Annals of Mathematics **186** (2017), 413–500
https://doi.org/10.4007/annals.2017.186.2.2

Bilinear forms with Kloosterman sums and applications

By Emmanuel Kowalski, Philippe Michel, and Will Sawin

Dedicated to Henryk Iwaniec

Abstract

We prove nontrivial bounds for general bilinear forms in hyper-Kloosterman sums when the sizes of both variables may be below the range controlled by Fourier-analytic methods (Pólya-Vinogradov range). We then derive applications to the second moment of cusp forms twisted by characters modulo primes, and to the distribution in arithmetic progressions to large moduli of certain Eisenstein-Hecke coefficients on GL_3. Our main tools are new bounds for certain complete sums in three variables over finite fields, proved using methods from algebraic geometry, especially ℓ-adic cohomology and the Riemann Hypothesis.

Contents

Keywords: Kloosterman sums, Kloosterman sheaves, monodromy, Riemann Hypothesis over finite fields, short exponential sums, moments of L-functions, arithmetic functions in arithmetic progressions

AMS Classification: Primary: 11T23, 11L05, 11N37, 11N75, 11F66, 14F20, 14D05.

Ph. M. and E. K. were partially supported by a DFG-SNF lead agency program grant (grant 200021L_153647). W. S. was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1148900. W.S. partially supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation.

1. Introduction

1.1. *Statements of results.* A number of important problems in analytic number theory can be reduced to nontrivial estimates for bilinear forms

$$B(K,\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_m \sum_n \alpha_m \beta_n K(mn) \tag{1.1}$$

for some arithmetic function K and complex coefficients $(\alpha_m)_{m\geqslant 1}$, $(\beta_n)_{n\geqslant 1}$. A particularly important case is when $K : \mathbf{Z} \longrightarrow \mathbf{Z}/q\mathbf{Z} \to \mathbf{C}$ runs over a sequence of q-periodic functions, which are bounded independently of q, and estimates are required in terms of q.

In dealing with these sums, the challenges lie (1) in handling coefficients (α_m), (β_n) which are as general as possible; and (2) in dealing with coefficients supported in intervals $1 \leqslant m \leqslant M$ and $1 \leqslant n \leqslant N$ with M, N as small as possible compared with q. In this respect, a major threshold is the *Fourier-theoretic range* (also called sometimes the Pólya-Vinogradov range), where M and N are both close to $q^{1/2}$, and especially when they are slightly smaller in logarithmic scale, so that applying the completion method and even best-possible bounds for the Fourier transform gives trivial results.

In particular, when dealing with problems related to the analytic theory of automorphic forms, one is often faced with the case where $K(n)$ is a hyper-Kloosterman sum $\mathrm{Kl}_k(n;q)$. We recall that these sums are defined, for $k \geqslant 2$ and $a \in (\mathbf{Z}/q\mathbf{Z})^\times$, by

$$\mathrm{Kl}_k(a;q) = \frac{1}{q^{(k-1)/2}} \sum_{\substack{x_1,\dots,x_k \in \mathbf{Z}/q\mathbf{Z} \\ x_1\cdots x_k = a}} e\Big(\frac{x_1+\cdots+x_k}{q}\Big).$$

A deep result of Deligne shows that $|\mathrm{Kl}_k(a;q)| \leqslant k^{\omega(q)}$ for all $a \in (\mathbf{Z}/q\mathbf{Z})^\times$. For any integer c coprime to q, we also denote by $[\times c]^*\,\mathrm{Kl}_k$ the function $a \mapsto \mathrm{Kl}_k(ca;q)$.

There are several intrinsic reasons why hyper-Kloosterman sums are ubiquitous in the theory of automorphic forms:

– they are closely related, via the Bruhat decomposition, to Fourier coefficients and Whittaker models of automorphic forms and representations, and therefore occur in the Kuznetsov–Petersson formula — see, for instance, the works of Deshouillers and Iwaniec [DI83], Bump–Friedberg–Goldfeld [BFG88] or Blomer [Blo13]);
– the hyper-Kloosterman sums are the inverse Mellin transforms of certain monomials in Gauss sums, and therefore occur in computations involving root numbers in families of L-functions (as in the paper of Luo, Rudnick and Sarnak [LRS95]);
– the hyper-Kloosterman sums are constructed by iterated multiplicative convolution (see Katz's book [Kat88] for the algebro-geometric version

of this construction), which explains why they occur after applying the Voronoi summation formula on GL_k.

Our main results provide new bounds for general bilinear forms in hyper-Kloosterman sums that go beyond the Fourier-theoretic range; see Theorems 1.1 and 1.3 below. To illustrate the potential of the results, we derive two applications of these bounds in this paper. Both are related to the third source of hyper-Kloosterman sums described above, but we believe that further significant applications will arise from the other perspectives (as well as from other directions).

1.2. *Bilinear forms with Kloosterman sums.* We will always assume that the sequences $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have finite support. We denote

$$\|\boldsymbol{\alpha}\|_1 = \sum_m |\alpha_m|, \qquad \|\boldsymbol{\alpha}\|_2 = \Big(\sum_m |\alpha_m|^2\Big)^{1/2},$$

the ℓ^1 and ℓ^2 norms.

Our main result for general bilinear forms is the following:

THEOREM 1.1 (General bilinear forms). *Let q be a prime. Let c be an integer coprime to q. Let M and N be real numbers such that*

$$1 \leqslant M \leqslant Nq^{1/4}, \quad q^{1/4} < MN < q^{5/4}.$$

Let $\mathcal{N} \subset [1, q-1]$ be an interval of length $\lfloor N \rfloor$, and let $\boldsymbol{\alpha} = (\alpha_m)_{m \leqslant M}$ and $\boldsymbol{\beta} = (\beta_n)_{n \in \mathcal{N}}$ be sequences of complex numbers.

For any $\varepsilon > 0$, we have

$$B([\times c]^* \mathrm{Kl}_k, \boldsymbol{\alpha}, \boldsymbol{\beta}) \ll q^{\varepsilon} \|\boldsymbol{\alpha}\|_2 \|\boldsymbol{\beta}\|_2 (MN)^{\frac{1}{2}} \Big(M^{-\frac{1}{2}} + (MN)^{-\frac{3}{16}} q^{\frac{11}{64}}\Big), \tag{1.2}$$

where the implied constant depend only on k and ε.

Remark 1.2. The bilinear form is easily bounded by $\|\boldsymbol{\alpha}\|_2 \|\boldsymbol{\beta}\|_2 (MN)^{\frac{1}{2}}$, which we view as the trivial bound; a more elaborate treatment yields the bound of Pólya-Vinogradov type (cf. [FKM14, Th. 1.17])

$$B([\times c]^* \mathrm{Kl}_k, \boldsymbol{\alpha}, \boldsymbol{\beta}) \ll_k \|\boldsymbol{\alpha}\|_2 \|\boldsymbol{\beta}\|_2 (MN)^{\frac{1}{2}} \Big(q^{-\frac{1}{4}} + M^{-\frac{1}{2}} + N^{-\frac{1}{2}} q^{\frac{1}{4}} \log q\Big), \tag{1.3}$$

which improves the trivial bound as long as $M \gg 1$ and $N \gg q^{1/2} \log^2 q$. We then see that for $M = N$, the bound (1.2) is nontrivial as long as $M = N \geqslant q^{11/24}$, which goes beyond the Fourier-theoretic range. In the special case $M = N = q^{1/2}$, the saving factor is $q^{-1/64+\varepsilon}$.

When $\boldsymbol{\beta}$ is the characteristic function of an interval (or more generally, by summation by parts, a "smooth" function; in classical terminology, this means that the bilinear form is a "type I" sum), we obtain a stronger result:

THEOREM 1.3 (Special bilinear forms). *Let q be a prime number. Let c be an integer coprime to q. Let M, $N \geqslant 1$ be such that*

$$1 \leqslant M \leqslant N^2, \quad N < q, \quad MN < q^{3/2}.$$

Let $\boldsymbol{\alpha} = (\alpha_m)_{m \leqslant M}$ be a sequence of complex numbers bounded by 1, and let $\mathcal{N} \subset [1, q-1]$ be an interval of length $\lfloor N \rfloor$.

For any $\varepsilon > 0$, we have

$$B([\times c]^* \operatorname{Kl}_k, \boldsymbol{\alpha}, 1_{\mathcal{N}}) \ll q^{\varepsilon} \|\boldsymbol{\alpha}\|_1^{1/2} \|\boldsymbol{\alpha}\|_2^{1/2} M^{\frac{1}{4}} N \times \Big(\frac{M^2 N^5}{q^3}\Big)^{-1/12}, \tag{1.4}$$

where the implied constant depend only on k and ε.

Remark 1.4. (1) A trivial bound in that case is $\|\boldsymbol{\alpha}\|_1^{1/2} \|\boldsymbol{\alpha}\|_2^{1/2} M^{1/4} N$, which explains why we stated the result in this manner. When $M = N$, we see that our bound (1.4) is nontrivial essentially when $M = N \geqslant q^{3/7}$, which goes even more significantly below the Fourier-theoretic range. In the special case $M = N = q^{1/2}$, the saving is $q^{-1/24+\varepsilon}$.

(2) For $k = 2$, a slightly stronger result is proved by Blomer, Fouvry, Kowalski, Michel and Milićević [BFK+17, Prop. 3.1]. This builds on a method of Fouvry and Michel [FM98, §VII], which is also the basic starting point of the analysis in this paper.

(3) If $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are both characteristic functions of intervals, a stronger result is proved by Fouvry, Kowalski and Michel in [FKM14, Th. 1.16] for a much more general class of summands K, namely, the trace functions of arbitrary geometrically isotypic Fourier sheaves, with an implied constant depending then on the conductor of these sheaves. (For $M = N$, it is enough there that $MN \geqslant q^{3/8}$, and for $M = N = q^{1/2}$, the saving is $q^{-1/16+\varepsilon}$.)

1.3. *Application* 1: *moments of twisted L-functions.* Let f and g be fixed Hecke-eigenforms (of level 1 say). A long-standing problem is the evaluation with power-saving error term of the average

$$\frac{1}{\varphi(q)} \sum_{\chi \,(\mathrm{mod}\, q)} L(f \otimes \chi, 1/2) L(g \otimes \overline{\chi}, 1/2),$$

where χ runs over Dirichlet characters of prime conductor q. When f and g are nonholomorphic Eisenstein series, the problem becomes that of evaluating the fourth moment of Dirichlet L-series at $1/2$. This was studied, for instance, by Heath-Brown [HB81] and by Soundararajan [Sou07], and it was solved by Young [You11]. For f and g cuspidal, this question was studied by Gao, Khan and Ricotta [GKR09] and, with different methods, by Hoffstein and Lee [HL]. Recently, the problem was revisited in full generality by Blomer and Milićević [BM15] and by Blomer, Fouvry, Kowalski, Michel and Milićević [BFK+17]. This last work solved the problem when one of the two

forms is noncuspidal. The general bilinear bound of Theorem 1.1 (for $k = 2$) is the final ingredient to the resolution of this problem in the case where f and g are cuspidal.

THEOREM 1.5 (Moments of twisted cuspidal L-functions). *Let q be a prime number. Let f, g be cuspidal Hecke eigenforms (holomorphic forms or Maass forms) of level 1 with respective root numbers $\varepsilon(f)$ and $\varepsilon(g)$ (equal to ± 1). If f and g are either both holomorphic forms or both Maass forms, assume also that $\varepsilon(f)\varepsilon(g) = 1$.*

Let $\delta < 1/144$. If $f \neq g$, we have

$$\frac{1}{\varphi(q)} \sum_{\chi \,(\mathrm{mod}\, q)} L(f \otimes \chi, 1/2)\overline{L(g \otimes \chi, 1/2)} = \frac{2L(f \otimes g, 1)}{\zeta(2)} + O(q^{-\delta}), \tag{1.5}$$

where $L(f \otimes g, 1) \neq 0$ is the value at 1 of the Rankin–Selberg convolution of f and g, and the implied constant depends only on f, g and δ.

If $f = g$, then there exists a constant $\beta_f \in \mathbf{C}$ such that we have

$$\frac{1}{\varphi(q)} \sum_{\chi \,(\mathrm{mod}\, q)} |L(f \otimes \chi, 1/2)|^2 = \frac{2L(\mathrm{sym}^2 f, 1)}{\zeta(2)}(\log q) + \beta_f + O(q^{-\delta}), \tag{1.6}$$

where $L(\mathrm{sym}^2 f, s)$ denotes the symmetric square L-function of f, and the the implied constant depends only on f and δ.

Proof. In [BFK+17, §7.2], Theorem 1.5 (which is Theorem 1.3 in loc. cit.) was shown to follow from a certain bound on a bilinear sum of Kloosterman sums; cf. the statement of [BFK+17, Prop. 3.1]. That bound is exactly the case $k = 2$ and $c = 1$ of Theorem 1.1. □

Remark 1.6. (1) The assumption on the root number in Theorem 1.5 is necessary, since otherwise the special values vanish and the sums are identically 0.

(2) It is well established that an asymptotic formula with a power saving error term for some moment in a family of L-functions typically implies the possibility of evaluating asymptotically some additional "twisted" moments, in this case those of the shape

$$\frac{1}{\varphi(q)} \sum_{\chi \,(\mathrm{mod}\, q)} L(f \otimes \chi, 1/2)\overline{L(g \otimes \chi, 1/2)}\chi(\ell/\ell'),$$

where $1 \leqslant \ell, \ell' \leqslant L$ are coprime integers that are also coprime with q and $L = q^\eta$ for some fixed absolute constant $\eta > 0$.

Using such a formula for $f = g$, we may apply the *mollification method* and the *resonance method* and obtain further results on the special values for this family of L-functions (estimates for the distribution of the order of vanishing

at $s = 1/2$, existence of large values, for instance). This will be taken up in the forthcoming paper [BFK$^+$] jointly with Blomer, Fouvry and Milićević.

1.4. *Application* 2: *arithmetic functions in arithmetic progressions.* In our second application, we use the bound for special bilinear forms when $K = \mathrm{Kl}_3$ to study the distribution in arithmetic progressions to large moduli of certain arithmetic functions that are closely related to the triple divisor function.

THEOREM 1.7. *Let f be a holomorphic primitive cusp form of level* 1 *with Hecke eigenvalues $\lambda_f(n)$, normalized so that $|\lambda_f(n)| \leqslant d_2(n)$. For $n \geqslant 1$, let*

$$(\lambda_f \star 1)(n) = \sum_{d|n} \lambda_f(d).$$

For $x \geqslant 2$, for any $\eta < 1/102$, for any prime $q \leqslant x^{1/2+\eta}$, for any integer a coprime to q and for any $A \geqslant 1$, we have

$$\sum_{\substack{n\leqslant x\\ n\equiv a \,(\mathrm{mod}\, q)}} (\lambda_f \star 1)(n) - \frac{1}{\varphi(q)} \sum_{\substack{n\leqslant x\\ (n,q)=1}} (\lambda_f \star 1)(n) \ll \frac{x}{q}(\log x)^{-A},$$

where the implied constant depends only on (f, η, A).

When f is replaced by a specific nonholomorphic Eisenstein series, we obtain as coefficients $(\lambda_f \star 1)(n) = (d_2 \star 1)(n) = d_3(n)$, the triple (or ternary) divisor function. In that case, a result with exponent of distribution $> 1/2$ as above was first obtained (for general moduli) by Friedlander and Iwaniec [FI85]. This was subsequently improved by Heath-Brown [HB86] and more recently (for prime moduli) by Fouvry, Kowalski and Michel [FKM15b].

The approach of [FKM15b] relied ultimately on bounds for the bilinear sums $B(\mathrm{Kl}_3, \boldsymbol{\alpha}, \boldsymbol{\beta})$ when *both* sequences $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are smooth. Indeed, as already recalled, a very general estimate for $B(K, \boldsymbol{\alpha}, \boldsymbol{\beta})$ was proved in that case in [FKM14]. Here, in the cuspidal case, the splitting $d_2(n) = (1 \star 1)(n)$ is not available and we need instead a bound where only one sequence is smooth, which is given by Theorem 1.3. (We could of course also use Theorem 1.1, with a slightly weaker result.)

The functions $n \mapsto d_3(n) = (1 \star 1 \star 1)(n)$ and $n \mapsto (\lambda_f \star 1)(n)$ are the Hecke eigenvalues of certain noncuspidal automorphic representation of $\mathrm{GL}_{3,\mathbf{Q}}$, namely the isobaric representations $1 \boxplus 1 \boxplus 1$ and $\pi_f \boxplus 1$. The methods of [FI85], [HB86], [FKM15b] and of the present paper can be generalized straightforwardly to show that the n-th Hecke eigenvalue function of any fixed noncuspidal automorphic representation of $\mathrm{GL}_{3,\mathbf{Q}}$ has exponent of distribution $> 1/2$ for individual prime moduli. Extending this further to cuspidal $\mathrm{GL}_{3,\mathbf{Q}}$-representation is a natural and interesting challenge.

Theorem 1.7 is proved in Section 5.

1.5. *Further developments.* We describe here some possible extensions of our results, which will be the subject of future papers.

1.5.1. *Extension to other trace functions.* A natural problem is to try to extend Theorems 1.1 and 1.3 to more general trace functions K. In [FM98], Fouvry and Michel derived nontrivial bounds as in Theorem 1.3 (type I sums) when Kl_k is replaced by a *rational phase function* of the type

$$K_f(n) = \begin{cases} e_q(f(n)) & \text{if } n \text{ is not a pole of } f, \\ 0 & \text{otherwise,} \end{cases}$$

where q is prime, $e_q(x) = \exp(2\pi i \frac{x}{q})$ and $f \in \mathbf{F}_q(X)$ is some rational function that is not a polynomial of degree $\leqslant 2$. They proved bounds similar to Theorem 1.1 (type II sums) for K given by a *quasi-monomial* phase, defined as above with

$$f = aX^d + bX$$

for some $a, b \in \mathbf{F}_q$, $a \neq 0$ and $d \in \mathbf{Z} - \{0, 1, 2\}$. While both cases relied on arguments from algebraic geometry, they were different from, and far simpler than, those involved in the present work.

It is plausible that the methods developed in the present paper would allow for an extension of Theorems 1.3 and 1.1 to many of the families of exponential sums studied in great details in the books of Katz (in particular, in [Kat88], [Kat90]). Other potentially interesting variants that could be treated by the methods presented here are bilinear sums of the shape

$$\sum_{m,n} \alpha_m \beta_n K((m^d n)^{\pm 1}), \ d \geqslant 1 \text{ fixed.}$$

Again the case where K is a hyper-Kloosterman sum (possibly including multiplicative characters) seem particularly interesting for number theoretic applications. (See, for instance, the recent work of Nunes [Nun].)

1.5.2. *Extension to composite moduli.* In this paper, we have focused our attention on bilinear forms associated to functions K that are periodic modulo a prime q. This is in some sense the hardest case, but nevertheless it would be very useful for many applications to have bounds similar to those of Theorems 1.3 and 1.1 when the modulus q is arbitrary, or at least squarefree.

For instance, Blomer and Milićević [BM15, Th. 1] proved the analogue of the asymptotic formula in Theorem 1.5 with power saving error term when the modulus q admits a factorization $q = q_1 q_2$ where q_1 and q_2 are neither close to 1 (excluding therefore the case when q is prime, which is now solved by Theorem 1.5) nor to $q^{1/2}$. This excludes the case when q is a product of two distinct primes that are close to each other; it would be possible to treat this if a version of Theorem 1.1 for composite moduli were available.

Another direct application would be a version of Theorem 1.7 for general moduli q. This would immediately imply the following shifted convolution bound: there exists a constant $\delta > 0$, independent of f and h, such that for all $N \geqslant 1$ and $h \geqslant 0$, we have

$$\sum_{n \leqslant N} (\lambda_f \star 1)(n) d_2(n+h) \ll_f N^{1-\delta},$$

where the implied constant is independent of h. We refer to the works of Munshi [Mun13a], [Mun13b] and Topacogullari [Top16] for related results.

Other potential applications are to problems involving the Petersson–Kuznetsov trace formula (the first of the three items listed in the beginning of this introduction) as well as to the study of arithmetic functions (like the primes) in arithmetic progressions to large moduli, as suggested by Theorem 1.7.

1.6. *Structure of the proofs.* We now discuss the essential features of the proofs of our bounds for bilinear sums, in the more difficult case of general coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Several aspects of the proof are not specific to the case of hyper-Kloosterman sums. In view of possible extensions to new cases, we describe the various steps in a general setting and indicate those that are currently restricted to the case of hyper-Kloosterman sums.

Let q be a prime, and let K be the q-periodic trace function of some ℓ-adic sheaf $\mathcal{F}$ on $\mathbf{A}^1_{\mathbf{F}_q}$, which we assume to be a middle-extension pure of weight 0, geometrically irreducible and of conductor $\mathbf{c}(\mathcal{F})$. We think of q varying, while the conductor $\mathbf{c}(\mathcal{F})$ is bounded independently of q. (For the case of hyper-Kloosterman sums, the sheaf $\mathcal{F} = \mathrm{Kl}_k$ is the Kloosterman sheaf, defined by Deligne and studied by Katz [Kat88].) We denote by ψ a fixed nontrivial additive character of $\mathbf{F}_q$.

The problem of bounding the general bilinear sums $B(K, \boldsymbol{\alpha}, \boldsymbol{\beta})$, with nontrivial bounds slightly below the Fourier-theoretic range, can be handled by the following steps.

(1) We consider auxiliary functions $\mathbf{K}$ and $\mathbf{R}$, of the "sum of product" type, defined by

$$\mathbf{K}(r, s, \lambda, \boldsymbol{b}) = \psi(\lambda s) \prod_{i=1}^{2} K(s(r+b_i)) \overline{K(s(r+b_{i+2}))}$$

and

$$\mathbf{R}(r, \lambda, \boldsymbol{b}) := \sum_{s \in \mathbf{F}_q} \mathbf{K}(r, s, \lambda, \boldsymbol{b}),$$

where r, s and λ are in $\mathbf{F}_q$, and $\boldsymbol{b} = (b_1, b_2, b_3, b_4) \in \mathbf{F}_q^4$.

Building on methods developed in [FM98] (also inspired by the work of Friedlander–Iwaniec [FI85] and the "shift by ab" trick of Vinogradov and

Karatsuba), we reduce the problem in Section 2 to that of obtaining square-root cancellation bounds for two complete exponential sums involving $\mathbf{K}$ and $\mathbf{R}$. Precisely, we need to obtain bounds of the type

$$\sum_{r \,(\mathrm{mod}\, q)} \mathbf{K}(r, s, 0, \boldsymbol{b}) \ll q^{1/2}, \tag{1.7}$$

for $s \in \mathbf{F}_q^\times$, as well as *generic* bounds

$$\sum_{r \,(\mathrm{mod}\, q)} \mathbf{R}(r, \lambda, \boldsymbol{b}) \ll q, \tag{1.8}$$

$$\sum_{r \,(\mathrm{mod}\, q)} \mathbf{R}(r, \lambda, \boldsymbol{b})\overline{\mathbf{R}(r, \lambda', \boldsymbol{b})} = \delta(\lambda, \lambda')q^2 + O(q^{3/2}). \tag{1.9}$$

Here, "generic" means that the bounds should hold for every $\lambda \in \mathbf{F}_q$ provided $\boldsymbol{b}$ does not belong to some proper subvariety of $\mathbf{A}^4$. Of course, the implied constants in all these estimates must be controlled by the conductor of $\mathcal{F}$, but this can be achieved relatively easily in all cases using general arguments to bound suitable Betti numbers independently of q.

