 5$. We conclude that $|P(\gamma; R_m)| \leq |D(\mathbb{Z}_p)|\sqrt{p}^{3m+h}$, where D represents the union of the two schemes of critical points corresponding to the two choices of α. It is easy to see that $D(\mathbb{Z}_p) = \widetilde{D}(\mathbb{Z}_p)$, just as before, so our previous analysis applies.

This proves (3.10) if $\delta > 0$ ($\delta \geq 4$ and $m + \delta \geq 8$ if $p = 2$; $m \geq 5$ if $p = 3$). It also proves (3) and completes the proof of (1) and (2).

Next assume that $v_p(b)$, $v_p(3c) > 0$. We dealt with the case $p = 3 \nmid c$ just above, so let $v = \min\{v_p(b), v_p(c)\}$ and assume that $0 < v < m$. Note that a, $d \in \mathbb{Z}_p^\times$ and so $\delta = 0$, thanks to (3.2). In R_m,

$$\gamma(x + p^{m-v}z) = \frac{a(x + p^{m-v}z) + b}{cx+d} = \gamma(x) + p^{m-v}\frac{a}{d}z.$$

As x runs through $R_m^\times$ and z runs through R_v, $x + p^{m-v}z$ runs through $R_m^\times$, p^v times and so

$$\begin{aligned}&P(\gamma; R_m)\\ &\quad= \frac{1}{p^v}\sum_{x\in R_m^\times, z\in R_v} K(x + p^{m-v}z; R_m)K(\gamma(x) + p^{m-v}(a/d)z; R_m)\\ &\quad= \sum_{x,s,t\in R_m^\times} e_m(s + x/s + t + \gamma(x)/t)\cdot\frac{1}{p^v}\sum_{z\in R_v} e_v((1/s + a/(dt))z).\end{aligned}$$

The inner sum vanishes unless $as/dt \equiv -1 \pmod{p^v}$, in which case it gives p^v. Setting $s/t = -d/a + p^v w$, we get

$$(3.11)\quad P(\gamma; R_m) = \sum_{\substack{x,t\in R_m^\times\\ w\in R_{m-v}}} e_m\left([1 - d/a + p^v w]t + \left[\gamma(x) + \frac{x}{-d/a + p^v w}\right]\frac{1}{t}\right).$$

The sum over t gives

$$S\left(\frac{a-d}{a}+p^v w, \frac{ax+b}{cx+d}+\frac{ax}{-d+p^v aw}; p^m\right).$$

Since $v<m$ and the second argument of this Kloosterman sum is congruent to $ax/d - ax/d \pmod{p^v}$, (2.10) shows that the sum vanishes unless $v_p(a-d) \geq v$, as claimed. (We will see below that we need $v_p(a-d)=v$ if $v<m-1$.)

Assume, therefore, that $1-d/a=(a-d)/a=p^v\alpha$, with $\alpha \in \mathbb{Z}_p$, and set $b=p^v b'$, $c=p^v c'$. Let $s=t(-1+p^v\alpha+p^v w) \in \mathbb{Z}_p[t,w]$; alternatively, we could proceed by using four variables x, s, t, w with the relation $as+dt=p^v awt$. We have

$$s+\frac{x}{s}+t+\frac{\gamma(x)}{t}=p^v f(x,t,w)$$

with

$$f(x,t,w)=\left((\alpha+w)t+\frac{b's+awtx+c'x^2t}{(cx+d)st}\right) \in \mathbb{Z}_p[x,t,w,1/((cx+d)st)]$$

and so

$$P(\gamma;R_m)=p^{2v}\sum_{\substack{x,t\in R_{m-v}^\times\\ w\in R_{m-v}}} e_{m-v}(f(x,t,w)). \tag{3.12}$$

If $m-v=1$ then $a=d$ and $s=-t$ in $R_{m-v}=\mathbb{F}_p$ and the sum over w is elementary: it vanishes unless

$$t \equiv \frac{atx}{(cx+d)t^2} \equiv \frac{x}{t} \pmod{p},$$

in which case it gives p. Therefore

$$P(\gamma;R_m)=p^{2v+1}\sum_{t\in\mathbb{F}_p^\times} e_1(\alpha t+(b'/d)t^{-1}-(c'/d)t^3).$$