We will obtain the bounds (1.7), (1.8) and (1.9) from Deligne's general form of the Riemann Hypothesis over finite fields [Del80]. A crucial feature is that we can interpret the functions $\mathbf{K}$ and $\mathbf{R}$ themselves as trace functions of suitable ℓ-adic sheaves denoted $\mathcal{K}$ (on $\mathbf{A}^7$) and $\mathcal{R}$ (on $\mathbf{A}^6$) respectively. We call the latter the *sum-product transform sheaf* associated to the input sheaf $\mathcal{F}$, to emphasize the structure of its trace functions and the "$+ab$" trick.

Using the Grothendieck–Lefschetz trace formula and Deligne's form of the Riemann Hypothesis, we see that the bounds will result if we can show the following properties of these sheaves:

- the sheaf representing $r \mapsto \mathbf{K}(r, s, 0, \boldsymbol{b})$ is geometrically irreducible and geometrically nontrivial;
- the sheaf $\mathcal{R}_{\lambda,\boldsymbol{b}}$ with trace function $r \mapsto \mathbf{R}(r, \lambda, \boldsymbol{b})$ is geometrically irreducible, and $\mathcal{R}_{\lambda,\boldsymbol{b}}$ is not geometrically isomorphic to $\mathcal{R}_{\lambda',\boldsymbol{b}}$ if $\lambda' \neq \lambda$.

This is a natural and well-established approach, but the implementation of this strategy will require very delicate geometric analysis of the ℓ-adic sheaves involved.

(2) The first bound (1.7) is proved in great generality in Section 3 using the ideas of Katz around the Goursat–Kolchin–Ribet criterion (see [Kat90, Prop. 1.8.2]) following the general discussion of sums of products by Fouvry, Kowalski and Michel in [FKM15c]. Indeed, it is sufficient that the original sheaf $\mathcal{F}$ with trace function K be a "bountiful" sheaf in the sense of [FKM15c, Def. 1.2], a class that contains many interesting sheaves in analytic number theory (in particular, Kloosterman sheaves).

(3) To prove that the sheaf representing $r \mapsto \mathbf{R}(r, \lambda, \boldsymbol{b})$ is geometrically irreducible is much more involved. As a first step, we prove (also in Section 3) a weaker generic irreducibility property, where both $\boldsymbol{b}$ and λ are variables. Indeed, using Katz's diophantine criterion for irreducibility [Kat96, §7]), it suffices to evaluate asymptotically the second moment of the relevant trace function over all finite extensions $\mathbf{F}_{q^d}$ of $\mathbf{F}_q$, and to prove that

$$\frac{1}{(q^d)^5} \sum_{(r,\boldsymbol{b})\in\mathbf{F}_{q^d}^5} |\mathbf{R}(r, 0, \boldsymbol{b}; \mathbf{F}_{q^d})|^2 = q^d(1 + o(1)),$$

$$\frac{1}{(q^d)^2} \sum_{(r,\lambda)\in\mathbf{F}_{q^d}^2} |\mathbf{R}(r, \lambda, \boldsymbol{b}; \mathbf{F}_{q^d})|^2 = q^d(1 + o(1))),$$

as $d \to +\infty$. Again, the methods are those of [FKM15c] and require only that $\mathcal{F}$ be a bountiful sheaf.

(5) The next and final step is the crucial one and is the deepest part of this work. In the very long Section 4, we show that one can "upgrade" the generic irreducibility of $\mathcal{R}$ from the previous step to *pointwise* irreducibility of the sheaf deduced from $\mathcal{R}$ by fixing the values of λ and $\boldsymbol{b}$, where only $\boldsymbol{b}$ is required to be outside some exceptional set. This step uses such tools as Deligne's semicontinuity theorem and vanishing cycles. It requires quite precise information on the ramification properties of $\mathcal{K}$ and $\mathcal{R}$. At this stage, we need to build on the precise knowledge of the local monodromy of Kloosterman sheaves $\mathcal{K}\ell_k$, which is again due to Katz [Kat88]. We will give some indications of the ideas involved in Section 4.

Notation. We write $\delta(x, y)$ for the Kronecker delta symbol. For any prime number ℓ, we assume fixed an isomorphism $\iota : \overline{\mathbf{Q}}_\ell \to \mathbf{C}$. Let q be a prime number. Given an algebraic variety $X_{\mathbf{F}_q}$, a prime $\ell \neq q$ and a constructible $\overline{\mathbf{Q}}_\ell$-sheaf $\mathcal{F}$ on X, we denote by $t_{\mathcal{F}} : X(\mathbf{F}_q) \longrightarrow \mathbf{C}$ its trace function, defined by

$$t_{\mathcal{F}}(x) = \iota(\mathrm{Tr}(\mathrm{Fr}_{x,\mathbf{F}_q} \mid \mathcal{F}_x)),$$

where $\mathcal{F}_x$ denotes the stalk of $\mathcal{F}$ at x. More generally, for any finite extension $\mathbf{F}_{q^d}/\mathbf{F}_q$, we denote by $t_{\mathcal{F}}(\cdot; \mathbf{F}_{q^d})$ the trace function of $\mathcal{F}$ over $\mathbf{F}_{q^d}$, namely,

$$t_{\mathcal{F}}(x; \mathbf{F}_{q^d}) = \iota(\mathrm{Tr}(\mathrm{Fr}_{x,\mathbf{F}_{q^d}} \mid \mathcal{F}_x)).$$

We will usually omit writing ι; in any expression where some element z of $\overline{\mathbf{Q}}_\ell$ has to be interpreted as a complex number, we mean to consider $\iota(z)$.

We denote by $\mathbf{c}(\mathcal{F})$ the conductor of a constructible ℓ-adic sheaf $\mathcal{F}$ on $\mathbf{A}^1_{\mathbf{F}_q}$ as defined in [FKM15a] (with adaptation to deal with sheaves that may not

be middle-extensions). Recall that this is the nonnegative integer given by

$$\mathbf{c}(\mathcal{F}) = \operatorname{rank}(\mathcal{F}) + |\operatorname{Sing}(\mathcal{F})| + \sum_{x\in\operatorname{Sing}(\mathcal{F})} \operatorname{Swan}_x(\mathcal{F}) + \dim H^0_c(\mathbf{A}^1_{\overline{\mathbf{F}}_q}, \mathcal{F}),$$

where $\operatorname{Sing}(\mathcal{F}) \subset \mathbf{P}^1(\overline{\mathbf{F}}_q)$ is the set of ramification points of $\mathcal{F}$ and $\operatorname{Swan}_x(\mathcal{F})$ is the Swan conductor at x.

For convenience, we recall the general version of the Riemann Hypothesis over finite fields that will be the source of our estimates.

PROPOSITION 1.8. *Let $\mathbf{F}_q$ be a finite field with q elements. Let $\mathcal{F}$ and $\mathcal{G}$ be constructible ℓ-adic sheaves on $\mathbf{A}^1_{\mathbf{F}_q}$ that are geometrically irreducible, mixed of weights $\leqslant 0$ and pointwise pure of weight 0 on a dense open subset. We have*

$$\sum_{x\in\mathbf{F}_q} t_{\mathcal{F}}(x;\mathbf{F}_q)\overline{t_{\mathcal{G}}(x;\mathbf{F}_q)} \ll, \sqrt{q},$$

unless $\mathcal{F}$ is geometrically isomorphic to $\mathcal{G}$, and

$$\sum_{x\in\mathbf{F}_q} |t_{\mathcal{F}}(x;\mathbf{F}_q)|^2 = q + O(\sqrt{q}).$$

The implied constants depend only on the conductors of $\mathcal{F}$ and $\mathcal{G}$.

We denote by $\mathcal{F}^\vee$ the dual of a constructible sheaf $\mathcal{F}$; if $\mathcal{F}$ is a middle-extension sheaf, we will use the same notation for the middle-extension dual.

Let ψ (resp. χ) be a nontrivial additive (resp. multiplicative) character of $\mathbf{F}_q$. We denote by $\mathcal{L}_\psi$ (resp. $\mathcal{L}_\chi$) the associated Artin–Schreier (resp. Kummer) sheaf on $\mathbf{A}^1_{\mathbf{F}_q}$ (resp. on $(\mathbf{G}_m)_{\mathbf{F}_q}$), as well (by abuse of notation) as their middle extension to $\mathbf{P}^1_{\mathbf{F}_q}$. The trace functions of the latter are given by

$$t_\psi(x;\mathbf{F}_{q^d}) = \psi(\operatorname{Tr}_{\mathbf{F}_{q^d}/\mathbf{F}_q}(x)) \quad \text{if } x\in\mathbf{F}_{q^d}, \quad t_\psi(\infty;\mathbf{F}_{q^d}) = 0,$$

$$t_\chi(x;\mathbf{F}_{q^d}) = \chi(\operatorname{Nr}_{\mathbf{F}_{q^d}/\mathbf{F}_q}(x)) \quad \text{if } x\in\mathbf{F}^\times_{q^d}, \quad t_\chi(0;\mathbf{F}_{q^d}) = t_\chi(\infty;\mathbf{F}_{q^d}) = 0$$

(which we denote also by $\psi_{q^d}(x)$ and by $\chi_{q^d}(x)$, respectively). For the trivial additive or multiplicative character, the trace function of the middle-extension is the constant function 1.

Given $\lambda\in\mathbf{F}_{q^d}$, we denote by $\mathcal{L}_{\psi_\lambda}$ the Artin–Schreier sheaf of the character of $\mathbf{F}_{q^d}$ defined by $x\mapsto\psi(\operatorname{Tr}_{\mathbf{F}_{q^d}/\mathbf{F}_q}(\lambda x))$.

If $q\geqslant 3$, we denote by χ_2 the Legendre symbol on $\mathbf{F}_q$.

If $X_{\mathbf{F}_q}$ is an algebraic variety, ψ (resp. χ) is an ℓ-adic additive character of $\mathbf{F}_q$ (resp. ℓ-adic multiplicative character) and $f : X\longrightarrow\mathbf{A}^1$ (resp. $g : X\longrightarrow\mathbf{G}_m$) is a morphism, we denote by either $\mathcal{L}_{\psi(f)}$ or $\mathcal{L}_\psi(f)$ (resp. by $\mathcal{L}_{\chi(g)}$ or $\mathcal{L}_\chi(g)$) the pullback $f^*\mathcal{L}_\psi$ of the Artin–Schreier sheaf associated to ψ (resp. the pullback $g^*\mathcal{L}_\chi$ of the Kummer sheaf). These are lisse sheaves on X with trace functions $x\mapsto\psi(f(x))$ and $x\mapsto\chi(g(x))$, respectively. The meaning of the notation $\mathcal{L}_\psi(f)$, which we use when putting f as a subscript would be

typographically unwieldy, will always be unambiguous, and no confusion with Tate twists will arise.

Given a variety $X/\mathbf{F}_q$, an integer $k \geqslant 1$ and a function c on X, we denote by $\mathcal{L}_\psi(cs^{1/k})$ the sheaf on $X \times \mathbf{A}^1$ (with coordinates (x, s)) given by

$$\alpha_* \mathcal{L}_{\psi(c(x)t)},$$

where α is the covering map $(x, s, t) \mapsto (x, s)$ on the k-fold cover

$$\{(x, s, t) \in X \times \mathbf{A}^1 \times \mathbf{A}^1 \mid t^k = s\}.$$

Given a field extension $L/\mathbf{F}_p$, and elements $\alpha \in L^\times$ and $\beta \in L$, we denote by $[\times\alpha]$ the scaling map $x \mapsto \alpha x$ on $\mathbf{A}^1_L$, and by $[+\beta]$ the additive translation $x \mapsto x + \beta$. For a sheaf $\mathcal{F}$, we denote by $[\times\alpha]^*\mathcal{F}$ (resp. $[+\alpha]^*\mathcal{F}$) the respective pull-back operation. More generally, given an element $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{PGL}_2$, we denote by $\gamma^*\mathcal{F}$ the pullback under the fractional linear transformation on $\mathbf{P}^1$ given by

$$\gamma \cdot x = \frac{ax + b}{cx + d}.$$

We usually omit mention of any necessary base change to L if the matrix involved is in $\mathrm{PGL}_2(L)$ for some extension $L/\mathbf{F}_q$.

We will usually not indicate base points in étale fundamental groups; whenever this occurs, it will be clear that the properties under consideration are independent of the choice of a base point.

As mentioned above, a large portion of our argument is valid for a more general class of functions K than hyper-Kloosterman sums. We now state the definition of the relevant class of sheaves, which is a slight extension of [FKM15c, Def. 1.2]. Let $\mathcal{G}$ be a middle-extension sheaf on $\mathbf{A}^1$ of rank $k \geqslant 2$, which is pure of weight 0. Let $U = \mathbf{A}^1 - S_\mathcal{G}$ denote the maximal open subset where $\mathcal{G}$ is lisse, and let $\mathbf{c}(\mathcal{G})$ be the conductor of $\mathcal{G}$. Let $\mathcal{F}$ be *either* $\mathcal{G}$ or the extension by zero to $\mathbf{A}^1$ of $\mathcal{G}|U$.

Definition 1.9. We say that $\mathcal{F}$ is *bountiful* (resp. *bountiful with respect to the upper-triangular Borel subgroup* $\mathrm{B} \subset \mathrm{PGL}_2$) if

- The geometric and arithmetic monodromy groups of the lisse sheaf $\mathcal{F}|U$, or equivalently of $\mathcal{G}|U$, coincide and are equal either SL_k if $k \geqslant 3$ or to Sp_k. Accordingly, we will say that $\mathcal{F}$ (or $\mathcal{G}$) is of Sp or SL type.
- For any nontrivial element $\gamma \in \mathrm{PGL}_2(\overline{\mathbf{F}}_q)$ (resp. in $\mathrm{B}(\overline{\mathbf{F}}_q)$), the sheaf $\gamma^*\mathcal{G}$ is not geometrically isomorphic to $\mathcal{G} \otimes \mathcal{L}$ for any rank 1 sheaf $\mathcal{L}$.
- If $\mathcal{F}$ is of SL-type, there is at most one $\xi \in \mathrm{PGL}_2(\overline{\mathbf{F}}_q)$ (resp. $\xi \in \mathrm{B}(\overline{\mathbf{F}}_q)$) such that we have a geometric isomorphism

$$\xi^*\mathcal{G} \simeq \mathcal{G}^\vee \otimes \mathcal{L}$$

for some rank 1 sheaf $\mathcal{L}$. If the element ξ exists, it is called the special involution of $\mathcal{F}$. It is *exactly* of order 2 and in the Borel case, is of the shape

$$\xi_{\mathcal{F}} = \begin{pmatrix} -1 & b_{\mathcal{F}} \\ & 1 \end{pmatrix}.$$

Remark 1.10. We take this occasion to address a minor slip in [FKM15c] pointed by one of the referees: the original definition of a bountiful sheaf should have required the rank of the sheaf to be $\geqslant 3$ in the SL case, since SL_2 should be viewed as a symplectic group in this context (because its standard representation is self-dual). Correspondingly [FKM15c, Th. 1.5] should include this condition as well. This has no impact on applications since the resulting corollaries all included that condition in their statement.

Remark 1.11. Another difference with [FKM15c] is that we allow the possibility that $\mathcal{F}$ be the extension by zero of $\mathcal{G}$ and do not require that $\mathcal{F}$ be necessarily a middle extension. It is immediate that the results of [FKM15c] that we use extend to this slightly more general class of sheaves: the arguments there are either performed on a dense open subset where all sheaves involved are lisse, or only depend on the bound $|t_{\mathcal{G}}(x)| \leqslant \mathrm{rank}(x)$ for a middle-extension sheaf $\mathcal{G}$ (see, e.g., [FKM15c, p. 21, proof of Prop. 1.1]). We refer to Remark 4.7 for a justification of this change in the definition of [FKM15c].

The Kloosterman sheaves $\mathcal{K}\ell_k$ (defined here as extension by zero of the Kloosterman sheaves on $\mathbf{G}_m$) are examples of bountiful sheaves. They are of Sp-type if k is even and of SL-type if k is odd (cf. [Kat88, FKM15c]), and in that case, there is a special involution given by $\xi = \left(\begin{smallmatrix} -1 & \\ & 1 \end{smallmatrix}\right)$, and indeed $\xi^*\mathcal{K}\ell_k \simeq \mathcal{K}\ell_k^\vee$. All this will be recalled with references in Section 4.2.

Acknowledgements. We acknowledge the deep influence of É. Fouvry on this work. The ideas of our collaborators concerning the problem of averages of twisted L-functions in [BFK+17] (É. Fouvry, V. Blomer and D. Milićević) were also of great importance in motivating our work on this paper. We also thank P. Nelson and I. Petrow for many discussions.

We are extremely thankful to the referee who read Sections 3 and 4 and pointed out many minor slips and a few more significant issues in the first and second version of this paper. We thank him or her in particular for giving very useful references to certain papers of L. Fu that corrected and simplified some of our local monodromy computations.

A large portion of this paper was written while Ph. M. and W. S. were enjoying the hospitality of the Forschungsinstitut für Mathematik at ETH Zürich, and it was continued during a visit of Ph. M. at Caltech. We would like to thank both institutions for providing excellent working conditions.

2. **Reduction to complete exponential sums**

In this section, we perform the first step of the proof of Theorems 1.3 and 1.1: the reduction to estimates for complete sums over finite fields. The two subsections below are essentially independent; the first one concerns special bilinear forms ("type I," as in Theorem 1.3) and the second discusses the case of general bilinear forms ("type II," as in Theorem 1.1).

2.1. *Special bilinear forms.* We follow the method of [FM98], as generalized in [BFK+17, §6.2]. Let q be a prime number, and let $\mathcal{F}$ be a bountiful sheaf on $\mathbf{A}^1_{\mathbf{F}_q}$ (with respect to the Borel subgroup). Let $k \geqslant 2$ be the rank of $\mathcal{F}$ and $\mathbf{c}(\mathcal{F})$ its conductor.

We fix some $c \in \mathbf{F}_q^\times$ and denote $K_c = [\times c]^* K$. We consider the special bilinear form

$$B(K_c, \boldsymbol{\alpha}, \mathcal{N}) = \sum_{m \leqslant M} \sum_{n \in \mathcal{N}} \alpha_m K(cmn),$$

where $\mathcal{N}$ is an interval in $[1, q-1]$ of length $\lfloor N \rfloor$ and $\boldsymbol{\alpha} = (\alpha_m)_{m \leqslant M}$ with

$$1 \leqslant M \leqslant N^2, \ N < q, \ MN < q^{3/2}. \tag{2.1}$$

Remark 2.1. The condition $MN < q^{3/2}$ is somewhat restrictive. It arises from the estimate of the possible "bad" parameter $\boldsymbol{b}$. (See the proof of Theorem 2.4 below.) However, for $MN \geqslant q^{3/2}$, other methods lead to nontrivial estimates for these bilinear forms (e.g., the bound (1.3)).

Given auxiliary integral parameters $A, B \geqslant 1$ such that

$$2B < q, \quad AB \leqslant N, \quad AM < q, \tag{2.2}$$

we have

$$\begin{aligned} B(K_c, \boldsymbol{\alpha}, \mathcal{N}) &= \frac{1}{AB} \sum_{\substack{A < a \leqslant 2A \\ B < b \leqslant 2B}} \sum_{m \leqslant M} \alpha_m \sum_{n + ab \in \mathcal{N}} K_c(m(n+ab)) \\ &= \frac{1}{AB} \sum_{\substack{A < a \leqslant 2A \\ B < b \leqslant 2B}} \sum_{m \leqslant M} \alpha_m \sum_{n + ab \in \mathcal{N}} K_c(am(\overline{a}n + b)). \end{aligned}$$

We get

$$B(K_c, \boldsymbol{\alpha}, \mathcal{N}) \ll_\varepsilon \frac{q^\varepsilon}{AB} \sum_{\substack{r \,(\mathrm{mod}\, q) \\ s \leqslant 2AM}} \sum \nu(r,s) \Big| \sum_{B < b \leqslant 2B} \eta_b K_c(s(r+b)) \Big|,$$

where

$$\nu(r,s) = \sum_{\substack{A < a \leqslant 2A, \ m \leqslant M, \ n \in \mathcal{N} \\ am = s, \ \overline{a}n \equiv r \,(\mathrm{mod}\, q)}} \sum \sum |\alpha_m|$$

and $(\eta_b)_{B<b\leqslant 2B}$ are some complex numbers such that $|\eta_b| \leqslant 1$. We have clearly

$$\sum_{r,s} \nu(r,s) \ll AN \sum_{m\leqslant M} |\alpha_m|.$$

We also have

$$\sum_{r,s} \nu(r,s)^2 = \sum_{\substack{a,m,n,a',m',n' \\ am=a'm' \\ a'n=an' \,(\mathrm{mod}\, q)}} \cdots \sum |\alpha_m||\alpha_{m'}|.$$

Observe that, once a and m are given, the equation $am = a'm'$ determines a' and m' up to $O(q^\varepsilon)$ possibilities; furthermore, for each such pair (a,m) and each $n \in \mathcal{N}$, the congruence $a'n = an' \,(\mathrm{mod}\, q)$ determines n' uniquely, as n' varies over an interval of length $\leqslant q$. Therefore we get

$$\sum_{r,s} \nu(r,s)^2 \ll \sum_{a,m} |\alpha_m|^2 \sum_{\substack{n,a',m',n' \\ am=a'm' \\ a'n=an' \,(\mathrm{mod}\, q)}} \cdots \sum 1 \ll_\varepsilon q^\varepsilon AN \sum_m |\alpha_m|^2,$$

where we have used the inequality $|\alpha_m||\alpha_{m'}| \leqslant |\alpha_m|^2 + |\alpha_{m'}|^2$.

We next apply Hölder's inequality in the form

$$\begin{aligned}
&\sum_{\substack{r \,(\mathrm{mod}\, q) \\ 1\leqslant s\leqslant 2AM}} \sum \nu(r,s) \Big| \sum_{B<b\leqslant 2B} \eta_b K_c(s(r+b)) \Big| \\
&\qquad \leqslant \Big(\sum_{r,s} \nu(r,s) \Big)^{\frac12} \Big(\sum_{r,s} \nu(r,s)^2 \Big)^{\frac14} \\
&\qquad\qquad \times \Big(\sum_{r,s} \Big| \sum_{B<b\leqslant 2B} \eta_b K_c(s(r+b)) \Big|^4 \Big)^{\frac14} \\
&\qquad \ll_\varepsilon q^\varepsilon (AN)^{\frac34} \|\boldsymbol{\alpha}\|_1^{\frac12} \|\boldsymbol{\alpha}\|_2^{\frac12} \Big(\sum_{r,s} \Big| \sum_{B<b\leqslant 2B} \eta_b K_c(s(r+b)) \Big|^4 \Big)^{\frac14}.
\end{aligned}$$

Expanding the fourth power, we have

$$\sum_{r,s} \Big| \sum_{B<b\leqslant 2B} \eta_b K_c(s(r+b)) \Big|^4 \leqslant \sum_{\boldsymbol{b}\in\mathcal{B}} \Big| \Sigma(K_c, \boldsymbol{b}; AM) \Big|, \tag{2.3}$$

where $\mathcal{B}$ denotes the set of tuples $\boldsymbol{b} = (b_1, b_2, b_3, b_4)$ of integers satisfying $B < b_i \leqslant 2B$ $(i = 1, \dots, 4)$, and

$$\Sigma(K_c, \boldsymbol{b}; AM) = \sum_{\substack{r \,(\mathrm{mod}\, q) \\ 1\leqslant s\leqslant 2AM}} \sum \prod_{i=1}^{2} K_c(s(r+b_i)) \overline{K_c(s(r+b_{i+2}))}. \tag{2.4}$$

This is a sum over r and s of a product of four values of the trace function K, which we will later specialize to hyper-Kloosterman sums. At this stage, we

have proved the bound

$$B(K_c,\boldsymbol{\alpha},\mathcal{N})\ll q^{\varepsilon}\frac{N^{3/4}}{A^{1/4}B}\|\boldsymbol{\alpha}\|_1^{1/2}\|\boldsymbol{b}\|_2^{1/2}\Big(\sum_{\boldsymbol{b}\in\mathcal{B}}\Big|\Sigma(K_c,\boldsymbol{b};AM)\Big|\Big)^{1/4} \tag{2.5}$$

for any $\varepsilon>0$, where the implied constant depends on ε and on the conductor of $\mathcal{F}$.

To continue, we first define the "diagonal" in the space of the parameters $\boldsymbol{b}\in\mathcal{B}$. We recall that sheaves of Sp-type or of SL-type were introduced in Definition 1.9.

Definition 2.2. Let $\mathcal{V}^{\Delta}$ be the affine variety of 4-uples

$$\boldsymbol{b}=(b_1,b_2,b_3,b_4)\in\mathbf{A}^4_{\mathbf{F}_q}$$

defined by the following conditions:

– if $\mathcal{F}$ is of Sp-type, then for any $i\in\{1,\dots,4\}$, the cardinality

$$|\{j=1,\dots,4\mid b_j=b_i\}|$$

is even;

– if $\mathcal{F}$ is of SL-type, then for any $i\in\{1,2\}$, we have

$$|\{j=1,2\mid b_j=b_i\}|-|\{j=3,4\mid b_j=b_i\}|=0.$$

We now denote by $\mathcal{B}^{\Delta}$ the subset of tuples of integers $\boldsymbol{b}\in\mathcal{B}$ such that

$$\boldsymbol{b}\,(\mathrm{mod}\,q)\in\mathcal{V}^{\Delta}(\mathbf{F}_q).$$

Since $k\geqslant 2$ and $2B<q$ (by (2.2)), we have $|\mathcal{B}^{\Delta}|=O(B^2)$. For $\boldsymbol{b}\in\mathcal{B}^{\Delta}$, we estimate $\Sigma(K_c,\boldsymbol{b};AM)$ trivially using the bound $|K(cx)|\leqslant\mathbf{c}(\mathcal{F})$. The contribution to (2.3) of all $\boldsymbol{b}\in\mathcal{B}^{\Delta}$ satisfies

$$\sum_{\boldsymbol{b}\in\mathcal{B}^{\Delta}}|\Sigma(K_c,\boldsymbol{b};AM)|\ll AB^2Mq, \tag{2.6}$$

where the implied constant depends only on the conductor of $\mathcal{F}$.

In Section 3, we will establish two estimates concerning the contribution of $\boldsymbol{b}\notin\mathcal{B}^{\Delta}$. For the first argument, we fix the value of s with $1\leqslant s\leqslant 2AM$ and we average over r.

Lemma 2.3. *For $\boldsymbol{b}\in\mathcal{B}\backslash\mathcal{B}^{\Delta}$ and any $s\in\mathbf{F}_q^{\times}$, we have*

$$\sum_{r\,(\mathrm{mod}\,q)}\prod_{i=1}^{2}K_c(s(r+b_i))\overline{K_c(s(r+b_{i+2}))}\ll_k q^{1/2},$$

where the implied constant depends only on $\mathbf{c}(\mathcal{F})$.

In particular, for any subset $\mathcal{B}'\subset\mathcal{B}\backslash\mathcal{B}^{\Delta}$, we have

$$\sum_{\boldsymbol{b}\in\mathcal{B}'}|\Sigma(K_c,\boldsymbol{b};AM)|\ll_k AM|\mathcal{B}'|q^{1/2}, \tag{2.7}$$

where the implied constant depends only on $\mathbf{c}(\mathcal{F})$.

This result gives a saving of a factor $q^{1/2}$ over the trivial bound. We refer to Section 3.1 for the proof.

The second argument is much deeper, and we can only bring it to completion for hyper-Kloosterman sums. We apply the discrete Plancherel formula to complete the sum with respect to the variable s. (See, for instance, [IK04, Lemma 12.1 and following].) Recall that ψ is a fixed nontrivial additive character of $\mathbf{F}_q$. For any function $L\colon \mathbf{F}_q \to \mathbf{C}$, define

$$\hat{\Sigma}(L, \boldsymbol{b}, \lambda) = \sum_{r\in\mathbf{F}_q} \mathbf{R}(L, r, \lambda, \boldsymbol{b})$$

with

$$\mathbf{R}(L, r, \lambda, \boldsymbol{b}) = \sum_{s\in\mathbf{F}_q^\times} \psi(\lambda s) \prod_{i=1}^{2} L(s(r+b_i))\overline{L(s(r+b_{i+2}))}. \tag{2.8}$$

Then, observing that for any $c \in \mathbf{F}_q^\times$, we have

$$\mathbf{R}(K_c, r, \lambda, \boldsymbol{b}) = \mathbf{R}(K, r, \lambda/c, \boldsymbol{b}), \qquad \hat{\Sigma}(K_c, \boldsymbol{b}, \lambda) = \hat{\Sigma}(K, \boldsymbol{b}, \lambda/c), \tag{2.9}$$

and the completion yields the bound

$$\Sigma(K_c, \boldsymbol{b}; AM) \ll (\log q) \max_{\lambda\in\mathbf{F}_q} |\hat{\Sigma}(K, \boldsymbol{b}, \lambda)|,$$

where the implied constant is absolute.

Taking $\mathcal{F}$ to be the Kloosterman sheaf with trace function $K = \mathrm{Kl}_k$, we will obtain an additional saving of $q^{1/2}$ in comparison with Lemma 2.3, from the cancellation in the completed variable s, leading to a net saving $AMq^{1/2}$.

THEOREM 2.4. *Let $k \geqslant 2$, and let $K = \mathrm{Kl}_k$. There exists a codimension 1 subvariety $\mathcal{V}^{\mathrm{bad}} \subset \mathbf{A}^4_{\mathbf{F}_q}$ containing $\mathcal{V}^{\Delta}$, with degree bounded independently of q, such that for any $\lambda \in \mathbf{F}_q$ and any $\boldsymbol{b} \notin \mathcal{V}^{\mathrm{bad}}(\mathbf{F}_q)$, we have*

$$\Sigma([\times c]^* \mathrm{Kl}_k, \boldsymbol{b}; AM) \ll q \log q$$

for any $c \in \mathbf{F}_q^\times$. The implied constant depends only on k.

This follows from Theorem 4.11 in Section 4.

Now, assuming Lemma 2.3 and Theorem 2.4, we can conclude the proof of Theorem 1.3. Indeed, set

$$\mathcal{B}^{\mathrm{bad}} = \mathcal{B} \cap \{\boldsymbol{b} \in \mathcal{B} \mid \boldsymbol{b} \,(\mathrm{mod}\, q) \in \mathcal{V}^{\mathrm{bad}}(\mathbf{F}_q)\}, \qquad \mathcal{B}^{\mathrm{gen}} = \mathcal{B}\backslash\mathcal{B}^{\mathrm{bad}}.$$

Since $\mathcal{V}^{\mathrm{bad}}$ has degree bounded in terms of k only, independently of q, we have $|\mathcal{B}^{\mathrm{bad}}| = O_k(B^3)$. (In fact, $|\mathcal{B}^{\mathrm{bad}}| \leqslant (\deg \mathcal{V}^{\mathrm{bad}})|B|^3$ by the so-called Schwarz–Zippel Lemma.)

Thus, applying Theorem 2.4 for $\boldsymbol{b} \in \mathcal{B}^{\mathrm{gen}}$, the bound (2.7) from Lemma 2.3 for $\boldsymbol{b} \in \mathcal{B}^{\mathrm{bad}} - \mathcal{B}^{\Delta}$, and finally (2.6) for $\boldsymbol{b} \in \mathcal{B}^{\Delta}$, we obtain

$$\sum_{\boldsymbol{b}\in\mathcal{B}} \Big|\Sigma([\times c]^* \operatorname{Kl}_k, \boldsymbol{b}; AM)\Big| \ll_k (B^4 q + AB^3 M q^{1/2} + AB^2 M q)(\log q).$$

Upon choosing

$$A = M^{-\frac{1}{3}} N^{\frac{2}{3}}, \quad B = (MN)^{\frac{1}{3}},$$

(which satisfy (2.2) by (2.1)), we see that the first and third terms in parenthesis coincide and are equal to $(MN)^{4/3}q$, while the second term is equal to

$$(MN)^{4/3}q \times (MNq^{-3/2})^{1/3} \leqslant (MN)^{4/3}q$$

by (2.1). Therefore we deduce from (2.5) that

$$\begin{aligned} B([\times c]^* \operatorname{Kl}_k, \boldsymbol{\alpha}, \mathcal{N}) &\ll_{k,\varepsilon} \frac{q^{\varepsilon}}{AB}(AN)^{\frac{3}{4}} \|\boldsymbol{\alpha}\|_1^{\frac{1}{2}} \|\boldsymbol{\alpha}\|_2^{\frac{1}{2}} B q^{1/4} \\ &\ll_{k,\varepsilon} q^{\varepsilon} \|\boldsymbol{\alpha}\|_1^{\frac{1}{2}} \|\boldsymbol{\alpha}\|_2^{\frac{1}{2}} M^{1/4} N \Big(\frac{M^2 N^5}{q^3}\Big)^{-1/12}. \end{aligned}$$

This proves Theorem 1.3, subject to the proof of Lemma 2.3 and of Theorem 2.4.

2.2. *General bilinear forms.* We now consider the situation of Theorem 1.1. Again we begin with a prime q and a bountiful sheaf $\mathcal{F}$ on $\mathbf{A}^1_{\mathbf{F}_q}$ with respect to the Borel subgroup. Let $k \geqslant 2$ be the rank of $\mathcal{F}$ and $\mathbf{c}(\mathcal{F})$ its conductor.

Given $M, N \geqslant 1$ satisfying

$$1 \leqslant M \leqslant N q^{1/4}, \ q^{1/4} < MN < q^{5/4}, \tag{2.10}$$

an interval $\mathcal{N} \subset [1, q-1]$ of length $\lfloor N \rfloor$ and sequences $\boldsymbol{\alpha} = (\alpha_m)_{m\leqslant M}$ and $\boldsymbol{\beta} = (\beta_n)_{n\in\mathcal{N}}$, we consider the general bilinear form

$$B(K_c, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{m\leqslant M,\, n\in\mathcal{N}} \sum \alpha_m \beta_n K(cmn).$$

We begin once more as in [FM98], [BFK$^+$17]. We choose auxiliary parameters $A, B \geqslant 1$ satisfying (2.2). The argument of [BFK$^+$17, §5.5] leads to the estimate

$$|B(K_c, \boldsymbol{\alpha}, \boldsymbol{\beta})|^2 \ll \|\boldsymbol{\alpha}\|_2^2 \|\boldsymbol{\beta}\|_2^2 \Big(N + \frac{q^{\varepsilon}}{AB}(AN)^{3/4} M^{1/2} \Big(\sum_{\boldsymbol{b}} |\Sigma^{\neq}(K_c, \boldsymbol{b}; AM)|\Big)^{1/4}\Big) \tag{2.11}$$

for any $\varepsilon > 0$, where the implied constant depends only on $\mathbf{c}(\mathcal{F})$ and ε, and where

$$\Sigma^{\neq}(K_c, \boldsymbol{b}; AM) = \sum_{r \,(\mathrm{mod}\, q)} \sum_{\substack{1 \leqslant s_1, s_2 \leqslant AM \\ s_1 \neq s_2 \,(\mathrm{mod}\, q)}} \sum \prod_{i=1}^{2} K_c(s_1(r+b_i))\overline{K}_c(s_2(r+b_i)) \overline{K_c(s_1(r+b_{i+2}))\overline{K}_c(s_2(r+b_{i+2}))}$$

for $\boldsymbol{b}$ running over the set $\mathcal{B}$ of quadruples of integers (b_1, b_2, b_3, b_4) satisfying $B < b_i \leqslant 2B$. Note that, in the case $K = \mathrm{Kl}_k$, we now have a sum, over the three variables (r, s_1, s_2), of a product of eight hyper-Kloosterman sums.

We will estimate the inner triple sum over r, s_1, s_2 in different ways depending on the value taken by $\boldsymbol{b}$.

First, for $\boldsymbol{b} \in \mathcal{B}^{\Delta}$ (as defined in Definition 2.2), we use the trivial bound from $|K(cx)| \leqslant \mathbf{c}(\mathcal{F})$ and obtain

$$\sum_{\boldsymbol{b} \in \mathcal{B}^{\Delta}} |\Sigma^{\neq}(K_c, \boldsymbol{b}; AM)| \ll qA^2B^2M^2, \tag{2.12}$$

where the implied constant depends only on $\mathbf{c}(\mathcal{F})$.

We next have an analogue of Lemma 2.3, where we sum over the variable r for fixed (s_1, s_2):

LEMMA 2.5. *For $\boldsymbol{b} \in \mathcal{B} \backslash \mathcal{B}^{\Delta}$ and any $s_1,\ s_2 \in \mathbf{F}_q^{\times}$ with $s_1 \neq s_2$, we have*

$$\sum_{r \,(\mathrm{mod}\, q)} \prod_{i=1}^{2} K_c(s_1(r+b_i))\overline{K}_c(s_2(r+b_i)) \overline{K_c(s_1(r+b_{i+2}))\overline{K}_c(s_2(r+b_{i+2}))} \ll q^{1/2}, \tag{2.13}$$

where the implied constant depends only on $\mathbf{c}(\mathcal{F})$.

In particular, for any subset $\mathcal{B}' \subset \mathcal{B} \backslash \mathcal{B}^{\Delta}$, we have

$$\sum_{\boldsymbol{b} \in \mathcal{B}'} |\Sigma^{\neq}(K_c, \boldsymbol{b}; AM)| \ll (AM)^2 |\mathcal{B}'| q^{1/2},$$

where the implied constant depends only on $\mathbf{c}(\mathcal{F})$.

This is proved in Section 3.1.

Finally, we use discrete Fourier analysis. We detect the condition $s_1 \not\equiv s_2 \,(\mathrm{mod}\, q)$ using additive characters:

$$1 - \frac{1}{q} \sum_{\lambda \,(\mathrm{mod}\, q)} e_q(\lambda(s_1 - s_2)) = \begin{cases} 1 & \text{if } s_1 \neq s_2 \text{ in } \mathbf{F}_q, \\ 0 & \text{otherwise.} \end{cases}$$

We further complete the sums over s_1 and s_2 using additive characters. For any $L\colon \mathbf{F}_q \to \mathbf{C}$, we define

$$\mathcal{C}(L,\lambda_1,\lambda_2,\boldsymbol{b}) = \sum_{r \,(\mathrm{mod}\, q)} \mathbf{R}(L,r,\lambda_1,\boldsymbol{b})\overline{\mathbf{R}(L,r,\lambda_2,\boldsymbol{b})},$$

where $\mathbf{R}(L,r,\lambda,\boldsymbol{b})$ is the sum defined in (2.8). Then let

$$\hat{\Sigma}(L,\boldsymbol{b},\lambda_1,\lambda_2,) = \mathcal{C}(L,\lambda_1,\lambda_2,\boldsymbol{b}) - \frac{1}{q}\sum_{\lambda \,(\mathrm{mod}\, q)} \mathcal{C}(L,\lambda_1+\lambda,\lambda_2+\lambda,\boldsymbol{b}).$$

Observing, as in (2.9), that for $c \in \mathbf{F}_q^\times$ we have

$$\hat{\Sigma}(K_c,\boldsymbol{b},\lambda_1,\lambda_2) = \hat{\Sigma}(K,\lambda_1/c,\lambda_2/c,\boldsymbol{b}),$$

the completion leads to the bound

$$\Sigma^{\neq}(K_c,\boldsymbol{b};AM) \ll (\log q)^2 \max_{\lambda_1,\lambda_2\in\mathbf{F}_q} |\hat{\Sigma}(K,\boldsymbol{b},\lambda_1,\lambda_2)|$$

for any $c \in \mathbf{F}_q^\times$, where the implied constant is absolute.

We must now assume as before that $\mathcal{F} = \mathcal{K}\ell_k$ is the Kloosterman sheaf of rank k with trace function $K = \mathrm{Kl}_k$. We will prove below our final bound:

THEOREM 2.6. *Let $k \geqslant 2$, and let $K = \mathrm{Kl}_k$. There exists a codimension 1 subvariety $\mathcal{V}^{\mathrm{bad}} \subset \mathbf{A}^4_{\mathbf{F}_q}$ containing $\mathcal{V}^\Delta$, with degree bounded independently of q, such that for any $\boldsymbol{b} \notin \mathcal{V}^{\mathrm{bad}}(\mathbf{F}_q)$ and every distinct λ_1, $\lambda_2 \in \mathbf{F}_q$, we have*

$$|\hat{\Sigma}(\mathrm{Kl}_k,\boldsymbol{b},\lambda_1,\lambda_2)| \ll q^{3/2}, \tag{2.14}$$

where the constant depends only on k.

This follows from Theorem 4.11 in Section 4. In fact, the subvariety $\mathcal{V}^{\mathrm{bad}}$ is the same as in Theorem 2.4.

Assuming these results, we conclude the proof of Theorem 1.1 in the same manner as in the previous section. For

$$\mathcal{B}^{\mathrm{bad}} = \mathcal{B} \cap \{\boldsymbol{b} \in \mathcal{B} \mid \boldsymbol{b} \,(\mathrm{mod}\, q) \in \mathcal{V}^{\mathrm{bad}}(\mathbf{F}_q)\}, \qquad \mathcal{B}^{\mathrm{gen}} = \mathcal{B}\backslash\mathcal{B}^{\mathrm{bad}},$$

we have the estimate $|\mathcal{B}^{\mathrm{bad}}| = O_k(B^3)$ since $\mathcal{V}^{\mathrm{bad}}$ has degree bounded independently of q.

We apply Theorem 2.6 for $\boldsymbol{b} \in \mathcal{B}^{\mathrm{gen}}$, the bound (2.13) of Lemma 2.5 for $\boldsymbol{b} \in \mathcal{B}^{\mathrm{bad}} - \mathcal{B}^\Delta$ and finally (2.12) for $\boldsymbol{b} \in \mathcal{B}^\Delta$. This gives

$$\sum_{\boldsymbol{b}} |\Sigma^{\neq}([\times c]^*\,\mathrm{Kl}_k,\boldsymbol{b};AM)| \ll (\log q)^2(B^4q^{3/2} + A^2B^3M^2q^{1/2} + A^2B^2M^2q),$$

where the implied constant depends only on k.

We select

$$A = q^{\frac{1}{8}}M^{-\frac{1}{2}}N^{\frac{1}{2}},\ B = q^{-\frac{1}{8}}M^{\frac{1}{2}}N^{\frac{1}{2}},$$

which satisfy (2.2) by (2.10). Then $AB = N$ and the first and third terms on the right-hand side are equal to $(MN)^2 q$. The second term is $(MN)^{\frac{5}{2}} q^{\frac{3}{8}} \leqslant (MN)^2 q$ by (2.10). Therefore we have

$$\sum_{\boldsymbol{b}} |\Sigma^{\neq}([\times c]^* \operatorname{Kl}_k, \boldsymbol{b}; AM)| \ll (MN)^2 q (\log q)^2,$$

and consequently we obtain from (2.11) the bound

$$\begin{aligned} |B([\times c]^* \operatorname{Kl}_k, \boldsymbol{\alpha}, \boldsymbol{\beta})|^2 &\ll \|\boldsymbol{\alpha}\|_2^2 \|\boldsymbol{\beta}\|_2^2 \Big(N + \frac{q^{\varepsilon}}{N} (AN)^{3/4} M^{1/2} q^{1/4} (MN)^{1/2} \Big) \\ &\ll q^{\varepsilon} \|\boldsymbol{\alpha}\|_2^2 \|\boldsymbol{\beta}\|_2^2 \Big(N + (MN)^{\frac{5}{8}} q^{\frac{11}{32}} \Big) \\ &\ll q^{\varepsilon} \|\boldsymbol{\alpha}\|_2^2 \|\boldsymbol{\beta}\|_2^2 MN \Big(M^{-1} + (MN)^{-\frac{3}{8}} q^{\frac{11}{32}} \Big), \end{aligned}$$

for any $\varepsilon > 0$, where the implied constant depends only on k and ε.

This concludes the proof of Theorem 1.1 modulo the proof of Lemma 2.5 and of Theorem 2.6.

Remark 2.7. As in [FM98] it is possible to apply the Hölder inequality that leads to (2.11) with higher exponent than $2l = 4$. Doing this leads to sums involving products of the shape

$$(r, s_1, s_2) \mapsto \prod_{i=1}^{l} K(s_1(r + b_i)) \overline{K}(s_2(r + b_i)) \overline{K(s_1(r + b_{i+l})) \overline{K}(s_2(r + b_{i+l}))}$$

for

$$\boldsymbol{b}_l = (b_1, \dots, b_l, b_{l+1}, \dots, b_{2l}) \in]B, 2B]^{2l}.$$

Except for heavier notational complexity, some of the arguments of this section (and of the next) do carry over and (assuming that (2.10) holds), one obtains for $l \geqslant 3$ and $MN \geqslant q^{7/8}$ the bound

$$|B(\operatorname{Kl}_k, \boldsymbol{\alpha}, \boldsymbol{\beta})|^2 \ll_{\varepsilon, k} q^{\varepsilon} \|\boldsymbol{\alpha}\|_2^2 \|\boldsymbol{\beta}\|_2^2 MN \Big(M^{-1} + (q^{l+4} (MN)^{-8})^{\frac{1}{4l(l+2)}} \Big).$$

This bound is only interesting when $l = 3$ and yields a nontrivial estimate in the range

$$MN \geqslant q^{\frac{7}{8} + \delta}, \qquad \delta > 0$$

compared with $MN \geqslant q^{\frac{11}{12} + \delta}$ in Remark 1.2.

In order for the Hölder inequality with higher exponents to give better estimates, one needs to improve the lower bound on the codimension of the variety $\mathcal{V}_l^{\mathrm{bad}} \subset \mathbf{A}^{2l}_{\mathbf{F}_q}$ in the corresponding generalization of Theorem 2.6. At the moment, we only know that this codimension is at least 1, but if one could prove that this variety has codimension 2, one could take $l = 5$ and obtain a nontrivial bound in the range $MN \geqslant q^{\frac{5}{6} + \delta}$.

The best possible result that might be achieved using this method would be if $\mathcal{V}_l^{\mathrm{bad}}$ had codimension l. This would lead to nontrivial bounds as long as

$$MN \geqslant q^{\frac{3l+5}{4(l+1)}+\delta}, \qquad \delta > 0.$$

Although we expect that the codimension of $\mathcal{V}_l^{\mathrm{bad}}$ is indeed l, this seems a difficult geometric problem when l is large. (Indeed, the method used in Section 4 do not seem to be sufficient, as they amount to "estimating" $\mathcal{V}_l$ by showing that it is a subvariety of another variety whose codimension we estimate, and one expects that the codimension of this auxiliary variety is exactly 1.)

By taking l very large, we thus see that the limit of the method is the range $MN \geqslant q^{\frac{3}{4}+\delta}$. Interestingly, this is the same range achieved in [FKM14, Th. 1.6] for the case where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are both smooth.

D. Bejleri, A. Christensen, B. Kadets, C.-Y. Hsu and Z. Yao have verified some of the claims in this remark while pursuing a research project during the 2016 Arizona Winter School.

3. **Bounds for complete exponential sums**

In this section we use methods from ℓ-adic cohomology to prove Lemmas 2.3 and 2.5, and we make the first steps towards Theorems 2.4 and 2.6. The proof of these last two theorems will be finished in Section 4.