We believe that this cubic sum was first estimated using the Riemann hypothesis for curves; it is easily analyzed using l-adic cohomology. First, suppose that $v_p(b)>v_p(3c)$. With the same notation as in the proof of Proposition 3.4, consider the sheaf $\mathcal{F}=\mathcal{L}_{e_1(\alpha t-(c'/d)t^3)}$ on $\mathbb{G}_m$. This $\mathcal{F}$ is lisse on $\mathbb{A}^1$ and has Swan conductor 3 at ∞. By the usual arguments, $H_c^0(\mathcal{F})$ and $H_c^2(\mathcal{F})$ both vanish; by the Lefschetz Trace Formula and the Euler–Poincaré Formula, one finds that the cubic sum is bounded by $3\sqrt{p}$. Next, suppose that $v_p(3c)>v_p(b)=m-1$; since we are assuming $v_p(c)<m$, this can only happen if $p=3$ and $v_p(c)=m-1$. In this case, $\mathcal{F}=\mathcal{L}_{e_1(\alpha t+b'/dt-(c'/d)t^3)}$ is lisse on $\mathbb{G}_m$; $\mathrm{Swan}_\infty(\mathcal{F})=3$ and $\mathrm{Swan}_0(\mathcal{F})=1$; in this case, one finds that the cubic sum is bounded by $4\sqrt{p}$. We conclude that $|P(\gamma;R_m)| \leq p^{2v+1}\cdot 4\sqrt{p}=4\sqrt{p}^{\,3m+v}$.

Now assume $m - v > 1$. Ignoring the messy expression for $f(x,t,w)$, we can calculate

$$\operatorname{grad} f = p^{-v}\left(\frac{1}{s} + \frac{u}{(cx+d)^2 t}, \left(1 - \frac{x}{s^2}\right)\frac{s}{t} + 1 - \frac{\gamma(x)}{t^2}, \left(1 - \frac{x}{s^2}\right)p^v t\right),$$

which shows that $D(\mathbb{Z}_p) = \widetilde{D}(\mathbb{Z}_p)$, as before. Further calculation gives, at a critical point, $\det H = 2(\text{unit})(3c' - b'/(x\gamma(x)))$. If $p = 2$ (respectively, $p = 3$ and $v_p(b) > v_p(3c)$) then $v_p(\det H) = 1$ and so Theorem 1.8(b) applies for $m - v \geq 8$ (resp., $m - v \geq 5$), giving

$$|P(\gamma; R_m)| \leq p^{2v}|D(\mathbb{Z}_p)|\sqrt{p}^{3(m-v)+1} \leq 8\sqrt{p}^{3m+v+1}.$$

In all other cases, D is étale and so Theorem 1.4 applies for all $m \geq 2$, giving

$$|P(\gamma; R_m)| \leq |D(\mathbb{Z}_p)|\sqrt{p}^{3m+v}.$$

Let $s \in \widetilde{D}(\mathbb{Z}_p)$ and $x = s^2$. The equation $u^2 x = (ax+b)(cx+d)^3$ implies that $a^2d^2\alpha x \equiv b'd^3 + 3ac'd^2x^2 \pmod{p^v}$. Thus $v_p(\alpha) = \min\{v_p(b'), v_p(3c')\}$, i.e., $v_p(a-d) = \min\{v_p(b), v_p(3c)\}$. If $p \neq 2$ then, since x is a square, so is $a^2\alpha/(b'd) = a(a-d)/(bd)$ (if $v_p(3c) > v_p(b)$) or $a\alpha/(3c') = (a-d)/(3c)$ (if $v_p(b) > v_p(3c)$). The same conclusion holds for $p = 2$, provided $v \geq 3$ and $|v_p(b) - v_p(3c)| \geq 3$.

Finally, we will deal with the cases $p = 2$ and $p = 3$. According to (3.5), $|K(x; R_m)| \leq 2\sqrt{p}^m$ if p is odd. Using this in the definition of $P(\gamma; R_m)$, we find that

$$|P(\gamma; R_m)| \leq 4\sqrt{p}^{4m+\delta}. \tag{3.13}$$

If $p = 2$ then there is an extra factor of $\sqrt{2}$ in the bound on the Kloosterman sum but the sum for $P(\gamma; R_m)$ has (at most) $\phi(p^m)$ terms; the factor of $1 - 1/p$, which we ignore for most primes p, exactly cancels the two factors of $\sqrt{2}$ when $p = 2$.