All results in this section apply for bountiful sheaves (with respect to the Borel subgroup). Thus we fix a prime q and such a sheaf $\mathcal{F}$ on $\mathbf{A}^1_{\mathbf{F}_q}$. We denote by K the trace function of $\mathcal{F}$, and by $\mathrm{Sing}(\mathcal{F}) \subset \mathbf{P}^1(\bar{\mathbf{F}}_q)$ the set of ramification points of $\mathcal{F}$.

For any finite extension $\mathbf{F}_{q^d}/\mathbf{F}_q$, for $\boldsymbol{b} \in (\mathbf{F}_{q^d})^4$, $\lambda \in \mathbf{F}_{q^d}$, and $(r,s) \in \mathbf{F}_{q^d} \times \mathbf{F}_{q^d}$, we denote

$$\mathbf{K}(r,s,\lambda,\boldsymbol{b};\mathbf{F}_{q^d}) = \psi_{\mathbf{F}_{q^d}}(\lambda s) \prod_{i=1}^{2} K(s(r+b_i);\mathbf{F}_{q^d})\overline{K(s(r+b_{i+2});\mathbf{F}_{q^d})}.$$

For $d=1$, we write simply $\mathbf{K}(r,s,\lambda,\boldsymbol{b}) = \mathbf{K}(r,s,\lambda,\boldsymbol{b};\mathbf{F}_q)$.

3.1. *One variable bounds.* The next proposition is a restatement of Lemmas 2.3 and 2.5, where the variable $c \in \mathbf{F}_q^\times$ is taken to be equal to 1; since we may replace s by cs (resp. (s_1,s_2) by (cs_1,cs_2)), this implies the case where $c \in \mathbf{F}_q^\times$ is arbitrary.

PROPOSITION 3.1. *Assume $q \neq 2$. Let $\mathcal{V}^\Delta \subset \mathbf{A}^4_{\mathbf{F}_q}$ be the affine variety given in Definition 2.2. For all $\boldsymbol{b} = (b_1,b_2,b_3,b_4) \in {\mathbf{F}_q}^4 - \mathcal{V}^\Delta(\mathbf{F}_q)$ and for all s, s_1, $s_2 \in \mathbf{F}_q^\times$, with $s_1 \neq s_2$, we have*

$$\sum_{r \in \mathbf{F}_q} \mathbf{K}(r,s,0,\boldsymbol{b}) \ll q^{1/2} \tag{3.1}$$

and

$$\sum_{r\in\mathbf{F}_q} \mathbf{K}(r,s_1,0,\boldsymbol{b})\overline{\mathbf{K}(r,s_2,0,\boldsymbol{b})} \ll q^{1/2}, \tag{3.2}$$

where the constant implied depends only on the conductor of $\mathcal{F}$.

Proof. This follows from the techniques surveyed in [FKM15c]. Precisely, for fixed $s \in \mathbf{F}_q^\times$ and $\boldsymbol{b} \notin \mathcal{V}^\Delta(\mathbf{F}_q)$, the sum in (3.1) is of the type discussed in [FKM15c, Cor. 1.6] with $k = 4$, $h = 0$, the 4-tuple

$$\boldsymbol{\gamma} = (\gamma_{s,1}, \dots, \gamma_{s,4}) \in \mathrm{PGL}_2(\mathbf{F}_q)^4$$

such that

$$\gamma_{s,i} = \begin{pmatrix} s & sb_i \\ & 1 \end{pmatrix},\ i = 1, \dots, 4$$

and (if $\mathcal{F}$ is of SL-type) the 4-tuple

$$\boldsymbol{\sigma} = (\sigma_i)_{i=1,\dots,4} \in \mathrm{Aut}(\mathbf{C}/\mathbf{R})^4,$$

where

$$\sigma_1 = \sigma_2 = \mathrm{Id}_{\mathbf{C}}, \quad \sigma_3 = \sigma_4 = c, \quad c = \text{complex conjugation}.$$

If $\mathcal{F}$ is of Sp type, the fact that $\boldsymbol{b}$ is not contained in $\mathcal{V}^\Delta(\mathbf{F}_q)$ implies that the tuple $\boldsymbol{\gamma}$ is *normal* in the sense of [FKM15c, Def. 1.3].

Similarly, if $\mathcal{F}$ is of SL-type with $\mathrm{rank}(\mathcal{F}) = r \geqslant 3$, and $\boldsymbol{b} \notin \mathcal{V}^\Delta(\mathbf{F}_q)$, the pair of tuples (γ, σ) is *r-normal*, including with respect to the special involution $\xi_{\mathcal{F}}$ of $\mathcal{F}$, if the latter exists. Indeed, because $q \neq 2$, $\gamma_{s,i}\gamma_{s,j}^{-1}$ is not an involution unless $b_i = b_j$ and so can only be equal to $\xi_{\mathcal{F}}$ if $b_i = b_j$. This means that conditions (2) and (3) of [FKM15c, Def. 1.3] are equivalent in our situation. Thus the bound (3.1) follows from [FKM15c, Cor. 1.6].

We now consider the bound (3.2), with $h = 0$, $k = 8$, the 8-tuple

$$\boldsymbol{\gamma} = (\gamma_{s_1,1}, \dots, \gamma_{s_1,4}, \gamma_{s_2,1}, \dots, \gamma_{s_2,4})$$

and (in the SL-type case) the 8-tuple

$$\boldsymbol{\sigma} = (\mathrm{Id}_{\mathbf{C}}, \mathrm{Id}_{\mathbf{C}}, c, c, c, c, \mathrm{Id}_{\mathbf{C}}, \mathrm{Id}_{\mathbf{C}}).$$

For $s_1 \neq s_2$ and $\boldsymbol{b} \notin \mathcal{V}^\Delta(\mathbf{F}_q)$, the 8-tuple $\boldsymbol{\gamma}$ is normal for $\mathcal{F}$ of Sp_k-type while for $\mathcal{F}$ of SL_r-type with $r \geqslant 3$, the tuples $(\boldsymbol{\gamma}, \boldsymbol{\sigma})$ are again r-normal (also possibly with respect to the special involution $\xi_{\mathcal{F}}$, if it exists). Indeed, the fact that $s_1 \neq s_2$ implies that the multiplicities involved in checking [FKM15c, Def. 1.3] are either multiplicities from the 4-tuple associated to s_1, or from that associated to s_2, and we are reduced to the situation in (3.1). Hence we obtain (3.2) by [FKM15c, Cor. 1.6]. □

By definition, the bound (3.1) gives Lemma 2.3, and (3.2) gives Lemma 2.5.

Remark 3.2. In the case of hyper-Kloosterman sums ($K = \mathrm{Kl}_k$), the statements we use are special cases of the bounds stated in [FKM15c, Cors. 3.2, 3.3].

3.2. *Second moment computations.* We now consider second moment averages. These estimates will be used in the next section to prove irreducibility of various sheaves.

For any finite extension $\mathbf{F}_{q^d}/\mathbf{F}_q$, any $\boldsymbol{b} \in (\mathbf{F}_{q^d})^4$ and $r \in \mathbf{F}_{q^d}$, we define

$$\mathbf{R}(r, \lambda, \boldsymbol{b}; \mathbf{F}_{q^d}) = \sum_{s \in \mathbf{F}_{q^d}} \mathbf{K}(r, s, \lambda, \boldsymbol{b}; \mathbf{F}_{q^d}). \tag{3.3}$$

Note that, as a function of λ, this is the discrete Fourier transform of $s \mapsto \mathbf{K}(r, s, 0, \boldsymbol{b}; \mathbf{F}_{q^d})$.

Lemma 3.3. *Suppose that the b_i, $1 \leqslant i \leqslant 4$, are pairwise distinct in $\mathbf{F}_q$. For any $d \geqslant 1$, we have*

$$\frac{1}{(q^d)^2} \sum_{r, \lambda \in \mathbf{F}_{q^d}} \sum |\mathbf{R}(r, \lambda, \boldsymbol{b}; \mathbf{F}_{q^d})|^2 = q^d + O(q^{d/2}), \tag{3.4}$$

where the implied constant depends only on the conductor of $\mathcal{F}$.

If $\mathcal{F}$ is of SL*-type and admits the special involution*

$$\xi = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.5}$$

then we have

$$\frac{1}{(q^d)^2} \sum_{r, \lambda \in \mathbf{F}_{q^d}} \mathbf{R}(r, \lambda, \boldsymbol{b}; \mathbf{F}_{q^d}) \overline{\mathbf{R}(r, -\lambda, \boldsymbol{b}; \mathbf{F}_{q^d})} = O(q^{d/2}), \tag{3.6}$$

where the implied constant depends only on the conductor of $\mathcal{F}$.

Proof. We abbreviate simply $\psi = \psi_{\mathbf{F}_{q^d}}$ and $K(x) = K(x; \mathbf{F}_{q^d})$ in the computations. Opening the squares in the left-hand sides of (3.4) and (3.6) and averaging over λ, we obtain

$$\begin{aligned} q^{-d} &\sum_{r, s \in \mathbf{F}_{q^d}} |\mathbf{K}(r, s, 0, \boldsymbol{b}; \mathbf{F}_{q^d})|^2 \\ &= q^{-d} \sum_{r, s \in \mathbf{F}_{q^d}} \prod_{i=1}^{4} |K(s(r + b_i))|^2 \\ &= q^{-d} \sum_{\substack{r \in \mathbf{F}_{q^d} \\ r + b_i \neq 0,\ i = 1, \dots, 4}} \sum_{s \in \mathbf{F}_{q^d}} \prod_{i=1}^{4} |K(s(r + b_i))|^2 + O(1) \end{aligned}$$

and

$$q^{-d}\sum_{r,s\in\mathbf{F}_{q^d}}\mathbf{K}(r,s,0,\boldsymbol{b};\mathbf{F}_{q^d})\overline{\mathbf{K}(r,-s,0,\boldsymbol{b};\mathbf{F}_{q^d})}$$

$$=q^{-d}\sum_{r,s\in\mathbf{F}_{q^d}}\prod_{i=1}^{2}K(s(r+b_i))\overline{K(s(r+b_{i+2}))K(-s(r+b_i))}K(-s(r+b_{i+2}))$$

$$=q^{-d}\sum_{\substack{r\in\mathbf{F}_{q^d}\\ r+b_i\neq 0,\ i=1,\dots,4}}\sum_{s\in\mathbf{F}_{q^d}}\prod_{i=1}^{2}K(s(r+b_i))\overline{K(s(r+b_{i+2}))}$$
$$\times\overline{K(-s(r+b_i))}K(-s(r+b_{i+2}))+O(1)$$

respectively, where the implied constant depends only on the conductor of $\mathcal{F}$.

Since $\xi^*\mathcal{F}$ is geometrically isomorphic to the tensor product of the dual of $\mathcal{F}$ with a rank 1 sheaf $\mathcal{L}$, by assumption, it follows that $K(-x)=\chi(x)\overline{K(x)}$ for some function χ with $|\chi(x)|=1$ for all x such that $\mathcal{F}$ is lisse at x. Hence the last sum is equal to

$$q^{-d}\sum_{\substack{r\in\mathbf{F}_{q^d}\\ r+b_i\neq 0,\ i=1,\dots,4}}\sum_{s\in\mathbf{F}_{q^d}}L(s)\prod_{i=1}^{2}K(s(r+b_i))^2\overline{K(s(r+b_{i+2}))}^2+O(1),$$

where

$$L(s)=\prod_{i=1}^{2}\overline{\chi(s(r+b_i))}\chi(s(r+b_{i+2}))$$

is the trace function of a rank 1 sheaf. Using the relation $\xi^*\mathcal{F}\simeq\mathcal{F}^\vee\otimes\mathcal{L}$, we see that the conductor of $\mathcal{L}$ is bounded in terms of the conductor of $\mathcal{F}$ only.

We proceed to evaluate the sum over s using again [FKM15c]. (More precisely, the final estimates follow from the extension to all finite fields of these results, which is immediate.)

For each i, let

$$\gamma_{r+b_i}=\begin{pmatrix}r+b_i & 0\\ 0 & 1\end{pmatrix}.$$

In the Sp-type case, since the $r+b_i$ are pairwise distinct for $1\leqslant i\leqslant 4$, the 8-tuple

$$\boldsymbol{\gamma}=(\gamma_{r+b_1},\dots,\gamma_{r+b_4},\gamma_{r+b_1},\dots,\gamma_{r+b_4})$$

consists of four distinct pairs (γ,γ); by [FKM15c, Cor. 1.7 (1)], it follows that for each r distinct from the $-b_i$, for $1\leqslant i\leqslant 4$, we have

$$\sum_{s\in\mathbf{F}_{q^d}}\prod_{i=1}^{4}|K((r+b_i)s)|^2=q^d+O(q^{d/2}),$$

and summing over r gives (3.4).

In the SL-type case with $r = \mathrm{rank}(\mathcal{F}) \geqslant 3$, the components of the pair of 8-tuples

$$\boldsymbol{\gamma} = (\gamma_{r+b_1}, \dots, \gamma_{r+b_4}, \gamma_{r+b_1}, \dots, \gamma_{r+b_4}),$$
$$\boldsymbol{\sigma} = (\mathrm{Id}_{\mathbf{C}}, \mathrm{Id}_{\mathbf{C}}, \mathrm{Id}_{\mathbf{C}}, \mathrm{Id}_{\mathbf{C}}, c, c, c, c)$$

satisfy the final assumption of [FKM15c, Cor. 1.7 (2)], and hence

$$\sum_{s\in\mathbf{F}_{q^d}} \prod_{i=1}^{4} |K((r+b_i)s)|^2 = q^d + O(q^{d/2})$$

also follows if $r + b_i$ is nonzero for each i. We therefore derive (3.4) again.

Finally we prove (3.6): Recall we are in the SL-type with the special involution ξ as in (3.5) and with the pair of 8-tuples

$$\boldsymbol{\gamma} = (\gamma_{r+b_1}, \dots, \gamma_{r+b_4}, \gamma_{r+b_1}, \dots, \gamma_{r+b_4}),$$
$$\boldsymbol{\sigma} = (\mathrm{Id}_{\mathbf{C}}, \mathrm{Id}_{\mathbf{C}}, c, c, \mathrm{Id}_{\mathbf{C}}, \mathrm{Id}_{\mathbf{C}}, c, c).$$

This pair is r-normal with respect to ξ (because the multiplicity of any element in the tuple is either 0 or 2). Arguing as in the proof of [FKM15c, Th. 1.5] (pp. 20–21, loc. cit.), we deduce that for each r distinct from the $-b_i$ for $1 \leqslant i \leqslant 4$, we have

$$\sum_{s\in\mathbf{F}_{q^d}} L(s) \prod_{i=1}^{2} K(s(r+b_i))^2 \overline{K(s(r+b_{i+2}))}^2 \ll q^{d/2},$$

where the implied constant depends only on the conductor of $\mathcal{F}$. □

Finally, we consider one more averaging over the r and $\boldsymbol{b}$ variables in the case when $\lambda = 0$.

Lemma 3.4. *For any $d \geqslant 1$, we have*

$$\frac{1}{(q^d)^5} \sum_{(r,\boldsymbol{b})\in\mathbf{F}_{q^d}^5} \sum |\mathbf{R}(r, 0, \boldsymbol{b}; \mathbf{F}_{q^d})|^2 = q^d + O(q^{d/2}),$$

where the implied constant depends only on the conductor of $\mathcal{F}$.

Proof. By a change of variables, we see that the sum is given by

$$q^{-4d} \sum_{b_1,b_2,b_3,b_4} \sum \Big| \sum_{s\in\mathbf{F}_{q^d}^\times} \prod_{i=1}^{2} K(sb_i; \mathbf{F}_{q^d}) \overline{K(sb_{i+2}; \mathbf{F}_{q^d})} \Big|^2$$
$$= \sum_{s,s'\in\mathbf{F}_{q^d}^\times} |\mathcal{C}(K, s, s'; \mathbf{F}_{q^d})|^2 |\mathcal{C}(K, s', s; \mathbf{F}_{q^d})|^2,$$

where $\mathcal{C}(K, s, s'; \mathbf{F}_{q^d})$ denote the correlation sum

$$\begin{aligned}\mathcal{C}(K, s, s'; \mathbf{F}_{q^d}) &= q^{-d} \sum_{b \in \mathbf{F}_{q^d}} K(sb; \mathbf{F}_{q^d})\overline{K}(s'b; \mathbf{F}_{q^d}) \\ &= q^{-d} \sum_{b \in \mathbf{F}_{q^d}} K((s/s')b; \mathbf{F}_{q^d})\overline{K}(b; \mathbf{F}_{q^d}).\end{aligned}$$

By assumption, the sheaf $\mathcal{F}$ is geometrically irreducible and is such that $[\times(s/s')]^*\mathcal{F}$ is geometrically isomorphic to $\mathcal{F}$ if and only if $s = s'$. Therefore by the usual application of the Riemann Hypothesis (see Proposition 1.8), we have

$$\mathcal{C}(K, s, s'; \mathbf{F}_{q^d}) = \delta(s, s') + O(q^{-d/2}),$$

where the implied constant depends only on the conductor of $\mathcal{F}$. It follows that

$$\begin{aligned}\sum_{s,s' \in \mathbf{F}_{q^d}^\times} |\mathcal{C}(K, s, s'; \mathbf{F}_{q^d})|^2 |\mathcal{C}(K, s', s; \mathbf{F}_{q^d})|^2 &= q^d + O(q^{d-d/2}) + O(q^{2d-4d/2}) \\ &= q^d + O(q^{d/2}),\end{aligned}$$

where the implied constant depends only on the conductor of $\mathcal{F}$. □

4. **Irreducibility of sum-product transform sheaves**

The goal of this long section, which is the most difficult of the paper, is to prove Theorems 2.4 and 2.6. In the whole section, we fix a prime q and a nontrivial additive character ψ of $\mathbf{F}_q$. We also fix an integer $k \geqslant 2$. We will also assume that q is sufficiently large depending on k. In particular, unless stated otherwise, we always assume that

$$q > k \geqslant 2.$$

We first begin by outlining the argument. The 7-variable function $\mathbf{K}$ and its sum $\mathbf{R}$ associated to the trace function of a sheaf $\mathcal{F}$ are first interpreted as trace functions of suitable sheaves in Section 4.1. The goal is then to prove that various specializations of these sheaves, which we call *sum-product* sheaves, are geometrically irreducible. This we can do when $\mathcal{F}$ is a Kloosterman sheaf. To do so requires quite delicate properties of these sheaves, which are recalled in Section 4.2. It also requires some relatively general tools, which are stated for convenience in Section 4.3. The argument splits in two parts, depending on whether we specialize with $\lambda = 0$ or with $\lambda \neq 0$, and these are handled separately in Sections 4.4 and 4.5.

4.1. *Sum-product sheaves.* Let $\mathcal{F}$ be an $\overline{\mathbf{Q}}_\ell$-sheaf on $\mathbf{A}^1_{\mathbf{F}_q}$, lisse of rank k and pure of weight 0 on a dense open subset, and mixed of weight $\leqslant 0$ on $\mathbf{A}^1$. (Examples of this include the extension by zero of a lisse and pure sheaf from an open subset or the middle extension of a lisse and pure sheaf [Del80, Cor. 1.8.9].)

On the affine space $\mathbf{A}^7 = \mathbf{A}^2 \times \mathbf{A}^1 \times \mathbf{A}^4$, with coordinates denoted $(r, s, \lambda, \boldsymbol{b})$, we define the projection $p_{2,3} : \mathbf{A}^7 \longrightarrow \mathbf{A}^1$ by

$$p_{2,3}(r, s, \lambda, b_1, \dots, b_4) = \lambda s$$

and morphisms $f_i : \mathbf{A}^7 \longrightarrow \mathbf{A}^1$ for $1 \leqslant i \leqslant 4$ by

$$f_i(r, s, \lambda, b_1, \dots, b_4) = s(r + b_i). \tag{4.1}$$

Let $\mathcal{K}$ be the $\overline{\mathbf{Q}}_\ell$-sheaf on $\mathbf{A}^7$ defined by

$$\mathcal{K} = p_{2,3}^*\mathcal{L}_\psi \otimes \bigotimes_{i=1}^{2} (f_i^*\mathcal{F} \otimes f_{i+2}^*\mathcal{F}^\vee). \tag{4.2}$$

The sheaf $\mathcal{K}$ is a constructible $\overline{\mathbf{Q}}_\ell$-sheaf of rank k^4 on $\mathbf{A}^7$, pointwise mixed of weights $\leqslant 0$. It is lisse and pointwise pure of weight 0 on the dense open set $U_{\mathcal{K}}$ that is the complement of the union of the divisors given by the equations

$$\{s(r + b_i) = \mu\} \quad \text{for } \mu \in S_{\mathcal{F}} \text{ and } i = 1, \dots, 4,$$

where $S_{\mathcal{F}}$ is the set of ramification points of $\mathcal{F}$ in $\mathbf{A}^1$. The trace function of $\mathcal{K}$ is

$$t_{\mathcal{K}}(r, s, \lambda, \boldsymbol{b}) = \mathbf{K}(r, s, \lambda, \boldsymbol{b})$$

for $(r, s, \lambda, \boldsymbol{b}) \in U_{\mathcal{K}}(\mathbf{F}_q)$.

Now we consider the projection $\pi^{(2)} : \mathbf{A}^7 \longrightarrow \mathbf{A}^6$ given by

$$\pi^{(2)}(r, s, \lambda, \boldsymbol{b}) = (r, \lambda, \boldsymbol{b}),$$

and the compactly-supported higher-direct image sheaves $R^i\pi^{(2)}_!\mathcal{K}$. Since the fibers of $\pi^{(2)}$ are curves, these sheaves are zero unless $0 \leqslant i \leqslant 2$.

LEMMA 4.1. *Assume that the sheaf $\mathcal{F}$ is bountiful with respect to the Borel subgroup.*

(1) *For $0 \leqslant i \leqslant 2$, the sheaf $R^i\pi^{(2)}_!\mathcal{K}$ on $\mathbf{A}^6_{\mathbf{F}_q}$ is mixed of weights $\leqslant i$.*
(2) *Let $\mathcal{V}^\Delta$ be the subvariety of $\mathbf{A}^4$ given in Definition 2.2. The sheaves $R^0\pi^{(2)}_!\mathcal{K}$ and $R^2\pi^{(2)}_!\mathcal{K}$ are supported on $\mathbf{A}^1 \times \mathbf{A}^1 \times \mathcal{V}^\Delta$.*
(3) *For $(r, \lambda, \boldsymbol{b})$ such that $\boldsymbol{b} \notin \mathcal{V}^\Delta$, the geometric monodromy representation of the sheaf $\mathcal{K}_{r,\lambda,\boldsymbol{b}}$ does not contain a trivial subrepresentation on a dense open subset of $\mathbf{A}^1$ where it is lisse.*

Proof. The first part is an application of Deligne's main theorem [Del80, Th. 1]. For the second part, by the proper base change theorem, the stalk of $R^i\pi_!^{(2)}\mathcal{K}$ at $x = (r, \lambda, \boldsymbol{b}) \in \mathbf{A}^6$ is

$$H_c^i(\mathbf{A}^1_{\overline{\mathbf{F}}_q}, \mathcal{L}_{\psi(s\lambda)} \otimes \bigotimes_{i=1}^{2} [\times(r+b_i)]^*\mathcal{K} \otimes [\times(r+b_{i+2})]^*\mathcal{K}^\vee),$$

where s is the coordinate on $\mathbf{A}^1$.

This cohomology group vanishes for $i = 0$ and any x. For $i = 2$ and $x \notin \mathcal{V}^\Delta$, its vanishing is given by [FKM15c, Th. 1.5] using (only) the assumption that $\mathcal{F}$ is bountiful in the sense of our definition.

For the last part, we first consider a closed point $x = (r, \lambda, \boldsymbol{b})$. Then the vanishing of the stalk

$$H_c^2(\mathbf{A}^1_{\overline{\mathbf{F}}_q}, \mathcal{L}_{\psi(s\lambda)} \otimes \bigotimes_{i=1}^{2} [\times(r+b_i)]^*\mathcal{K} \otimes [\times(r+b_{i+2})]^*\mathcal{K}^\vee)$$

of $R^2\pi_!^{(2)}\mathcal{K}$ implies that the geometric monodromy representation of $\mathcal{K}_{r,\lambda,\boldsymbol{b}}$ has no trivial subrepresentation where it is lisse (by the co-invariant formula and the semisimplicity that holds because the sheaf is pure of weight 0). The statement then extends to all points by Pink's Specialization Theorem [Kat90, Th. 8.18.2]. □

The sheaf $R^1\pi_!^{(2)}\mathcal{K}$, which is mixed of weights at most 1, is almost the sheaf we want to understand. However, some cleaning-up is required to facilitate the later arguments. Precisely, recall (see [Del80, Th. 3.4.1 (ii)]) that a lisse sheaf that is mixed of weight $\leqslant w$ is an extension of a lisse sheaf that is pure of weight w by a mixed sheaf of weight $\leqslant w-1$. Thus the following definition makes sense:

Definition 4.2 (Sum-product sheaf). Let $\mathcal{F}$ be a bountiful sheaf on $\mathbf{A}^1_{\mathbf{F}_q}$, and let $\mathcal{K}$ be the sheaf (4.2) and $\mathcal{R} = R^1\pi_!^{(2)}\mathcal{K}$. Consider the stratification $(X_i)_{1\leqslant i\leqslant m}$ of $\mathbf{A}^6_{\mathbf{F}_q}$ such that

- X_1 is the maximal open subset of $\mathbf{A}^6$ on which $\mathcal{R}$ is lisse;
- for $i \geqslant 2$, X_i is the maximal open subset of $\mathbf{A}^6 \setminus (X_1 \cup \cdots \cup X_{i-1})$ on which $\mathcal{R}$ is lisse.

We define the *sum-product transform sheaf* $\mathcal{R}^*$ associated to $\mathcal{F}$ as the constructible sheaf given as the sum over X_i of the maximal quotient of $\mathcal{R}|X_i$ that is pure of weight 1 extended by zero to all of $\mathbf{A}^6_{\mathbf{F}_q}$, so that $\mathcal{R}^*|X_i$ is the maximal quotient of $\mathcal{R}|X_i$ pure of weight 1.

For any $(\lambda, \boldsymbol{b}) \in \mathbf{A}^5$, we denote by $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ the pullback of $\mathcal{R}^*$ to the affine line given by the morphism $r \mapsto (r, \lambda, \boldsymbol{b})$, and we call $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ a *specialized sum-product sheaf*.

By construction, the sum-product sheaf is punctually pure of weight 1. A first property of this sheaf is as follows:

PROPOSITION 4.3. *For any $d \geqslant 1$, we have*

$$\frac{1}{(q^d)^5} \sum_{(r,\boldsymbol{b})\in \mathbf{F}_{q^d}^5} \sum |t_{\mathcal{R}^*}(r,0,\boldsymbol{b};\mathbf{F}_{q^d})|^2 = q^d + O(q^{d/2}).$$

Proof. Since

$$t_{\mathcal{R}^*}(r,0,\boldsymbol{b};\mathbf{F}_{q^d}) = t_{\mathcal{R}}(r,0,\boldsymbol{b};\mathbf{F}_{q^d}) + O(1),$$

by construction, it is enough to prove that

$$\frac{1}{(q^d)^5} \sum_{(r,\boldsymbol{b})\in \mathbf{F}_{q^d}^5} \sum |t_{\mathcal{R}}(r,0,\boldsymbol{b};\mathbf{F}_{q^d})|^2 = q^d + O(q^{d/2}).$$

Let $\mathcal{V}^\Delta$ be the subvariety of $\mathbf{A}^4$ of Definition 2.2. We have

$$\sum_{(r,\boldsymbol{b})\in \mathbf{F}_{q^d}^5} \sum |\mathbf{R}(r,0,\boldsymbol{b};\mathbf{F}_{q^d})|^2 = \sum_{\substack{(r,\boldsymbol{b})\in \mathbf{F}_{q^d}^5 \\ \boldsymbol{b}\notin \mathcal{V}^\Delta(\mathbf{F}_{q^d})}} \sum |\mathbf{R}(r,0,\boldsymbol{b};\mathbf{F}_{q^d})|^2 + \sum_{\substack{(r,\boldsymbol{b})\in \mathbf{F}_{q^d}^5 \\ \boldsymbol{b}\in \mathcal{V}^\Delta(\mathbf{F}_{q^d})}} \sum |\mathbf{R}(r,0,\boldsymbol{b};\mathbf{F}_{q^d})|^2.$$

Since $\mathcal{V}^\Delta$ has codimension 2 and $\mathbf{R}(r,0,\boldsymbol{b};\mathbf{F}_{q^d}) \ll_k q^d$, the second sum is bounded by $\ll_k q^{5d}$. On the other hand, the first sum equals

$$\sum_{\substack{(r,\boldsymbol{b})\in \mathbf{F}_{q^d}^5 \\ \boldsymbol{b}\notin \mathcal{V}^\Delta(\mathbf{F}_{q^d})}} \sum |t_{\mathcal{R}}(r,0,\boldsymbol{b};\mathbf{F}_{q^d})|^2.$$

By the same argument, we get

$$\sum_{\substack{(r,\boldsymbol{b})\in \mathbf{F}_{q^d}^5 \\ \boldsymbol{b}\notin \mathcal{V}^\Delta(\mathbf{F}_{q^d})}} \sum |t_{\mathcal{R}}(r,0,\boldsymbol{b};\mathbf{F}_{q^d})|^2 = \sum_{(r,\boldsymbol{b})\in \mathbf{F}_{q^d}^5} \sum |t_{\mathcal{R}}(r,0,\boldsymbol{b};\mathbf{F}_{q^d})|^2 + O(q^{5d}),$$

and the result then follows from Lemma 3.4. □

The following lemma is a fairly standard one:

LEMMA 4.4. *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two lisse ℓ-adic sheaves on a smooth geometrically connected scheme $X/\mathbf{F}_q$. Assume that $\mathcal{F}_1$ and $\mathcal{F}_2$ are both pure of some weight w and that for any $d \geqslant 1$ and any $x \in X(\mathbf{F}_{q^d})$, we have*

$$t_{\mathcal{F}_1}(x;\mathbf{F}_{q^d}) = t_{\mathcal{F}_2}(x;\mathbf{F}_{q^d}) + O(q^{d(w-1)/2}),$$

where the implied constant is absolute. Then the semisimplifications of $\mathcal{F}_1$ and $\mathcal{F}_2$ are isomorphic.

Proof. By the Chebotarev density theorem, it suffices to prove that the trace functions of $\mathcal{F}_1$ and $\mathcal{F}_2$ actually coincide (see, e.g., [Lau87, Prop. 1.1.2.1]). After applying a suitable Tate twist, we may assume that $w = 0$. Let $d \geqslant 1$, and let $x \in X(\mathbf{F}_{q^d})$. Denote by (α_i) (resp. (β_j)) the (complex) eigenvalues of the Frobenius at x relative to $\mathbf{F}_{q^d}$ on $\mathcal{F}_1$ (resp. $\mathcal{F}_2$). By assumption, for any integer $k \geqslant 1$, we have

$$\sum_i \alpha_i^k = \sum_j \beta_j^k + O(q^{-k/2}).$$

We multiply this by z^k and sum over $k \geqslant 1$, getting

$$\sum_i \frac{\alpha_i z}{1-\alpha_i z} = \sum_j \frac{\beta_j z}{1-\beta_j z} + R(z),$$

where $R(z)$ is holomorphic for $|z| < q^{1/2}$. Comparing poles, we deduce that the α_i's are a permutation of the β_j's, hence the result. □

We deduce from this an important duality property.

LEMMA 4.5. *For $\boldsymbol{b} = (b_1, b_2, b_3, b_4) \in \mathbf{A}^4$, let $\tilde{\boldsymbol{b}} = (b_3, b_4, b_1, b_2)$. For any λ and $\boldsymbol{b} \notin \mathcal{V}^\Delta$, the arithmetic semisimplifications of $\mathcal{R}^{*\vee}_{\lambda,\boldsymbol{b}}$ and $\mathcal{R}^*_{-\lambda,\tilde{\boldsymbol{b}}}(1)$ are isomorphic on any dense open subset where $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ is lisse.*

Proof. Let U be a dense open subset where $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ is lisse. We will check that the sheaves $\mathcal{R}^{*,\vee}_{\lambda,\boldsymbol{b}}$ and $\mathcal{R}^*_{-\lambda,\tilde{\boldsymbol{b}}}(1)$, which are both pure of weight -1, satisfy the conditions of the previous lemma. Indeed, let $d \geqslant 1$ and $x \in U(\mathbf{F}_{q^d})$ be given. We observe that

$$\begin{aligned} t_{\mathcal{R}^*_{-\lambda,\tilde{\boldsymbol{b}}}}(x;\mathbf{F}_{q^d}) &= t_{\mathcal{R}_{-\lambda,\tilde{\boldsymbol{b}}}}(x;\mathbf{F}_{q^d}) + O(1) = -\mathbf{R}(x,-\lambda,\tilde{\boldsymbol{b}}) + O(1) \\ &= -\overline{\mathbf{R}(x,\lambda,\boldsymbol{b})} + O(1) = \overline{t_{\mathcal{R}_{\lambda,\boldsymbol{b}}}(x;\mathbf{F}_{q^d})} + O(1) \\ &= \overline{t_{\mathcal{R}^*_{\lambda,\boldsymbol{b}}}(x;\mathbf{F}_{q^d})} + O(1). \end{aligned}$$

Since $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ is pure of weight 1 on U, we have further that

$$t_{\mathcal{R}^{*\vee}_{\lambda,\boldsymbol{b}}}(x;\mathbf{F}_{q^d}) = \frac{1}{q^d}\overline{t_{\mathcal{R}^*_{\lambda,\boldsymbol{b}}}(x;\mathbf{F}_{q^d})} = \frac{1}{q^d} t_{\mathcal{R}^*_{-\lambda,\tilde{\boldsymbol{b}}}}(x;\mathbf{F}_{q^d}) + O(q^{-d}).$$

The conclusion now follows. □

4.2. *Properties of Kloosterman sheaves.* We will study the sum-product transform of the Kloosterman sheaves. We first summarize the basic properties of the Kloosterman sheaves, which were originally defined by Deligne.

PROPOSITION 4.6 (Kloosterman sheaves). *Let $q > 2$ be a prime number, $\ell \neq q$ an auxiliary prime number and ψ a nontrivial ℓ-adic additive character of $\mathbf{F}_q$. Let $k \geqslant 2$ be an integer.*

There exists a constructible $\overline{\mathbf{Q}}_\ell$-sheaf $\mathcal{K}\ell_k = \mathcal{K}\ell_{\psi,k}$ on $\mathbf{P}^1_{\mathbf{F}_q}$ with the following properties:

(1) *For any $d \geqslant 1$ and any $x \in \mathbf{G}_m(\mathbf{F}_{q^d})$, we have*

$$t_{\mathcal{K}\ell}(x; \mathbf{F}_{q^d}) = \mathrm{Kl}_k(x; \mathbf{F}_{q^d}) = \frac{(-1)^{k-1}}{q^{d(k-1)/2}} \sum_{x_1 \cdots x_k = x} \psi_{\mathbf{F}_{q^d}}(x_1 + \cdots + x_k).$$

(2) *The sheaf $\mathcal{K}\ell_{\psi,k}$ is lisse of rank k on $\mathbf{G}_m$.*
(3) *On $\mathbf{G}_m$, the sheaf $\mathcal{K}\ell_{\psi,k}$ is geometrically irreducible and pure of weight 0.*
(4) *The sheaf $\mathcal{K}\ell_{\psi,k}$ is tamely ramified at 0 with unipotent local monodromy with a single Jordan block.*
(5) *The sheaf $\mathcal{K}\ell_{\psi,k}$ is wildly ramified at ∞, with a single break equal to $1/k$, and with Swan conductor equal to 1.*
(6) *There is an arithmetic isomorphism*

$$\mathcal{K}\ell^\vee_{\psi,k} \simeq [x \mapsto (-1)^k x]^* \mathcal{K}\ell_{\psi,k}$$

and, in particular, $\mathcal{K}\ell_{\psi,k}$ is arithmetically self-dual if k is even.
(7) *If $k \geqslant 2$, then the arithmetic and geometric monodromy groups of $\mathcal{K}\ell_{\psi,k}$ are equal. If k is even, they are equal to Sp_k, and if k is odd, then they are equal to SL_k.*
(8) *The stalks of $\mathcal{K}\ell_{\psi,k}$ at 0 and ∞ both vanish.*
(9) *If $\gamma \in \mathrm{PGL}_2(\overline{\mathbf{F}}_q)$ is nontrivial, there does not exist a rank 1 sheaf $\mathcal{L}$ such that we have a geometric isomorphism over a dense open set*

$$\gamma^* \mathcal{K}\ell_{\psi,k} \simeq \mathcal{K}\ell_{\psi,k} \otimes \mathcal{L}.$$

Proof. All this is essentially *mise pour mémoire* from [Kat88]. The sheaf $\mathcal{K}\ell_k$ is the sheaf denoted $\mathrm{Kl}_n(\psi)((k-1)/2)$ in [Kat88, 11.0.2]; precisely, properties (1) to (5) are stated with references in [Kat88, 11.0.2], property (6) is found in [Kat88, Cors. 4.1.3, 4.1.4], and the crucial property (7) is [Kat88, Th. 11.1, Cor. 11.3]. The sheaf constructed in [Kat88] is on $\mathbf{G}_m$, and we extend by zero from $\mathbf{G}_m$ to $\mathbf{P}^1$, making property (8) true by definition. The last property is explained, e.g., in [FKM15c, §3, (b), (c)]. □

Remark 4.7. (1) As a matter of definition, one possibility is to define $\mathcal{K}\ell_{\psi,k}$ as k-fold (Tate-twisted) multiplicative convolution of the basic Artin–Schreier sheaf $\mathcal{L}_\psi$, namely,

$$\mathcal{K}\ell_k = (\mathcal{L}_\psi \star \cdots \star \mathcal{L}_\psi)((k-1)/2),$$

see [Kat88, 5.5].

(2) Katz has also shown (see [Kat88, Cor. 4.1.2]) that property (1) characterizes $\mathcal{K}\ell_{\psi,k}$ as a lisse sheaf on $\mathbf{G}_m$, up to arithmetic isomorphism.

(3) It might seem more natural to define $\mathcal{K}\ell_{\psi,k}$ as the middle extension from $\mathbf{G}_m$ to $\mathbf{P}^1$ of the sheaf constructed by Katz. However, the property of being a middle extension is not preserved by tensor product, so we would not be able to use directly any of the properties of middle-extension sheaves when studying the sum-product transform sheaves. On the other hand, having stalk zero is preserved by tensor product, and it will turn out that this property simplifies certain technical arguments.

COROLLARY 4.8. *For $k \geqslant 2$, the sheaf $\mathcal{K}\ell_{\psi,k}$ is bountiful with respect to the full group* PGL_2; *it is of* Sp*-type if k is even, and of* SL*-type if k is odd. In the second case, $\mathcal{K}\ell_{\psi,k}$ has the special involution $\left(\begin{smallmatrix}-1 & 0\\ 0 & 1\end{smallmatrix}\right)$. Moreover, the conductor of $\mathcal{K}\ell_{\psi,k}$ is bounded in terms of k only.*

Proof. This is clear from Proposition 4.6 using the definition of bountiful sheaves and of the conductor of a sheaf. □

For convenience, we will most often simply denote $\mathcal{K}\ell_k = \mathcal{K}\ell_{\psi,k}$ since we assume that ψ is fixed.

The following lemma computes precisely the local monodromy of $\mathcal{K}\ell_k$ at ∞. This is a special case of a formula of L. Fu [Fu10, Prop. 0.8] (which also describes $\mathcal{K}\ell_k$ as a representation of the decomposition group, not just the inertia group).

LEMMA 4.9. *Assume $q > k \geqslant 2$. Denote by $\tilde{\psi}$ the additive character $x \mapsto \psi(kx)$ of $\mathbf{F}_q$. Then, as representations of the inertia group $I(\infty)$ at ∞, there exists an isomorphism*

$$\mathcal{K}\ell_k \simeq [x \mapsto x^k]_*(\mathcal{L}_{\chi_2^{k+1}} \otimes \mathcal{L}_{\tilde{\psi}}),$$

where we recall that χ_2 is the unique nontrivial character of order 2 of $\mathbf{F}_q^\times$.

Proof. According to the remark in [Kat88, 10.4.5], we have an isomorphism

$$\mathcal{K}\ell_k \simeq [x \mapsto x^k]_*(\mathcal{L}_{\tilde{\psi}})$$

as a representation of the wild inertia subgroup $P(\infty) \subset I(\infty)$. On the other hand, by [Kat88, §1.18] and [Kat90, Th. 8.6.3], an $I(\infty)$-representation that is totally wild with Swan conductor 1 is determined, up to scaling, by its rank and its determinant. (That is, if two such representations π_1 and π_2 have same rank and determinant, then there exists a nonzero c such that $\pi_2 \simeq [\times c]^*\pi_1$.)

Since $\det \mathcal{K}\ell_k$ is trivial (see Proposition 4.6(7)), it is sufficient to check that the determinant of the $I(\infty)$-representation $[x \mapsto x^k]_*(\mathcal{L}_{\chi_2^{k+1}} \otimes \mathcal{L}_{\tilde{\psi}})$ is trivial. But for any multiplicative character χ, we have a geometric isomorphism

$$\det([x \mapsto x^k]_*(\mathcal{L}_\chi \otimes \mathcal{L}_{\tilde{\psi}})) \simeq \chi\chi_2^{k+1},$$

and this is geometrically trivial if $\chi = \chi_2^{k+1}$. (This follows, e.g., from the Hasse–Davenport relations as in [Kat88, Prop. 5.6.2] or from the block-permutation matrix representation of an induced representation, similarly to the argument that appears later in Lemma 4.15.) □

Finally, we can state our main theorem concerning the sum-product sheaves associated to Kloosterman sheaves.

THEOREM 4.10 (Irreducibility of sum-product sheaves). *Let $k \geqslant 2$ be an integer. Let ℓ be a prime $\neq q$, and let $\mathcal{R}^*$ be the ℓ-adic sum-product transform sheaf of $\mathcal{K}\ell_k$ over $\mathbf{F}_q$.*

If q is sufficiently large with respect to k, there exists a closed subset $\mathcal{V}^{\mathrm{bad}} \subset \mathbf{A}^4_{\mathbf{F}_q}$ containing $\mathcal{V}^{\Delta}$, of codimension 1 and of degree bounded independently of q, stable under the automorphism $(b_1, b_2, b_3, b_4) \mapsto (b_3, b_4, b_1, b_2)$, such that for all $\boldsymbol{b} = (b_1, b_2, b_3, b_4)$ not in $\mathcal{V}^{\mathrm{bad}}$, the following properties hold:

(1) *for all λ, the specialized sum-product sheaf $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ is lisse and geometrically irreducible on a dense open subset of $\mathbf{A}^1$;*
(2) *for all λ, there does not exist a dense open subset U of $\mathbf{A}^1$ such that $\mathcal{R}^*_{\lambda,\boldsymbol{b}}|U$ is geometrically trivial;*
(3) *if $\lambda \neq \lambda'$, then there does not exist a dense open subset U of $\mathbf{A}^1$ such that $\mathcal{R}^*_{\lambda,\boldsymbol{b}}|U$ is geometrically isomorphic to $\mathcal{R}^*_{\lambda',\boldsymbol{b}}|U$.*
(4) *for all $\lambda_1, \lambda_2, \boldsymbol{b}_1, \boldsymbol{b}_2$, the dimensions of the stalks of the sheaf $\mathcal{R}_{\lambda_i,\boldsymbol{b}_i}$ and the dimensions of the cohomology groups*

$$H^i_c(\mathbf{A}^1_{\overline{\mathbf{F}}_q}, \mathcal{R}_{\lambda_1,\boldsymbol{b}_1})$$

and

$$H^i_c(\mathbf{A}^1_{\overline{\mathbf{F}}_q}, \mathcal{R}_{\lambda_1,\boldsymbol{b}_1} \otimes \mathcal{R}_{\lambda_2,\boldsymbol{b}_2})$$

are bounded in terms of k only — in particular, independently of q for k fixed.

After some preliminaries, the proof splits into two cases: the case $\lambda = 0$ in Section 4.4 and the case $\lambda \neq 0$ in Section 4.5.

First, let us recall how this theorem implies our desired Theorems 2.4 and 2.6.

THEOREM 4.11. *Let $k \geqslant 2$ be an integer. Let $\mathbf{R}(r, \lambda, \boldsymbol{b})$ be the function on $\mathbf{A}^6(\mathbf{F}_q)$ defined in* (3.3). *For any $\boldsymbol{b} \in \mathbf{F}_q^4 - \mathcal{V}^{\mathrm{bad}}(\mathbf{F}_q)$ and any $\lambda, \lambda' \in \mathbf{F}_q$, we have*

$$\sum_{r \in \mathbf{F}_q} \mathbf{R}(r, \lambda, \boldsymbol{b}) \ll q,$$

$$\sum_{r \in \mathbf{F}_q} \mathbf{R}(r, \lambda, \boldsymbol{b})\overline{\mathbf{R}(r, \lambda', \boldsymbol{b})} = \delta(\lambda, \lambda')q^2 + O(q^{3/2}),$$

where the implied constant depends only on k.

Proof. It is sufficient to prove the theorem when q is sufficiently large with respect to k, since we can handle any finite set of primes by replacing the implied constant by a larger one using trivial bounds for the sums.

First of all, note that by the proper base change theorem and the Grothendieck–Lefschetz trace formula, we have

$$t_{\mathcal{R}}(r,\lambda,\boldsymbol{b}) = -\sum_{s\in\mathbf{F}_q} t_{\mathcal{K}}(r,s,\lambda,\boldsymbol{b}) = -\mathbf{R}(r,\lambda,\boldsymbol{b}) \tag{4.3}$$

for $\boldsymbol{b}\notin\mathcal{V}^{\Delta}(\mathbf{F}_q)$, where the implied constant depends only on k. Since $\mathcal{R}$ is mixed of weights $\leqslant 1$ and of rank bounded in terms of k only, we have

$$t_{\mathcal{R}}(r,\lambda,\boldsymbol{b}) \ll q^{1/2}$$

for $\boldsymbol{b}\notin\mathcal{V}^{\Delta}(\mathbf{F}_q)$.