Let $p = 2$. If neither b nor $3c$ is a unit then we have to consider $m \leq v+7$. The trivial estimate of the sum in (3.12) leads to $|P(\gamma; R_m)| \leq 2^8\sqrt{p}^{3m+v+1}$. If b or $3c$ is a unit then we have to consider $m+\delta \leq 7$ or $\delta \leq 3$ and $m \leq 3\delta+7$. In all these cases, $m \leq 16$. Thus (3.13) gives $|P(\gamma; R_m)| \leq 4\sqrt{2}^{16} \cdot \sqrt{p}^{3m+\delta} \leq 2^9\sqrt{2} \cdot \sqrt{p}^{3m+h}$.

Now let $p = 3$. If $v_p(b) > v_p(3c) = 1$ then we have to consider $m \leq 4$. Then (3.13) gives $|P(\gamma; R_m)| \leq 4\sqrt{3}^3 \cdot \sqrt{p}^{3m+\delta+1} < 21\sqrt{p}^{3m+h}$. If $v_p(b) > v_p(3c)$ and $v = v_p(c) > 0$ then we have to consider $m \leq v + 4$. The trivial estimate of (3.12) leads to $|P(\gamma; R_m)| \leq p^{2v}\phi(p^{m-v})^2 p^{m-v} \leq 108\sqrt{3} \times \sqrt{p}^{3m+v+1}$. ■

We will now apply the results of this section to the GL(3)-Kloosterman sums we computed in Section 2.

Notation 3.6. Let

$$w_0 = \begin{pmatrix} & & 1 \\ & -1 & \\ 1 & & \end{pmatrix}, \qquad t = \begin{pmatrix} p^s & & \\ & p^{r-s} & \\ & & p^{-r} \end{pmatrix},$$

$$\psi \begin{pmatrix} 1 & x & z \\ & 1 & y \\ & & 1 \end{pmatrix} = e^{2\pi i(\nu_1 x + \nu_2 y)}, \qquad \psi' \begin{pmatrix} 1 & x & z \\ & 1 & y \\ & & 1 \end{pmatrix} = e^{2\pi i(\nu_1' x + \nu_2' y)}$$

as in Section 2 and assume that ν_1, ν_2, ν_1', and $\nu_2' \in \mathbb{Z}_p \setminus \{0\}$. Let $\varepsilon = \min\{r, v_p(\nu_1\nu_2' + \nu_2\nu_1')\}$.

Applying Theorem 3.2 to Theorem 2.11 one can see that if $\nu_1\nu_2'+\nu_2\nu_1' = 0$ and $r = s > 0$ then $|\mathrm{Kl}(w_0t, \psi, \psi')|$ is $O(p^{4r/3})$; otherwise, it is $O(p^{(2s+3r)/4})$. Using Propositions 3.3–3.5, we can be more precise. In particular, we will find a lot of cancellation in the case $r = s$ (too much for coincidence?) from the terms $p^{-m}P(\gamma_m; R_m)$ where m is small enough that Proposition 3.3 applies; and at most five of the terms with m large are non-zero.

Theorem 3.7. *Keep the notation of Theorem* 2.11.

(a) *If* $r = s \ge n_1 + n_2 + 3$ *then*, *letting* $\tilde r = r - n_1 - n_2$, $\mathrm{Kl}(w_0t, \psi, \psi')$ *is given by*

$$\begin{aligned}\Big[p^{-\tilde r/2}&P(\gamma_{\tilde r/2}; R_{\tilde r/2})\delta_{2|\tilde r} \\ &+ p^{-(\tilde r-\varepsilon+v_p(3))/2}P(\gamma_{(\tilde r-\varepsilon+v_p(3))/2}; R_{(\tilde r-\varepsilon+v_p(3))/2}) \\ &\times \delta_{2|\tilde r-\varepsilon+v_p(3)}\delta_{\varepsilon>v_p(3)}\delta_{\tilde r\ge 3\varepsilon+4+3v_p(48)} \\ &+ \sum_{(\tilde r+1)/3\le m\le (\tilde r+1)/3+v_p(12)} p^{-m}P(\gamma_m; R_m)\delta_{(\tilde r-\varepsilon)/2\le m<\tilde r/2} \\ &+ p^{\lfloor \tilde r/3\rfloor}\delta_{\tilde r\le 3\varepsilon+2}\Big]p^{r+n_1+n_2}\delta_{N_1=M_1}\delta_{N_2=M_2}.\end{aligned}$$