We begin the proof of the second bound. Thus let $\boldsymbol{b}\in\mathbf{F}_q^4-\mathcal{V}^{\mathrm{bad}}(\mathbf{F}_q)$ (in particular, $\boldsymbol{b}\notin\mathcal{V}^{\Delta}(\mathbf{F}_q)$) and λ, $\lambda'\in\mathbf{F}_q$ be given. First, we have

$$\overline{\mathbf{R}(r,\lambda,\boldsymbol{b})} = \mathbf{R}(r,-\lambda,\tilde{\boldsymbol{b}}),$$

where $\boldsymbol{b}=(b_3,b_4,b_1,b_2)\in\mathbf{F}_q^4-\mathcal{V}^{\mathrm{bad}}(\mathbf{F}_q)$. Thus the relation (4.3) and the Grothendieck–Lefschetz trace formula imply that

$$\sum_{r\in\mathbf{F}_q}\mathbf{R}(r,\lambda,\boldsymbol{b})\overline{\mathbf{R}(r,\lambda',\boldsymbol{b})} = \sum_{i=0}^{2}(-1)^i\operatorname{Tr}\left(\mathrm{Fr}_q \mid H_c^i(\mathbf{A}^1_{\overline{\mathbf{F}}_q},\mathcal{R}_{\lambda,\boldsymbol{b}}\otimes\mathcal{R}_{-\lambda',\tilde{\boldsymbol{b}}})\right).$$

Let $\mathcal{F}=\mathcal{R}_{\lambda,\boldsymbol{b}}\otimes\mathcal{R}_{-\lambda',\tilde{\boldsymbol{b}}}$ and $\mathcal{F}^*=\mathcal{R}^*_{\lambda,\boldsymbol{b}}\otimes\mathcal{R}^*_{-\lambda',\tilde{\boldsymbol{b}}}$. Since $\mathcal{R}$ is mixed of weight $\leqslant 1$, the tensor product sheaf $\mathcal{F}$ is mixed of weight $\leqslant 2$, so the i-th compactly supported cohomology group with coefficient in $\mathcal{F}$ is mixed of weight $\leqslant i+2$ by Deligne's Theorem [Del80].

The dimension of these cohomology groups are bounded in terms of k only by Theorem 4.10 (4). Thus we have

$$\sum_{r\in\mathbf{F}_q}\mathbf{R}(r,\lambda,\boldsymbol{b})\overline{\mathbf{R}(r,\lambda',\boldsymbol{b})} = \operatorname{Tr}(\mathrm{Fr}_q \mid W_{\lambda,\lambda'}) + O(q^{3/2}),$$

where $W_{\lambda,\lambda'}$ is the subspace of weight 4 in $H_c^2(\mathbf{A}^1_{\overline{\mathbf{F}}_q},\mathcal{F}) = H_c^2(U_{\overline{\mathbf{F}}_q},\mathcal{F})$, and the implied constant depends only on k. (Here U is any dense open set where $\mathcal{F}$ is lisse.)

By definition we have a short exact sequence

$$0\longrightarrow\mathcal{G}\longrightarrow\mathcal{F}\longrightarrow\mathcal{F}^*\longrightarrow 0$$

of lisse sheaves on U where $\mathcal{G}$ is mixed of weights <2. Taking the long cohomology exact sequence and again applying Deligne's Theorem, we see that $W_{\lambda,\lambda'}\simeq W^*_{\lambda,\lambda'}$, where $W^*_{\lambda,\lambda'}$ is the subspace of weight 4 in $H_c^2(U_{\overline{\mathbf{F}}_q},\mathcal{F}^*)$.

By the coinvariant formula, we have

$$H_c^2(U_{\overline{\mathbf{F}}_q},\mathcal{F}^*) = (\mathcal{F}^*_{\bar{\eta}})_{\pi_1(U_{\overline{\mathbf{F}}_q})}(-1),$$

so it is sufficient to prove that the weight 2 part of the $\pi_1(U_{\overline{\mathbf{F}}_q})$-coinvariants of $\mathcal{F}^*$ has dimension $\delta(\lambda,\lambda')$ and that the action of Fr_q is multiplication by q when $\lambda=\lambda'$.

The sheaves $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ and $\mathcal{R}^*_{\lambda',\boldsymbol{b}}$ are geometrically irreducible by Theorem 4.10(1); in particular, they are arithmetically semisimple. By Lemma 4.5, we have arithmetic isomorphisms

$$\mathcal{R}^*_{-\lambda',\tilde{\boldsymbol{b}}} \simeq \mathcal{R}^{*\vee}_{\lambda',\boldsymbol{b}}(-1),\ \mathcal{F}^* \simeq \mathcal{R}^*_{\lambda,\boldsymbol{b}} \otimes \mathcal{R}^{*\vee}_{\lambda',\boldsymbol{b}}(-1)$$

on U. Again by geometric irreducibility of $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ and $\mathcal{R}^*_{\lambda',\boldsymbol{b}}$, the monodromy coinvariants of that tensor product is one-dimensional if $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ and $\mathcal{R}^*_{\lambda',\boldsymbol{b}}$ are geometrically isomorphic and is zero otherwise. By Theorem 4.10(3), the sheaves are geometrically isomorphic if and only if $\lambda=\lambda'$. In that later case the space of (geometric) coinvariants of $\mathcal{R}^*_{\lambda,\boldsymbol{b}}\otimes\mathcal{R}^{*\vee}_{\lambda',\boldsymbol{b}}$ is one-dimensional, generated by the trace, and Fr_q acts trivially on it; therefore Fr_q acts by multiplication by q on $\mathcal{F}^*$.

The argument for the first bound is similar but simpler. We work with the cohomology groups $H^i_c(\mathbf{A}^1_{\overline{\mathbf{F}}_q},\mathcal{R}_{\lambda,\boldsymbol{b}})$, which are mixed of weights $\leqslant i+1$. It is sufficient to show that the weight 3 part of $H^2_c(\mathbf{A}^1_{\overline{\mathbf{F}}_q},\mathcal{R}_{\lambda,\boldsymbol{b}})$ vanishes, and thus it is sufficient to show that the weight 1 part of the monodromy coinvariants of $\mathcal{R}_{\lambda,\boldsymbol{b}}$ vanishes. Because $\mathcal{R}^*$ is the weight 1 part of $\mathcal{R}$, this is the same as showing that the monodromy coinvariants of $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ vanish. But $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ is irreducible and nontrivial as a monodromy representation, by Theorem 4.10(2) (3), so it has no coinvariants. □

4.3. *Preliminaries.* In this section we collect a number of results and definitions that we will use in the proof of our results. At first reading, it might be easier to only survey the statements before going to the next section.

We will derive the irreducibility statement of Theorem 4.10 for $\lambda\neq 0$ from the second of the following criteria:

LEMMA 4.12. *Let X_0 and Y_0 be normal varieties over $\mathbf{F}_q$. Let $f: Y_0\to X_0$ be a smooth proper morphism whose fibers are curves and whose geometric fibers are geometrically connected. Let $D_0\subset Y_0$ be a divisor. Write X, Y and D for the corresponding varieties over $\overline{\mathbf{F}}_q$.*

For a lisse $\overline{\mathbf{Q}}_\ell$-sheaf $\mathcal{F}$ on Y_0-D_0, consider the three following conditions:

(1) *the sheaf $\mathcal{F}$ is geometrically irreducible and pure of some weight*;
(2) *for the generic point η of X, there exists a point z of D_η defined over the function field $\kappa(\eta)$ of X such that there exists an irreducible component of multiplicity 1 of the restriction of the monodromy representation of $\mathcal{F}_\eta$ to the inertia group at z whose isomorphism class is preserved by the action of the Galois group of $\kappa(\eta)$ by conjugation on representations of the inertia group*;

(3) *the divisor D is finite and flat over X, and the function*

$$x \mapsto \sum_{y \in Y_x - D_x} (\mathrm{Swan}_y(\mathcal{F} \otimes \mathcal{F}^\vee) + \mathrm{rank}(\mathcal{F} \otimes \mathcal{F}^\vee))$$

is locally constant on X.

Then the following statements are true:

(a) *if* (1) *and* (2) *hold, then for all x in a dense open subset of X, the restriction $\mathcal{F}_x = \mathcal{F}|(Y_x - D_x)$ to a fiber $Y_x - D_x$ is geometrically irreducible*;

(b) *if* (1), (2) *and* (3) *hold, then for all x in X, the restriction $\mathcal{F}|(Y_x - D_x)$ to a fiber $Y_x - X_x$ is geometrically irreducible.*

Proof. We assume that conditions (1) and (2) hold. Let η' be the generic point of Y. By [GR71, V, Prop. 8.2], the natural homomorphism $\pi_1(\eta') \longrightarrow \pi_1(Y - D)$ is surjective. Since it factors through the natural homomorphism

$$\pi_1(Y_\eta - D_\eta) \to \pi_1(Y - D),$$

it follows that the latter is also surjective. In particular, condition (1) shows that the restriction of $\mathcal{F}$ to $Y_\eta - D_\eta$ corresponds to an irreducible representation of $\pi_1(Y_\eta - D_\eta)$. Thus $\mathcal{F}_\eta = \mathcal{F}|(Y_\eta - D_\eta)$ is an irreducible lisse sheaf on $Y_\eta - D_\eta$.

Consider now a geometric point $\bar{\eta}$ over η, the geometric fibers $Y_{\bar{\eta}}$ and $D_{\bar{\eta}}$ and the pullback $\mathcal{F}_{\bar{\eta}}$ of $\mathcal{F}_\eta$ to $(Y - D)_{\bar{\eta}}$. We will show that condition (2) implies that $\mathcal{F}_{\bar{\eta}}$ is irreducible.

Indeed, the representation of $\pi_1(Y_{\bar{\eta}} - D_{\bar{\eta}})$ corresponding to $\mathcal{F}_{\bar{\eta}}$ is semisimple because it is the restriction to a normal subgroup of an irreducible, hence semisimple, representation. Let

$$\mathcal{F}_{\bar{\eta}} = \bigoplus_{i \in I} n_i V_i$$

be a decomposition of this representation of $\pi_1(Y_{\bar{\eta}} - D_{\bar{\eta}})$ into irreducible subrepresentations, where $n_i \geqslant 1$ and the V_i are pairwise nonisomorphic. The quotient

$$G = \pi_1(Y_\eta - D_\eta)/\pi_1(Y_{\bar{\eta}} - D_{\bar{\eta}})$$

is isomorphic to the Galois group of the function field $\kappa(\eta)$ of X since f has geometrically connected generic fiber. It acts on the set $\{V_i\}$ of irreducible subrepresentations of $\mathcal{F}_{\bar{\eta}}$. Since $\mathcal{F}_\eta$ is an irreducible representation of $\pi_1(Y_\eta - D_\eta)$, this action is transitive. Hence, for any point y of D_η defined over $\kappa(\eta)$, the restriction of $\mathcal{F}_\eta$ to the inertia group at y has the property that it is a direct sum of $n = \sum n_i$ subrepresentations that are G-conjugates (but not necessarily irreducible or even indecomposable). In particular, any irreducible subrepresentation of the inertia group whose isomorphism class is fixed by G appears with multiplicity divisible by n. By condition (2), this means that $n = 1$, so that $\mathcal{F}_{\bar{\eta}}$ is irreducible.

By Pink's Specialization Theorem (see [Kat90, Th. 8.18.2]), it follows that $\mathcal{F}_x$ is geometrically irreducible for all x in some dense open subset, which gives (a).

Now assume further that condition (3) holds. For a closed point $x \in X$, the fiber $\mathcal{F}_x$ is geometrically irreducible if and only if the cohomology group $H^2_c((Y-D)_{x,\overline{\mathbf{F}}_q}, \mathcal{F}_x \otimes \mathcal{F}_x^\vee)$ is one-dimensional, by the coinvariant formula for the second cohomology group on a curve (see, e.g., [Kat88, 2.0.4]) and the fact that $\mathcal{F}_x$, being pure by condition (1), is geometrically semisimple (see [Del80, Th. 3.4.1(iii)]). Equivalently, by the proper base change theorem, the specialized sheaf $\mathcal{F}_x$ is geometrically irreducible if and only if the stalk of $R^2 f_!(\mathcal{F} \otimes \mathcal{F}^\vee)$ at x is one-dimensional. Condition (3) and Deligne's semicontinuity theorem [Lau81, Cor. 2.1.2] imply that the sheaf $R^2 f_!(\mathcal{F} \otimes \mathcal{F}^\vee)$ is lisse on X. Since it has rank 1 at all closed points in an open set, by what we proved before, it has rank 1 on all of X, which means that $\mathcal{F}_x$ is geometrically irreducible for all closed points x in X. By Pink's Specialization Theorem (see [Kat90, Th. 8.18.2]), $\mathcal{F}_x$ is geometrically irreducible for all points in X. □

Remark 4.13. Our proof of condition (1) below generalizes to quite general (bountiful) sheaves, but the proofs of conditions (2) and (3) involve careful calculations that depend on specific properties of the Kloosterman sheaves. This means that our results do not easily generalize to other sheaves.

However, condition (2) is a "generic" condition that should hold for a "random" sheaf. Thus it should be possible to prove it in a number of different concrete cases. The last condition (3) is more subtle; although it is always true on a dense open subset (hence is generic in that sense), the closed complement where it fails will usually have codimension 1. However, it should often be possible to explicitly compute that subset and to use this information for further study; for instance, cf. Remark 2.7.

In this paper, we will only use the second criterion of Lemma 4.12 in the proof of Theorem 4.10 to show that for all $\boldsymbol{b}$ outside of a proper subvariety, the specialized sheaves $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ are geometrically irreducible for *every* nonzero λ. However, the first criterion might be useful in other applications. (In the first draft of this paper, we used it to deal with sum-product sheaves where $\lambda = 0$, but we later found a simpler argument to deal with this case.)

To verify the first condition of the lemma, we will use Katz's diophantine criterion for geometric irreducibility (compare [Kat96, Lemma 7.0.3]).

LEMMA 4.14 (Diophantine criterion for irreducibility). *Let Y be a normal variety over $\mathbf{F}_q$, $U \subset Y$ a dense open subset and $\mathcal{F}$ a sheaf on Y that is lisse on U. Assume, moreover, that $\mathcal{F}|U$ is pure of some weight w and that $\mathcal{F}$ is*

mixed of weights $\leqslant w$ on Y. Then $\mathcal{F}|U$ is geometrically irreducible if

$$\frac{1}{q^{d\dim Y}}\sum_{y\in Y(\mathbf{F}_{q^d})}|t_{\mathcal{F}}(y)|^2 = q^{dw}(1+o(1))$$

as d tends to infinity.

Proof. Using a Tate twist, we may assume that $w = 0$. Let n be the dimension of Y and $D = Y - U$. We have

$$\frac{1}{q^{nd}}\sum_{y\in Y(\mathbf{F}_{q^d})}|t_{\mathcal{F}}(y)|^2 = \frac{1}{q^{nd}}\sum_{y\in U(\mathbf{F}_{q^d})}|t_{\mathcal{F}}(y)|^2 + \frac{1}{q^{nd}}\sum_{y\in D(\mathbf{F}_{q^d})}|t_{\mathcal{F}}(y)|^2.$$

The second sum is bounded by $O(q^{-d}) = o(1)$ using our assumption on the weights of $\mathcal{F}$ on Y (and the reduction to $w = 0$), and hence the assumption implies that

$$\frac{1}{q^{nd}}\sum_{y\in U(\mathbf{F}_{q^d})}|t_{\mathcal{F}}(y)|^2 \to 1$$

as $d \to +\infty$. On the other hand, the Grothendieck–Lefschetz Trace Formula and the Riemann Hypothesis imply that

$$\sum_{y\in U(\mathbf{F}_{q^d})}|t_{\mathcal{F}}(y)|^2 = \mathrm{Tr}(\mathrm{Fr}_{\mathbf{F}_{q^d}} \mid H_c^{2n}(Y_{\overline{\mathbf{F}}_q}, \mathcal{F}\otimes\mathcal{F}^\vee)) + O(q^{d(n-1/2)}),$$

and therefore

$$\frac{1}{q^{nd}}\sum_{y\in U(\mathbf{F}_{q^d})}|t_{\mathcal{F}}(y)|^2 = \mathrm{Tr}(\mathrm{Fr}_{\mathbf{F}_{q^d}} \mid H_c^{2n}(Y_{\overline{\mathbf{F}}_q}, \mathcal{F}\otimes\mathcal{F}^\vee)(n)) + o(1).$$

By the semisimplicity of $\mathcal{F}$ (see [Del80, Th. 3.4.1 (iii)]) and the coinvariant formula

$$H_c^{2n}(Y_{\overline{\mathbf{F}}_q}, \mathcal{F}\otimes\mathcal{F}^\vee) \simeq (\mathcal{F}\otimes\mathcal{F}^\vee)_{\pi(U_{\overline{\mathbf{F}}_q})}(-n),$$

we deduce by combining these formulas that the geometric invariant subspace of $\mathcal{F}\otimes\mathcal{F}^\vee$ is one-dimensional, which by Schur's Lemma means that $\mathcal{F}$ is geometrically irreducible. □

We will use the following lemma from elementary representation theory to describe the local monodromy of tensor products of Kloosterman sheaves.

Lemma 4.15. *Let G be a group and E an arbitrary field. Let H be a normal subgroup of G of finite index. Consider the usual action $\sigma\cdot V = \sigma(V)$ of G/H on E-representations of H, where $x\in H$ acts on $\sigma(V)$ by the action of $\sigma^{-1}x\sigma$ on V.*

For any finite-dimensional E-representations $V_1, \dots, V_n$ of H, we have a canonical isomorphism

$$\bigotimes_{i=1}^{n} \operatorname{Ind}_H^G V_i \simeq \bigoplus_{(\sigma_2,\dots,\sigma_n)\in (G/H)^{n-1}} \operatorname{Ind}_H^G\Big(V_1 \otimes \bigotimes_{i=2}^{n} \sigma_i(V_i)\Big).$$

Proof. We proceed by induction on n. The case $n = 1$ is a tautology. For $n = 2$, we need to prove that

$$\operatorname{Ind}_H^G V_1 \otimes \operatorname{Ind}_H^G V_2 \simeq \bigoplus_{\sigma\in G/H} \operatorname{Ind}_H^G \left(V_1 \otimes \sigma(V_2)\right).$$

To see this, first apply the projection formula

$$\operatorname{Ind}_H^G(V_1 \otimes \operatorname{Res}_G^H \operatorname{Ind}_H^G V_2) = \operatorname{Ind}_H^G V_1 \otimes \operatorname{Ind}_H^G V_2$$

and then the fact that

$$\operatorname{Res}_G^H \operatorname{Ind}_H^G V_2 = \bigoplus_{\sigma\in G/H} \sigma(V_2),$$

which follows from the definition of induction. (See, e.g., [Kow14, Props. 2.3.15, 2.3.18] for these standard facts.)

We easily complete the proof for $n \geqslant 3$ by induction using the case $n=2$. □

As a corollary, we now obtain the local monodromy at infinity for the sheaves $\mathcal{K}_{r,\lambda,\boldsymbol{b}}$. To state the result, we recall from the introduction the notation $\mathcal{L}_\psi(cs^{1/k})$ for a variety $X/\mathbf{F}_q$, an integer $k \geqslant 1$ and a function c on X: this is the sheaf on $X \times \mathbf{A}^1$ (with coordinates (x, s)) given by

$$\mathcal{L}_\psi(cs^{1/k}) = \alpha_* \mathcal{L}_{\psi(c(x)t)},$$

where α is the map

$$\begin{cases} X \times \mathbf{A}^1 \to X \times \mathbf{A}^1, \\ (x, t) \mapsto (x, t^k). \end{cases}$$

Lemma 4.16. *Assume $q > k \geqslant 2$, and denote by $\tilde{\psi}$ the character $x \mapsto \psi(kx)$. Fix $r, \boldsymbol{b}, \lambda$ such that $r + b_i \neq 0$ for all i. Let $(r + b_i)^{1/k}$ be a fixed k-th root of $r + b_i$ in $\overline{\mathbf{F}}_q$.*

Then the local monodromy at $s = \infty$ of $\mathcal{K}_{r,\lambda,\boldsymbol{b}}$ is isomorphic to the local monodromy at $s = \infty$ of the sheaf

$$\mathcal{L}_{\psi(\lambda s)} \otimes \bigoplus_{\zeta_2,\zeta_3,\zeta_4\in \boldsymbol{\mu}_k} \mathcal{L}_{\tilde{\psi}}\Big(\Big((r + b_1)^{1/k} + \zeta_2(r + b_2)^{1/k} - \zeta_3(r + b_3)^{1/k} - \zeta_4(r + b_4)^{1/k}\Big)s^{1/k}\Big), \tag{4.4}$$

where $\boldsymbol{\mu}_k$ is the group of k-th roots of unity in $\overline{\mathbf{F}}_q$.

More generally, for fixed λ and $\boldsymbol{b}$, for any algebraic variety $U_{\mathbf{F}_q}$, let $f : U \longrightarrow \mathbf{A}^1 - \{-b_1, \dots, -b_4\}$ be a morphism, and assume there are morphisms $r_i : U \longrightarrow \mathbf{A}^1$ such that $[x \mapsto x^k] \circ r_i = [x + b_i] \circ f$. Assume that k is odd or that there exist a constant c and a function g on U such that $r_1 r_2 r_3 r_4 = cg^2$. Then the local monodromy of the sheaf $(f \times \mathrm{Id})^* \mathcal{K}_{\lambda,\boldsymbol{b}}$ on $U \times \mathbf{A}^1$ along the divisor $U \times \{\infty\}$ is isomorphic to the local monodromy of the sheaf

$$\mathcal{L}_{\psi(\lambda s)} \otimes \bigoplus_{\zeta_2,\zeta_3,\zeta_4 \in \boldsymbol{\mu}_k} \mathcal{L}_{\tilde{\psi}} \left((r_1 + \zeta_2 r_2 - \zeta_3 r_3 - \zeta_4 r_4)\, s^{1/k} \right)$$

along $U \times \{\infty\}$

Proof. We have

$$\mathcal{K}_{r,\lambda,\boldsymbol{b}} = \mathcal{L}_{\psi(\lambda s)} \otimes \bigotimes_{i=1}^{2} [\times(r + b_i)]^* \mathcal{K}\ell_k \otimes [\times(r + b_{i+2})]^* \mathcal{K}\ell_k^\vee,$$

so that it is enough to treat the case $\lambda = 0$. Furthermore, the first statement is the special case of the second where U is a single point (the second assumption holds with $c = r_1 r_2 r_3 r_4$, $g = 1$), so it is enough to handle the second case. By definition, we have

$$\begin{aligned}(f \times \mathrm{Id})^* \mathcal{K}_{\lambda,\boldsymbol{b}} &= (f \times \mathrm{Id})^* \bigotimes_{i=1}^{2} (f_i^* \mathcal{K}\ell_k \otimes f_{i+2}^* \mathcal{K}\ell_k^\vee) \\ &= \bigotimes_{i=1}^{2} \left((f \times \mathrm{Id})^* f_i^* \mathcal{K}\ell_k \otimes (f \times \mathrm{Id})^* f_{i+1}^* \mathcal{K}\ell_k^\vee \right),\end{aligned}$$

where f_i is the map $(r, s) \mapsto s(r + b_i)$.

Let $\alpha : \mathbf{A}^1 \to \mathbf{A}^1$ be the morphism $t \mapsto t^k$. By Lemma 4.9, the local monodromy of $\mathcal{K}\ell_k$ at ∞ is $\alpha_*(\mathcal{L}_{\chi_2^{k+1}} \otimes \mathcal{L}_{\tilde{\psi}})$.

Let $V = \mathbf{A}^1 - \{-b_1, \dots, -b_4\}$. For each i, we have the Cartesian diagram

$$\begin{array}{ccccc} U \times \mathbf{A}^1 & & \xrightarrow{\quad (u,t) \mapsto r_i(u)t \quad} & & \mathbf{A}^1 \\ \downarrow{\scriptstyle \mathrm{Id}_U \times \alpha} & & & & \downarrow{\scriptstyle \alpha} \\ U \times \mathbf{A}^1 & \xrightarrow{f \times \mathrm{Id}_{\mathbf{A}^1}} & V \times \mathbf{A}^1 & \xrightarrow{f_i} & \mathbf{A}^1. \end{array}$$

By proper base change, this implies that the local monodromy at ∞ of $(f \times \mathrm{Id})^* f_i^* \mathcal{K}\ell_k$ is the same as the local monodromy at ∞ of

$$(\mathrm{Id}_U \times \alpha)_* \left(\mathcal{L}_{\chi_2^{k+1}}(r_i t) \otimes \mathcal{L}_{\tilde{\psi}}(r_i t) \right).$$

In terms of representation theory, this means that the local monodromy representation at ∞ is induced from the normal subgroup H of $G = \pi_1((U \times \mathbf{G}_m)_{\overline{\mathbf{F}}_q})$ corresponding to the covering $\mathrm{Id}_U \times \alpha$ (which we will simply denote α by slight abuse of notation).

The quotient group G/H is naturally isomorphic to the Galois group of the covering, which is isomorphic to $\boldsymbol{\mu}_k$ by the homomorphism sending a root of unity $\zeta \in \boldsymbol{\mu}_k$ to the maps $(s,t) \mapsto (s, \zeta t)$. One checks easily that the action of ζ on representations of H is given by

$$\zeta \cdot \mathcal{L}_{\chi_2^{k+1}} = \mathcal{L}_{\chi_2^{k+1}}, \qquad \zeta \cdot \mathcal{L}_{\tilde{\psi}} = [\times\zeta]^* \mathcal{L}_{\tilde{\psi}}.$$

Hence by Lemma 4.15, the local monodromy at ∞ of $\mathcal{K}_{r,0,\boldsymbol{b}}$ is isomorphic to that of

$$\begin{aligned}\bigoplus_{\zeta_2,\zeta_3,\zeta_4 \in \boldsymbol{\mu}_k} \alpha_*\Big(&\mathcal{L}_{\chi_2^{k+1}(r_1 t)} \otimes \mathcal{L}_{\tilde{\psi}(r_1 t)} \otimes \mathcal{L}_{\chi_2^{k+1}(r_2 t)} \otimes \mathcal{L}_{\tilde{\psi}(\zeta_2 r_2 t)} \\ &\otimes \mathcal{L}_{\chi_2^{k+1}(r_3 t)} \otimes \mathcal{L}_{\tilde{\psi}(-\zeta_3 r_3 t)} \otimes \mathcal{L}_{\chi_2^{k+1}(r_4 t)} \otimes \mathcal{L}_{\tilde{\psi}(-\zeta_4 r_4 t)}\Big) \\ \simeq \bigoplus_{\zeta_2,\zeta_3,\zeta_4 \in \boldsymbol{\mu}_k} \alpha_*\Big(&\mathcal{L}_{\chi_2^{k+1}}(r_1 r_2 r_3 r_4 t^4) \mathcal{L}_{\tilde{\psi}}(r_1 t + \zeta_2 r_2 t - \zeta_3 r_3 t - \zeta_4 r_4 t)\Big).\end{aligned}$$

If k is odd, then χ_2^{k+1} is trivial. Otherwise $\chi_2^{k+1} = \chi_2$. Since $r_1, \dots, r_4$ are nonvanishing on U, the sheaf $\mathcal{L}_{\chi_2}(r_1 r_2 r_3 r_4 t^4)$ is lisse on $U \times \mathbf{G}_m \subseteq U \times \mathbf{A}^1$. By assumption, we have $r_1 r_2 r_3 r_4 = cg^2$, so $r_1 r_2 r_3 r_4 t^4 = c(gt^2)^2$, and thus $\mathcal{L}_{\chi_2}(r_1 r_2 r_3 r_4 t^4)$ is geometrically trivial on $U \times \mathbf{G}_m$. Therefore we may ignore that term, and we obtain (4.4). □

The following lemma about Kloosterman sheaves will prove useful to compute the monodromy at $r = \infty$ of sum-product sheaves.

LEMMA 4.17. *Let R be a strictly Henselian regular local ring of characteristic $q > 2$ with fraction field K and maximal ideal $\mathfrak{m}$. Assume that $q \nmid k$.*

(1) *If $a \in R - \{0\}$ and $b \in \mathfrak{m}$, then we have*

$$a^* \mathcal{K}\ell_k \simeq (a + ab)^* \mathcal{K}\ell_k,$$

where we view a and $a + ab$ as maps $\mathrm{Spec}(R) \longrightarrow \mathbf{A}^1_{\overline{\mathbf{F}}_q}$.

(2) *If $a \in K^\times$ is such that $a^{-1} \in \mathfrak{m}$, and $b \in R$, then we have*

$$a^* \mathcal{K}\ell_k \simeq (a + b)^* \mathcal{K}\ell_k,$$

where we view a and $a + b$ as maps $\mathrm{Spec}(R) \longrightarrow \mathbf{P}^1_{\overline{\mathbf{F}}_q}$.

Proof. (1) There are two cases: either $a \in \mathfrak{m}$ or $a \in R^\times$. If $a \in \mathfrak{m}$, we first observe that as $1 + b \in R^\times$, the ideals (a) and $(a + ab)$ are the same, and hence

$$Z = a^{-1}(\{0\}) = (a + ab)^{-1}(\{0\}) \subset \mathrm{Spec}(R).$$

Let U be the open complement of Z in $\mathrm{Spec}(R)$. Let j be the open immersion $U \to \mathrm{Spec}\, R$. As $\mathcal{K}\ell_k$ is zero at 0 according to our definition, both $a^*\mathcal{K}\ell_k$ and $(a + ab)^*\mathcal{K}\ell_k$ are zero on Z. Thus $a^*\mathcal{K}\ell_k$ is the extension by zero of $j^* a^* \mathcal{K}\ell_k$, and $(a + ab)^*\mathcal{K}\ell_k$ is the extension by zero of $j^*(a + ab)^*\mathcal{K}\ell_k$. So it is sufficient

to check that $j^*a^*\mathcal{K}\ell_k$ is isomorphic to $j^*(a+ab)^*\mathcal{K}\ell_k$ on U, and then applying $j_!$ gives the isomorphism on $\operatorname{Spec} R$.

As $\mathcal{K}\ell_k$ is lisse on $\mathbf{G}_m$, the sheaves $j^*a^*\mathcal{K}\ell_k$ and $j^*(a+ab)^*\mathcal{K}\ell_k$ are both lisse on U. We next check that these two sheaves are isomorphic as lisse sheaves on U, or equivalently that they are isomorphic as representations of $\pi_1(U)$.

First, a and $a+ab$, viewed as maps from $\operatorname{Spec}(R)$ to $\mathbf{A}^1_{\overline{\mathbf{F}}_q}$, both factor through the étale local ring at 0. So on the complement U of the inverse image of zero, both maps factor through the generic point.

By Proposition 4.6(4), the local monodromy representation associated to $\mathcal{K}\ell_k$ at 0 is tame, hence it factors through the tame fundamental group

$$\pi_1^t \simeq \varprojlim_{(n,q)=1} \boldsymbol{\mu}_n(\overline{\mathbf{F}}_q),$$

(see, e.g., [Mil80, Exams. I.5.2(c)]) corresponding to coverings obtained by adjoining n-th roots of the coordinate with $(n,q)=1$. To show that $a^*\mathcal{K}\ell_k$ and $(a+ab)^*\mathcal{K}\ell_k$ are isomorphic on U, it is therefore enough by the Galois correspondence to prove that, for any n with $(n,q)=1$, the pullbacks under a and $a+ab$ of the covers obtained by n-th roots of the coordinate are isomorphic. But $1+b$ is a unit and R is a strict Henselian local ring, so that R contains an n-th root of $1+b$, and the equation

$$(a+ab)^n = a^{1/n}(1+b)^{1/n}$$

gives such an isomorphism.

On the other hand, if $a \in R^\times$, then $a+ab \in R^\times$. Hence both a and $a+ab$, as maps from $\operatorname{Spec}(R)$ to $\mathbf{A}^1_{\overline{\mathbf{F}}_q}$, send the special point to a point $y \in \mathbf{G}_m$. Therefore the pullbacks $a^*\mathcal{K}\ell_k$ and $(a+ab)^*\mathcal{K}\ell_k$ are both locally constant on $\operatorname{Spec}(R)$ and hence correspond to representations of $\pi_1(\operatorname{Spec}(R))$. These are all trivial since $\pi_1(\operatorname{Spec}(R))$ is trivial for R strictly Henselian (see, e.g., [Mil80, Ex. I.5.2(b)]), and since $a^*\mathcal{K}\ell_k$ and $(a+ab)^*\mathcal{K}\ell_k$ have the same rank, they are isomorphic.

(2) Assume now that $a^{-1} \in \mathfrak{m}$. Then

$$u = \frac{a+b}{a} = 1 + \frac{b}{a} \in R^\times,$$

and hence $(a+b)^{-1} = u^{-1}a^{-1} \in \mathfrak{m}$. So both a and $a+b$ (now viewed as maps $\operatorname{Spec}(R) \longrightarrow \mathbf{P}^1_{\overline{\mathbf{F}}_q}$) send the special point of $\operatorname{Spec}(R)$ to $\infty \in \mathbf{P}^1$. Furthermore the inverse image $Z \subset \operatorname{Spec}(R)$ of $\infty \in \mathbf{P}^1_{\overline{\mathbf{F}}_q}$ is the same under both maps, since multiplying by a unit does not change whether a function is infinite at a point. Because the sheaves $a^*\mathcal{K}\ell_k$ and $(a+b)^*\mathcal{K}\ell_k$ are 0 on Z and lisse on the complement $U = \operatorname{Spec}(R) - Z$, they are both the extensions by zero of their restrictions to U, so it is enough to check that they are isomorphic on U as lisse sheaves, or as representations of the fundamental group $\pi_1(U)$.

As representations of the fundamental group, both sheaves are pullbacks of the local monodromy representation of $\mathcal{K}\ell_k$. By Lemma 4.9, the local monodromy of $\mathcal{K}\ell_k$ at ∞ is isomorphic to that of the sheaf

$$[x \mapsto x^k]_*(\mathcal{L}_{\chi_2^{k+1}} \otimes \mathcal{L}_{\tilde{\psi}}),$$

where $\tilde{\psi}(x) = \psi(kx)$. It is therefore enough to show that the pullbacks of this sheaf along a and $a+b$ are isomorphic.

Let $C_a = \mathrm{Spec}(R[a^{-1/k}])$ and $C_{a+b} = \mathrm{Spec}(R[(a+b)^{-1/k}])$, viewed as étale covers of U. Then, because $u = (a+b)/a \in R^\times$ is a unit congruent to 1 modulo $\mathfrak{m}$ (and k is coprime to q), there exists a k-th root (say v) of u in $R^\times$ that is congruent to 1 modulo $\mathfrak{m}$. The two covers are isomorphic via the map

$$C_{a+b} \longrightarrow C_a$$

induced by $y \mapsto vy$. Let $f : C_a \longrightarrow U$ be the covering map. We then have

$$a^*([x \mapsto x^k]_*(\mathcal{L}_{\chi_2^{k+1}} \otimes \mathcal{L}_{\tilde{\psi}})) \simeq f_*\Big((a^{1/k})^*\mathcal{L}_{\chi_2^{k+1}} \otimes (a^{1/k})^*\mathcal{L}_{\tilde{\psi}}\Big),$$

$$(a+b)^*([x \mapsto x^k]_*(\mathcal{L}_{\chi_2^{k+1}} \otimes \mathcal{L}_{\tilde{\psi}})) \simeq f_*\Big((va^{1/k})^*\mathcal{L}_{\chi_2^{k+1}} \otimes (va^{1/k})^*\mathcal{L}_{\tilde{\psi}}\Big).$$

It is therefore sufficient to prove that

$$(a^{1/k})^*\mathcal{L}_{\chi_2^{k+1}} \otimes (a^{1/k})^*\mathcal{L}_{\tilde{\psi}} \simeq (va^{1/k})^*\mathcal{L}_{\chi_2^{k+1}} \otimes (va^{1/k})^*\mathcal{L}_{\tilde{\psi}}.$$

Indeed, since $q \neq 2$ and v is a unit, we first have

$$(a^{1/k})^*\mathcal{L}_{\chi_2^{k+1}} \simeq (va^{1/k})^*\mathcal{L}_{\chi_2^{k+1}}$$

since v is a unit. Furthermore, since $v - 1 = w$ belongs to $\mathfrak{m}$, we get

$$(va^{1/k})^*\mathcal{L}_{\tilde{\psi}} \simeq (a^{1/k})^*\mathcal{L}_{\tilde{\psi}} \otimes (wa^{1/k})^*\mathcal{L}_{\tilde{\psi}}.$$

Now we claim that the second factor is trivial on $R[a^{-1/k}]$, which concludes the proof. Indeed, w is in the ideal generated by a^{-1} (by the power series $v = 1 + bk^{-1}a^{-1} + \cdots$), so $wa^{1/k}$ is in the ideal generated by $a^{-(k-1)/k}$ and thus in the maximal ideal of $R[a^{-1/k}]$. Hence it sends $\mathrm{Spec}\, R[a^{-1/k}]$ to a neighborhood of 0 in $\mathbf{A}^1_{\overline{\mathbf{F}}_q}$, where $\mathcal{L}_{\tilde{\psi}}$ is lisse and hence locally trivial, so the pullback $(wa^{1/k})^*\mathcal{L}_{\tilde{\psi}}$ is trivial. □

We will need some simple facts about hypergeometric sheaves in the sense of Katz [Kat90] — more precisely, about a particular hypergeometric sheaf.

Definition 4.18. For $k \geqslant 2$ an integer, we denote by $\mathcal{H}_{k-1}$ the middle extension to $\mathbf{A}^1$ with coordinate ξ of the ℓ-adic sheaf on $\mathbf{G}_m$ given by

$$\mathcal{H}_{k-1} = [\xi \mapsto \xi^{-1}]^* j^* \,\mathrm{FT}_\psi(\mathcal{L}_{\tilde{\psi}}(x^{1/k})),$$

where $j : \mathbf{G}_m \longrightarrow \mathbf{A}^1$ is the open immersion, and we recall that $\tilde{\psi}(x) = \psi(kx)$.

It is important for later purpose to note the following lemma:

LEMMA 4.19. *The sheaf $\mathcal{H}_{k-1}$ is a multiplicative translate of a hypergeometric sheaf of type $(k-1,0)$ in the sense of Katz. More precisely, it is geometrically isomorphic to*

$$\mathrm{Hyp}_{(-1)^k}(!,\bar{\psi};\{\chi|\chi^k=1,\ \chi\neq 1\};\emptyset),$$

with the notation of [Kat90, 8.2.2, 8.2.13]. *The inertia representation of $\mathcal{H}_{k-1}$ at infinity is absolutely irreducible.*

We thank the referee for giving a proof that is simpler than our original.

Proof. Since both $\mathcal{H}_{k-1}$ and hypergeometric sheaves are middle-extension sheaves (recall that $k\geqslant 2$), it is enough to prove the isomorphism after restriction to $\mathbf{G}_m$. We compute

$$\begin{aligned}
j^*\mathcal{H}_{k-1} &= [\xi\mapsto\xi^{-1}]^*j^*\mathrm{FT}_\psi(\mathcal{L}_{\tilde{\psi}}(x^{1/k}))\\
&\simeq [\xi\mapsto\xi^{-1}]^*j^*\mathrm{FT}_\psi(j_*\mathrm{Hyp}_1(!,\psi;\{\chi|\chi^k=1\};\emptyset)) \qquad \text{[Kat88, 5.6.2]}\\
&\simeq [\xi\mapsto\xi^{-1}]^*\mathrm{Hyp}_{(-1)^k}(!,\psi;\emptyset;\{\chi\neq 1,\ \chi^k=1\})) \qquad \text{[Kat88, 5.6.2]}\\
&\simeq \mathrm{Hyp}_{(-1)^k}(!,\overline{\psi};\{\chi\neq 1,\ \chi^k=1\};\emptyset),
\end{aligned}$$

where $\simeq$ always denotes geometric isomorphisms.

The last assertion now follows from [Kat90, Th. 8.4.2(6)], which shows that the inertia representation at ∞ is of dimension $k-1$ with unique break $1/(k-1)$ and [Kat88, Prop. 1.14], which shows that such a representation of the inertia group at ∞ is absolutely irreducible. □

We will need some properties of the local monodromy at ∞ of $\mathcal{H}_{k-1}$. To state them, we need the following definition:

Definition 4.20. Let K be a local field, and let σ be an automorphism of K. Let $n\geqslant 1$ be an integer, and let π be a uniformizer of K. We say that σ is *a reparametrization of order n* if $\sigma(\pi)$ is a uniformizer of K such that

$$\sigma(\pi)\equiv\pi\,(\mathrm{mod}\,\pi^n).$$

Note that since an order n reparametrization acts on K, it also defines an outer automorphism of the Galois group of K: each extension $\bar{\sigma}$ of σ to a separable closure $\bar{K}$ of K gives an automorphism of $\mathrm{Gal}(\bar{K}/K)$, and the ambiguity in the possible choices of this extension amounts to conjugating $\bar{\sigma}$ with an element of $\mathrm{Gal}(\bar{K}/K)$, so that the corresponding outer automorphism of the Galois group is well defined. This outer automorphism defines an action of σ on the set of isomorphism classes of representations of the Galois group. More abstractly, σ defines an automorphism of the category of finite étale covers of $\mathrm{Spec}(K)$ by pullback, and hence acts on the category of étale sheaves on $\mathrm{Spec}(K)$, which is equivalent to the category of Galois representations.

LEMMA 4.21. *Assume that $q > k \geqslant 2$.*

(1) *The local monodromy representation at infinity of $\mathcal{H}_{k-1}$ is invariant under reparametrizations of order 2.*
(2) *The local monodromy representation at infinity of $\mathcal{H}_{k-1}$, restricted to the wild inertia group, is a direct sum of pairwise nonisomorphic characters with multiplicity 1. The tame inertia group acts transitively on these characters.*
(3) *Let α_1, α_2 be elements of an algebraically closed extension of $\mathbf{F}_q$ such that the wild local monodromy representation at infinity of $[\times\alpha_1]^*\mathcal{H}_{k-1}$ and $[\times\alpha_2]^*\mathcal{H}_{k-1}$ have a common irreducible component. Then $\alpha_1 = \alpha_2$.*

Proof. The integer q is coprime with $2(k-1)$ since $q > k \geqslant 2$. By [Fu10, Th. 0.1(iii)] (which is more precise), we derive isomorphisms of $I(\infty)$-representations

(4.5)
$$\begin{aligned}\mathcal{H}_{k-1|I(\infty)} &\simeq \mathrm{FT}_\psi(\mathcal{L}_{\tilde{\psi}}(x^{1/k}))_{|I(0)}\\ &\simeq \mathrm{FT}_\psi \operatorname{loc}(\infty,0)([t\mapsto t^k]_*\mathcal{L}_{\tilde{\psi}}) \simeq [t\mapsto -t^{k-1}]_*(\mathcal{L}_{\psi((k-1)t)}\otimes\mathcal{L}_{\chi_2}),\end{aligned}$$

where $\mathrm{FT}_\psi \operatorname{loc}(\cdot,\cdot)$ denotes Laumon's local Fourier transform functors (see, e.g, [Kat90, 7.4]).

To prove (1), it is therefore enough to prove that for any additive character η and any multiplicative character χ, the local monodromy representation at ∞ of $[t\mapsto t^{k-1}]_*(\mathcal{L}_\eta\otimes\mathcal{L}_\chi)$ is invariant under reparametrizations of order 2.

Let V denote this representation. Let R be the strict henselization at ∞, let K be its field of fractions, and let π be a uniformizer of R. Let $g : \mathrm{Spec}(K) \longrightarrow \mathrm{Spec}(K)$ be the map corresponding to $t\mapsto t^{k-1}$. A representation obtained from V by applying a reparametrization of order 2 is of the form $\sigma^*V = (\sigma^{-1})_*V$, where σ is an automorphism K such that $\sigma(\pi)\equiv\pi \pmod{\pi^2}$. We view σ and σ^{-1} as automorphisms $\mathrm{Spec}(K)\longrightarrow\mathrm{Spec}(K)$.

Let $W = \mathcal{L}_\eta\otimes\mathcal{L}_\chi$; we have $V\simeq g_*W$ and hence $(\sigma^{-1})_*V = \tau_*W$, where $\tau = \sigma^{-1}\circ g$. There exists an automorphism σ_1 such that $\tau = g\circ\sigma_1$, and σ_1 is a reparametrization of order k. We can see this in coordinates by solving the equation

$$\sigma_1(t)^{k-1} = \sigma^{-1}(t^{k-1}) = t^{k-1} + a_1t^{2(k-1)} + \cdots$$

with

$$\sigma_1(t) = t + \frac{a_1}{k-1}t^k + \cdots.$$

Thus $\sigma^*V\simeq g_*(\sigma_{1,*}W)$ and, in particular, we obtain $\sigma^*V\simeq V$, provided W is invariant under reparametrizations of order k. In fact, we will show that both $\mathcal{L}_\eta$ and $\mathcal{L}_\chi$ are invariant under any reparametrization σ_1 of order $k\geqslant 2$, which will be enough.

For multiplicative characters, this amounts to saying that for d coprime to q, the covering $\mathrm{Spec}(K(\pi^{-1/d})) \longrightarrow \mathrm{Spec}(K)$ is invariant under σ_1, which is clear because if we write $\sigma_1(\pi) = \pi + b\pi^2$ for some $b \in R$, we get

$$\sigma_1(\pi)^{-1/d} = \pi^{-1/d}(1 + b\pi)^{-1/d}$$

and $(1 + b\pi)^{-1/d} \in K$. For additive characters, this amounts to proving that the Artin–Schreier covering with equation $y^q - y = \pi^{-1}$ is invariant, and this follows because the equation

$$z^q - z = \frac{1}{\sigma_1(\pi)} - \frac{1}{\pi} = -\frac{b}{1 + b\pi}$$

is solvable in K.

(2) By (4.5), the local wild monodromy representation of $\mathcal{H}_{k-1}$ at ∞ is isomorphic to

$$[t \mapsto -t^{k-1}]_*(\mathcal{L}_{\psi((k-1)t)}).$$

It is equivalent to study this after pulling back by any tame cover. In particular, after pulling back along the map $t \mapsto t^{k-1}$, we have to deal with

$$\bigoplus_{\xi^{k-1}=1} \mathcal{L}_{\psi(\xi(k-1)t)}, \tag{4.6}$$

which is indeed a sum of one-dimensional characters. They are pairwise non-isomorphic. (If we have, say, an isomorphism $\mathcal{L}_{\psi(\xi_1(k-1)t)} \simeq \mathcal{L}_{\psi(\xi_2(k-1)t)}$ as representations of the wild inertia group, then $\mathcal{L}_{\psi((\xi_1-\xi_2)(k-1)t)}$ is tamely ramified, which means that $\xi_1 = \xi_2$ since otherwise $\mathcal{L}_{\psi((\xi_1-\xi_2)(k-1)t)}$ is a nontrivial additive character sheaf with Swan conductor 1.)

Since $\mathcal{H}_{k-1}$ is an irreducible representation of the full inertia group at infinity (Lemma 4.19), the tame inertia group acts transitively by conjugation on the set of characters in (4.6). (The direct sum of any subset of the characters that is stable under the tame inertia group would define an inertia-invariant subspace.)

(3) Let $L/\mathbf{F}_q$ be an algebraically closed extension. We use the same notation $\mathcal{H}_{k-1}$ and $\mathcal{L}_\psi$ for the sheaves base-changed to L so that, for instance, $[\times\alpha]^*\mathcal{H}_{k-1}$ and $\mathcal{L}_{\psi(\beta t)}$ are well defined for α and $\beta \in L^\times$.