(*Note that there are at most* $1+v_p(12) \le 3$ *terms in the sum.*) *In particular*,

$$\begin{aligned}|\mathrm{Kl}(w_0t, \psi, \psi')| \le [&O(1)p^{\tilde r/4}\delta_{2|\tilde r} \\ &+ O(1)p^{(\tilde r+\varepsilon-v_p(3))/4}\delta_{2|\tilde r-\varepsilon+v_p(3)}\delta_{\varepsilon>v_p(3)}\delta_{\tilde r\ge 3\varepsilon+4+3v_p(48)} \\ &+ O(1)(\delta_{3|\tilde r+1} + v_p(12))p^{\tilde r/3-1/6}\delta_{\tilde r\le 3\varepsilon+2+6v_p(12)} \\ &+ p^{\lfloor \tilde r/3\rfloor}\delta_{\tilde r\le 3\varepsilon+2}]p^{r+n_1+n_2}\delta_{N_1=M_1}\delta_{N_2=M_2},\end{aligned}$$

where $O(1) = 6\delta_{p>3} + 324\delta_{p=3} + 2^{10}\delta_{p=2}$.

(b) *If* $r = s = n_1+n_2+2$ *then* $|\mathrm{Kl}(w_0t, \psi, \psi')| \le p^{r+n_1+n_2}[8p^{1/2}+1+1/p]$.

(c) *If* $r = s \le n_1+n_2+1$ *then* $|\mathrm{Kl}(w_0t, \psi, \psi')| \le p^{2r}\min\{n_1+1, n_2+1\}$.

(d) *If* $s > r \ge N_1 + N_2 + 3$ *then*

$$|\mathrm{Kl}(w_0t, \psi, \psi')| \le O(1)p^{[2s+3(r+N_1+N_2)]/4}\delta_{N_1=M_1}\delta_{N_2=M_2}\delta_{2|r-N_1-N_2},$$

where $O(1)$ *is as in part* (a).

(e) *If* $s > r = N_1 + N_2 + 2$ *then*
$$|\mathrm{Kl}(w_0 t, \psi, \psi')| \le 6p^{[s+3\min\{N_1+M_2+1, N_2+M_1+1\}]/2}.$$

(f) *If* $r < s$ *and* $r \le N_1 + N_2 + 1$ *then* $|\mathrm{Kl}(w_0 t, \psi, \psi')|$ *is bounded by*
$$6p^{[s+3r+M_1+M_2]/2} + p^{r+s}\left[\min\{r-1, n_1, n_2, n_1+n_2+1-r\} + \frac{2}{p} + \frac{1}{p^2}\right]$$
$$\times\, \delta_{s\le\min\{N_1+M_2+2, N_2+M_1+2\}}\delta_{2s-r\le M_1+M_2+1}\delta_{s-r\le\min\{M_1,M_2\}}.$$

Proof. (a) According to Theorem 2.11, we must evaluate
$$\sum_{1\le m\le \tilde r/2} p^{-m}P(\gamma_m; R_m).$$
The terms with $1 \le m \le \tilde r/3$ are given explicitly by Proposition 3.3; this portion of the sum telescopes, leaving $p^{\lfloor \tilde r/3\rfloor}\delta_{\tilde r\le 3\varepsilon+2} - 1 - 1/p$. The remaining terms are all estimated by Proposition 3.5. According to Proposition 3.5, condition (4), the terms with $\widetilde r/3 < m < (\widetilde r + 2)/3 + v_p(12)$ vanish unless $\widetilde r - 2m \le \varepsilon$ and those with $(\widetilde r + 2)/3 + v_p(12) \le m < \widetilde r/2$ vanish unless $\widetilde r - 2m + v_p(3) = \varepsilon$. Finally, if $\widetilde r$ is even then there is a term with $m = \widetilde r/2$.

(b) If $N_1 = M_1$ and $N_2 = M_2$ then we apply Proposition 3.4. Otherwise, Theorem 2.11 gives
$$|\mathrm{Kl}(w_0 t, \psi, \psi')| \le p^{r+n_1+n_2}\left(1 - \frac{1}{p}S(1, \nu; p)\right)$$
for some unit ν and the bounds (3.5) of Weil and Salié show that this is at most $2p^{r+n_1+n_2+1/2}$.

(c) This follows easily from Theorem 2.11.

(d) This follows from Proposition 3.5.

(e) This is similar to (b).

(f) This follows easily from Theorem 2.11. ■

References

[B] N. Bourbaki, *Commutative Algebra*, Chapters 1–7, Springer, Berlin, 1989.

[B-F-G] D. Bump, S. Friedberg and D. Goldfeld, *Poincaré series and Kloosterman sums for* SL(3, ℤ), Acta Arith. 50 (1988), 31–89.

[D-R] R. Dąbrowski and M. Reeder, *Kloosterman sets in reductive groups*, J. Number Theory, to appear.

[Da] H. Davenport, *Multiplicative Number Theory*, 2nd ed., Springer, Berlin, 1980.

[D] P. Deligne, *Applications de la formule des traces aux sommes trigonométriques*, in: SGA 4 1/2, Lecture Notes in Math. 569, Springer, Berlin, 1977.

[D-G] M. Demazure and P. Gabriel, *Introduction to Algebraic Geometry and Algebraic Groups*, North-Holland, Amsterdam, 1980.

[Fi] B. Fisher, *A note on Hensel's lemma in several variables*, Proc. Amer. Math. Soc., to appear.

[F] S. Friedberg, *Poincaré series for* GL(*n*): *Fourier expansion*, *Kloosterman sums and algebreo-geometric estimates*, Math. Z. 196 (1987), 165–188.

[G] M. Greenberg, *Rational points in Henselian discrete valuation rings*, Publ. Math. IHES 31 (1966), 59–64.

[EGA] A. Grothendieck *et al.*, *Éléments de Géométrie Algébrique*, EGA Chapter IV, part 4, Publ. Math. IHES 32, IHES, Paris, 1967.

[SGA] —, *Séminaire de Géométrie Algébrique du Bois-Marie*, SGA 1 Chapter II, Lecture Notes in Math. 224, Springer, Berlin, 1971.

[H] L. Hörmander, *The Analysis of Linear Partial Differential Operators I*, Springer, Berlin, 1983.

[K1] N. Katz, *Travaux de Laumon*, Astérisque 161–162 (1988), 105–132.

[K2] —, *Gauss Sums, Kloosterman Sums and Monodromy Groups*, Ann. of Math. Stud. 116, Princeton Univ. Press, Princeton, 1988.

[K3] —, *Exponential Sums and Differential Equations*, Ann. of Math. Stud. 124, Princeton Univ. Press, Princeton, 1990.

[Kl] H. D. Kloosterman, *On the representations of a number in the form $ax^2 + by^2 + cz^2 + dt^2$*, Acta Math. 49 (1926), 407–464.

[L] M. Larsen, Appendix in [B-F-G].

[Lo-Sm1] J. H. Loxton and R. A. Smith, *Estimates for multiple exponential sums*, J. Austral. Math. Soc. Ser. A 33 (1982), 125–134.

[Lo-Sm2] —, —, *On Hua's estimate for exponential sums*, J. London Math. Soc. 26 (1982), 15–20.

[Lo-V] J. H. Loxton and R. C. Vaughan, *The estimation of complete exponential sums*, Canad. Math. Bull. 28 (1985), 440–454.

[M-H] J. Milnor and D. Husemoller, *Symmetric Bilinear Forms*, Springer, Heidelberg, 1973.

[M] D. Mumford, *The Red Book of Varieties and Schemes*, Lecture Notes in Math. 1358, Springer, Berlin, 1988.

[Sa] H. Salié, *Über die Kloostermanschen Summen $S(u, v; q)$*, Math. Z. 34 (1931), 91–109.

[Se] A. Selberg, *On the estimation of Fourier coefficients of modular forms*, in: Proc. Sympos. Pure Math. 8, Amer. Math. Soc., 1965, 1–15.

[Sm1] R. A. Smith, *On n-dimensional Kloosterman sums*, J. Number Theory 11 (1979), 324–343.

[Sm2] —, *Estimates for exponential sums*, Proc. Amer. Math. Soc. 79 (1980), 365–368.

[S] G. Stevens, *Poincaré series on* GL(r) *and Kloosterman sums*, Math. Ann. 277 (1987), 25–51.

[W1] A. Weil, *On some exponential sums*, Proc. Nat. Acad. Sci. U.S.A. 34 (1948), 204–207.

[W2] —, *Sur certains groupes d'opérateurs unitaires*, Acta Math. 111 (1964), 143–211.

Department of Mathematics
Columbia University
New York, New York 10027
U.S.A.
E-mail: rdab@math.columbia.edu
benji@math.columbia.edu

Received on 12.12.1995
and in revised form on 6.9.1996 (2902)