Adding a multiplicative shift to the computation of (2), the pullback along $[t \mapsto -t^{k-1}]$ of the local wild monodromy representation of $[\times\alpha]^*\mathcal{H}_{k-1}$ at ∞ is isomorphic to

$$\bigoplus_{\beta^{k-1}=\alpha} \mathcal{L}_{\psi((k-1)\beta t)}.$$

If the local wild monodromy representations of $[\times\alpha_1]^*\mathcal{H}_{k-1}$ and $[\times\alpha_2]^*\mathcal{H}_{k-1}$ at ∞ have a common irreducible component, then one of the additive characters appearing in one of the two sums must also appear in the other, so there exists β such that $\alpha_1 = \beta^{k-1} = \alpha_2$. □

The following lemma is quite standard but we include a proof for lack of a suitable reference:

LEMMA 4.22. (1) *Let $U_{\mathbf{F}_q}$ be a dense open subset of a smooth projective curve $C_{\mathbf{F}_q}$, and let $\mathcal{F}$ be an ℓ-adic sheaf on C. Assume that $\mathcal{F}$ is lisse and pure of weight w on U, that it has no punctual sections, and that, viewed as a representation of the geometric fundamental group of U, it has no trivial subrepresentation.*

Then the subspace of weight $< w+1$ of $H^1(C_{\overline{\mathbf{F}}_q}, \mathcal{F})$ is equal to

$$\bigoplus_{x \in C-U} (\mathcal{F}_{\overline{\eta}}^{I_x}/\mathcal{F}_{\overline{x}}),$$

where I_x is the inertia group at x and $\mathcal{F}_{\overline{\eta}}$ is the stalk at the geometric generic point of $\mathcal{F}$.

(2) *Let $\pi : C \to X$ be a smooth projective morphism of relative dimension 1 over $\mathbf{F}_q$, and let $\mathcal{F}$ be an ℓ-adic sheaf on C. Assume that $\mathcal{F}$ is lisse and pure of weight w on e open subset $U \subset C$. Assume that for all $x \in X$ in some dense open subset, $\mathcal{F}|C_x$ has no punctual sections and that, when $\mathcal{F}|C_x$ is viewed as a representation of the geometric fundamental group of U_x, it has no trivial subrepresentation.*

On the dense open set where $R^1\pi_*\mathcal{F}$ is lisse, let $(R^1\pi_*\mathcal{F})^{<w+1}$ be the maximal lisse subsheaf of $R^1\pi_*\mathcal{F}$ of weight $< w+1$. Then for any point x in the dense open subset where $R^1\pi_*\mathcal{F}$ is lisse, we have an isomorphism

$$(R^1\pi_*\mathcal{F})_x^{<w+1} = \bigoplus_{y \in C_x - U_x} ((\mathcal{F}|C_x)_{\overline{\eta}}^{I_y}/(\mathcal{F}|C_x)_{\overline{y}}),$$

where $(\mathcal{F}|C_x)_{\overline{\eta}}$ is the stalk at the geometric generic point of the restriction of $\mathcal{F}$ to C_x.

Proof. (1) Let $j : U \longrightarrow C$ denote the open immersion. Because $\mathcal{F}$ has no punctual sections, the natural adjunction map $\mathcal{F} \to j_*j^*\mathcal{F}$ is injective. Let $\mathcal{G}$ be its cokernel. Then we have a long exact sequence

$$\cdots \longrightarrow H^i(C_{\overline{\mathbf{F}}_q}, \mathcal{F}) \to H^i(C_{\overline{\mathbf{F}}_q}, j_*j^*\mathcal{F}) \to H^i(C_{\overline{\mathbf{F}}_q}, \mathcal{G}) \longrightarrow \cdots . \tag{4.7}$$

By assumption on $\mathcal{F}$, we have

$$H^0(C_{\overline{\mathbf{F}}_q}, j_*j^*\mathcal{F}) = H^0(U_{\overline{\mathbf{F}}_q}, j^*\mathcal{F}) = 0.$$

Since $\mathcal{G}$ is supported on $C-U$, its cohomology vanishes in degree above 1, and hence we deduce a short exact sequence

$$0 \to H^0(C_{\overline{\mathbf{F}}_q}, \mathcal{G}) \to H^1(C_{\overline{\mathbf{F}}_q}, \mathcal{F}) \to H^1(C, j_*j^*\mathcal{F}) \to 0.$$

Because $j_*j^*\mathcal{F}$ is the middle extension of a lisse sheaf pure of weight w, a result of Deligne implies that its cohomology group $H^1(C_{\overline{\mathbf{F}}_q}, j_*j^*\mathcal{F})$ is pure of weight $w+1$ (see [Del80, Ex. 6.2.5(c) and Prop. 6.2.6]). Therefore the

weight $< w+1$ part of $H^1(C_{\overline{\mathbf{F}}_q}, \mathcal{F})$ is the same as the weight $< w+1$ part of $H^0(C_{\overline{\mathbf{F}}_q}, \mathcal{G})$. Since the sheaf $\mathcal{G}$ is punctual, we have

$$H^0(C_{\overline{\mathbf{F}}_q}, \mathcal{G}) = \bigoplus_{x \in C-U} \mathcal{G}_{\overline{x}} = \bigoplus_{x \in C-U} (j_* j^* \mathcal{F})_{\overline{x}} / \mathcal{F}_{\overline{x}}$$

(by definition of $\mathcal{G}$). We also have

$$(j_* j^* \mathcal{F})_{\overline{x}} = \mathcal{F}_{\overline{\eta}}^{I_x},$$

and [Del80, Lemma 1.8.1] shows that this space is of weight $\leqslant w$, so that all of $H^0(C_{\overline{\mathbf{F}}_q}, \mathcal{G})$ is the weight $< w+1$ part of $H^1(C_{\overline{\mathbf{F}}_q}, \mathcal{F})$, as claimed.

(2) Denote again by $j : U \longrightarrow C$ the open embedding. We want to apply (1) fiber by fiber. First (since pushforward does not commute with arbitrary base change), we let U_1 denote a dense open subset of X such that the adjunction map

$$\mathcal{F}|\pi^{-1}(U_1) \to j_* j^* \mathcal{F}|\pi^{-1}(U_1)$$

is injective. (The existence of such a dense open set follows from [Del77, Th. Finitude, Th. 1.9(2)], applied to the morphism j over the base X.) Let $\mathcal{G}$ be the quotient sheaf. Then we again take the long exact sequence

$$\cdots \longrightarrow R^i\pi_* \mathcal{F} \to R^i\pi_* j_* j^* \mathcal{F} \to R^i\pi_* \mathcal{G} \longrightarrow \cdots .$$

The fiber over any $x \in U_1$ of this exact sequence is the same as the exact sequence (4.7) for the fiber curve C_x, again using [Del77, Th. Finitude, Th. 1.9(2)]. In particular, for any point $x' \in U_1$ (closed or not), we have

$$\bigoplus_{y \in C_{x'} - U_{x'}} ((\mathcal{F}|C_{x'})_{\overline{\eta}}^{I_y} / (\mathcal{F}|C_{x'})_{\overline{y}}) = (R^0\pi_* \mathcal{G})_{x'}.$$

Thus (1) shows over any closed point $x' \in U_1$ that the weight $< w+1$ part of $(R^1\pi_* \mathcal{F})_{x'}$ is the image of $(R^0\pi_* \mathcal{G})_{x'}$ in $(R^1\pi_* \mathcal{F})_{x'}$. Over a possibly smaller dense open set $U_2 \subset U_1$ where $R^0\pi_* \mathcal{G}$ and $R^1\pi_* \mathcal{F}$ are both lisse, this implies that the maximal weight $< w+1$ lisse subsheaf of $R^1\pi_* \mathcal{F}$ is $R^0\pi_* \mathcal{G}$. Then for an arbitrary $x \in U_2$, we have

$$(R^1\pi_* \mathcal{F})_x^{<w+1} = (R^0\pi_* \mathcal{G})_x = \bigoplus_{y \in C_x - U_x} ((\mathcal{F}|C_x)_{\overline{\eta}}^{I_y} / (\mathcal{F}|C_x)_{\overline{y}}).$$

If x is the generic point, it is necessarily contained in the dense open subset U_2. If not, we can replace X by the closure of x in X and apply the same argument, obtaining the same identity (because the direct sum

$$\bigoplus_{y \in C_x - U_x} ((\mathcal{F}|C_x)_{\overline{\eta}}^{I_y} / (\mathcal{F}|C_x)_{\overline{y}})$$

depends only on the fiber over x, and the same for $(R^1\pi_* \mathcal{F})_x^{\leq w+1}$, since taking the weight $< w+1$ part commutes with restriction to a closed subscheme.) $\square$

The next lemma will be useful to bound in terms of q the degree of the subvariety $\mathcal{V}^{\mathrm{bad}}$ for $\lambda = 0$, by showing that this variety is defined over $\mathbf{Z}[1/\ell]$.

LEMMA 4.23. *Let X and Y be separated varieties of finite type over $\mathbf{Z}[1/\ell]$. Let $f : X \to Y$ and $g : X \to \mathbf{A}^1$ be morphisms. Let $p_2 : \mathbf{G}_m \times X \to X$ be the second projection.*

There exists an ℓ-adic complex K on Y such that, for any prime $q \neq \ell$ and any additive character ψ of $\mathbf{F}_q$, we have

$$R(f \circ p_2)_! \mathcal{L}_\psi(tg) = K|Y_{\mathbf{F}_q}.$$

Proof. We denote by t a coordinate on $\mathbf{G}_m$. As $R(f \circ p_2)_! = Rf_! Rp_{2,!}$, it is sufficient to prove that there exists a complex K' on X with

$$Rp_{2,!}\mathcal{L}_{\psi(tg)} = K'|X_{\mathbf{F}_q}$$

for all $q \neq \ell$ and all ψ, as we can then take $K = Rf_! K'$.

Let p_2' denote the second projection $\mathbf{G}_m \times \mathbf{A}^1 \to \mathbf{A}^1$. By the proper base change theorem, we have

$$Rp_{2,!}\mathcal{L}_{\psi(tg)} = g^* Rp'_{2,!}\mathcal{L}_{\psi(tx)}$$

for any $q \neq \ell$ and ψ, so it is sufficient to find a complex K^* on $\mathbf{A}^1_{\mathbf{Z}[1/\ell]}$ with

$$Rp'_{2!}\mathcal{L}_{\psi(tx)} = K^*|\mathbf{A}^1_{\mathbf{F}_q}$$

for all $q \neq \ell$ and all ψ, and to define $K' = g^* K^*$.

By the above reduction we may assume that $X = \mathbf{A}^1_{\mathbf{Z}[1/\ell]}$ and write p_2 for p_2'. Let $j : \mathbf{G}_m \to \mathbf{A}^1$ be the open immersion and $i : \{0\} \to \mathbf{A}^1$ the complementary closed immersion. Then $Rp_{2,!}\mathcal{L}_{\psi(tx)}$ is the Fourier transform of $j_!\overline{\mathbf{Q}}_\ell$ (as extension by zero commutes with pullback and tensor product). The existence of an ℓ-adic complex on $\mathbf{A}^1_{\mathbf{Z}[1/\ell]}$ that specializes to $\mathrm{FT}_\psi j_!\overline{\mathbf{Q}}_\ell = Rp_{2,!}\mathcal{L}_{\psi(tx)}$ in each positive characteristic $q \neq \ell$ is a special case of Laumon's homogeneous Fourier transform (see [Lau03, Th. 2.2]). In this special case, L. Fu (see [Fu16, Lemma 3.2]) showed that we can take the complex to be $j_*\overline{\mathbf{Q}}_\ell$. □

Finally, we can prove the last part of Theorem 4.10.

PROPOSITION 4.24. *For all $\lambda_1, \lambda_2 \in \mathbf{F}_q, \boldsymbol{b}_1, \boldsymbol{b}_2 \notin \mathcal{V}^\Delta(\mathbf{F}_q)$, the dimensions of the stalks of the sheaf $\mathcal{R}$ and the dimensions of the cohomology groups*

$$H^i_c(\mathbf{A}_{\overline{\mathbf{F}}_q}, \mathcal{R}_{\lambda_1,\boldsymbol{b}_1}), \qquad H^i_c(\mathbf{A}_{\overline{\mathbf{F}}_q}, \mathcal{R}_{\lambda_1,\boldsymbol{b}_1} \otimes \mathcal{R}_{\lambda_2,\boldsymbol{b}_2})$$

are all bounded in terms of k only.

Proof. We deal with the second of these cohomology groups. Fix λ_1, λ_2 in $\mathbf{F}_q$, $\boldsymbol{b}_1, \boldsymbol{b}_2 \notin \mathcal{V}^\Delta(\mathbf{F}_q)$. By construction of $\mathcal{R}$ and by interpreting sheaf-theoretically the definition of the hyper-Kloosterman sums, there exist an affine

variety $V_{\mathbf{Z}}$ and maps $f : V \to \mathbf{A}^1_{\mathbf{Z}}$ and $g : V \to \mathbf{A}^1_{\mathbf{Z}}$ such that, for any prime q, we have

$$(\mathcal{R}_{\lambda_1,\boldsymbol{b}_1} \otimes \mathcal{R}_{\lambda_2,\boldsymbol{b}_2})|\mathbf{A}^1_{\overline{\mathbf{F}}_q} = Rf_!g^*\mathcal{L}_\psi[2].$$

(See also Lemma 4.27 below for this construction.)

By the Leray spectral sequence, it is enough to bound the sum of Betti numbers

$$\sum_i \dim H^i_c(\tilde{V}_{\overline{\mathbf{F}}_q}, \mathcal{L}_{\psi(g)}),$$

where $\tilde{V}$ is the inverse image in V of either a line or a plane. Since (V, f, g) are defined over $\mathbf{Z}$, a suitable bound is given by the estimates of Bombieri and Katz for sums of Betti numbers. (See the version of Katz in [Kat01, Th. 12].)

A similar argument applies to $H^i_c(\mathbf{A}^1_{\overline{\mathbf{F}}_q}, \mathcal{R}_{\lambda,\boldsymbol{b}})$. □

Finally, we need a lemma on inertia groups that is probably well known but for which we do not know a convenient reference.

LEMMA 4.25. *Let $\pi\colon \mathbf{A}^5_{\overline{\mathbf{F}}_q} \to \mathbf{A}^4_{\overline{\mathbf{F}}_q}$ be the projection on the first four coordinates. Let $\bar{\pi}\colon \mathbf{P}^4 \times \mathbf{P}^1 \to \mathbf{P}^4$ be the analogue projection. For any divisor D in $\mathbf{P}^4$, the induced homomorphism from the inertia group at the generic point of $\bar{\pi}^{-1}(D)$ to the inertia group at the generic point of D is surjective.*

Proof. Let R (resp. R') be the étale local ring of D (resp. of $\bar{\pi}^{-1}(D)$) at its generic point, and let K (resp. K') be its field of fractions. Then the inertia group I_D of D is the Galois group of K and the inertia group $I_{\bar{\pi}^{-1}(D)}$ of $\bar{\pi}^{-1}(D)$ is the Galois group of K'. If the homomorphism $I_{\bar{\pi}^{-1}(D)} \to I_D$ of profinite groups is not surjective, then its image is contained in some proper open subgroup of I_D. By the Galois correspondence, this means that there exists some finite étale covering $E \to K$ without a section whose pullback to K' admits a section.

We will show that every finite étale covering $E \to \mathrm{Spec}(K)$ whose pullback to K' admits a section already has a section over K, implying by contradiction that the homomorphism is surjective, as claimed.

Let $E \to \mathrm{Spec}(K)$ be such a covering, and let s' be a section of the pullback to K'. The section s' is defined over $K' = R'[t^{-1}]$, where t is a uniformizer of R (and hence also a uniformizer of R'). Because R' is the étale local ring of the generic point of $\bar{\pi}^{-1}(D)$, it is the étale local ring of the generic point η of the special fiber $\mathbf{A}^1_R$. Because the section s' is necessarily defined over some finitely generated subring of $R'[t^{-1}]$, and R' is the limit of the rings of functions on all étale neighborhoods of η, the section s' is defined over the ring of functions on some étale neighborhood $X \to \mathbf{A}^1_R$ of η, after inverting t. The equations for s' over this ring describe a map $s\colon X - \{t = 0\} \to E$ over $\mathrm{Spec}(R)$.

The image of X in $\mathbf{A}^1_R$ contains a Zariski neighborhood of η, which contains all but finitely many closed points of the special fiber. Hence it contains the image of some section of the projection $\pi : \mathbf{A}^1_R \to \mathrm{Spec}(R)$. Let Y be the pullback of X along that section. Then there is a morphism $Y - \{t = 0\} \to E$, and Y is an étale cover of $\mathrm{Spec}(R)$, so it contains a copy of $\mathrm{Spec}(R)$, hence there is a map $\mathrm{Spec}(R) - \{t = 0\} \to E$, which means that E admits a section over $\mathrm{Spec}(K) = \mathrm{Spec}(R) - \{t = 0\}$. □

4.4. *Irreducibility of sum-product sheaves for $\lambda = 0$.* We now begin the study of sum-product sheaves in the case $\lambda = 0$. *We always assume that $q > k$.*

We denote by $\mathcal{R}_{\lambda=0}$ (resp. $\mathcal{R}^*_{\lambda=0}$) the pullback $i^*\mathcal{R}$ (resp. $i^*\mathcal{R}^*$) for the inclusion i of $\mathbf{A}^5$ in $\mathbf{A}^6$ such that $i(r, \boldsymbol{b}) = (r, 0, \boldsymbol{b})$, and similarly we define $\mathcal{K}_{\lambda=0}$.

The main result of this section establishes that for q large enough, and for generic values of $\boldsymbol{b}$ (i.e., outside some proper subvariety $\mathcal{V}^{\mathrm{bad}} \subset \mathbf{A}^4_{\mathbf{F}_q}$), the sheaf $\mathcal{R}^*_{\lambda=0,\boldsymbol{b}}$ is geometrically irreducible.

The strategy is as follows:

(1) A key observation (Lemma 4.27) is that, by homogeneity, $\mathcal{R}^*_{\lambda=0}$ is defined over $\mathbf{Z}[1/\ell]$. In particular, this implies that for q large enough, the sheaf $\mathcal{R}^*_{\lambda=0}$ is not wildly ramified.
(2) In Proposition 4.28, we use this fact together with the diophantine criterion of irreducibility (Lemma 4.14) and the the mean square average asymptotic formula of Proposition 4.3 to prove that $\mathcal{R}^*_{\lambda=0,\boldsymbol{b}}$ is generically geometrically irreducible.
(3) In Proposition 4.29, we conclude and show, using (1), that $\mathcal{V}^{\mathrm{bad}}$ is in fact defined over $\mathbf{Z}[1/\ell]$.

LEMMA 4.26. *Let Z be the subvariety of $\mathbf{A}^5 \times \mathbf{A}^4$ defined by*

$$\{(r, \boldsymbol{b}, x_1, x_2, x_3, x_4) \in \mathbf{A}^5 \times \mathbf{A}^4 \mid x_i^k = (r + b_i),\ i = 1, \dots, 4\}, \tag{4.8}$$

and let $\tilde{Z} \subset \mathbf{A}^5$ be the image of the subvariety of Z defined by the equation $x_1 + x_2 - x_3 - x_4 = 0$ under the projection

$$(r, \boldsymbol{b}, x_1, x_2, x_3, x_4) \in Z \mapsto (r, \boldsymbol{b}) \in \mathbf{A}^5.$$

(1) *The image $\tilde{Z}$ is closed and irreducible.*
(2) *Let U be the dense open complement of the union of $\tilde{Z}$ and of the divisors given by the equations $r = -b_i$ in $\mathbf{A}^5$. The sheaf $\mathcal{R}_{\lambda=0}$ is lisse on U.*
(3) *On any dense open subset $V \subset U$ where $\mathcal{R}^*_{\lambda=0}$ is lisse, the monodromy representation of $\mathcal{R}^*_{\lambda=0}$ factors through $\pi_1(U)$.*

Proof. (1) The projection $Z \to \mathbf{A}^5$ is finite because Z is defined by adjoining the coordinates x_1, x_2, x_3, x_4 to $\mathbf{A}^5$, and each satisfies a monic polynomial equation. Thus the closed subvariety of Z defined by the equation

$x_1 + x_2 - x_3 - x_4 = 0$ is also finite over $\mathbf{A}^5$, and hence its image $\tilde{Z}$ is closed. Moreover, $\tilde{Z}$ is the projection of the subvariety of $\mathbf{A}^9$ with equations

$$\begin{cases} x_i^k = r + b_i, \quad 1 \leqslant i \leqslant 4, \\ x_1 + x_2 - x_3 - x_4 = 0, \end{cases}$$

and hence this subvariety is isomorphic to the divisor in $\mathbf{A}^5$ with coordinates (x_1, x_2, x_3, x_4, r) given by the equation $x_1 + x_2 - x_3 - x_4 = 0$. In particular, it is irreducible, and therefore its image $\tilde{Z}$ is also irreducible.

(2) This will use Deligne's semicontinuity theorem [Lau81]. Precisely, as we already observed, the sheaf $\mathcal{K}_{\lambda=0}$ is lisse on the complement of the divisors given by the equations $r = -b_i$ and $s = 0$ in $\mathbf{A}^6$. We compactify the s-coordinate by $\mathbf{P}^1$ and work on

$$X = (\mathbf{A}^1 \times \mathbf{P}^1 \times \mathbf{A}^4) \cap \{(r, s, \boldsymbol{b}) \mid (r, \boldsymbol{b}) \in U\}.$$

By extending by 0, we view $\mathcal{K}_{\lambda=0}$ as a sheaf on X that is lisse on the complement in X of the divisors $s = 0$ and $s = \infty$ (because U is contained in the complement of the divisors $r = -b_i$ and thus X is as well). Let

$$\pi^{(2)} \, : \, X \longrightarrow U$$

denote the projection $(r, s, \boldsymbol{b}) \mapsto (r, \boldsymbol{b})$. Then $\pi^{(2)}$ is proper and smooth of relative dimension 1 and $\mathcal{R}_{\lambda=0}|U = R^1\pi^{(2)}_*\mathcal{K}$.

Since the restrictions of $\mathcal{K}_{\lambda=0}$ to the divisors $s = \infty$ and $s = 0$ are zero, this sheaf is the extension by zero from the complement of those divisors to the whole space of a lisse sheaf. Deligne's semicontinuity theorem [Lau81, Cor. 2.1.2] implies that the sheaf $\mathcal{R}_{\lambda=0}$ is lisse on U if the Swan conductor is constant on each of these two divisors.

On $s = 0$, the Kloosterman sheaf has tame ramification, hence any tensor product of Kloosterman sheaves has tame ramification. Thus $\mathcal{K}_{\lambda=0}$ has tame ramification at $s = 0$ and, in particular, the Swan conductor is zero, which is constant.

On the other hand, Lemma 4.16 gives a formula for the local monodromy representation at $s = \infty$ as a sum of pushforward representations from the tame covering $x \mapsto x^k$. Since the Swan conductor is additive and since the Swan conductor is invariant under pushforward by a tame covering (see, e.g., [Kat88, 1.13.2]), it follows that

$$\begin{aligned} &\mathrm{Swan}_\infty(\mathcal{K}_{r,\lambda=0,\boldsymbol{b}}) \\ &= \sum_{\zeta_2,\zeta_3,\zeta_4 \in \boldsymbol{\mu}_k} \mathrm{Swan}_\infty\Big(\mathcal{L}_\psi\Big(\big((r+b_1)^{1/k} + \zeta_2(r+b_2)^{1/k} \\ &\qquad\qquad - \zeta_3(r+b_3)^{1/k} - \zeta_4(r+b_4)^{1/k}\big)\, t\Big)\Big) = k^3, \end{aligned}$$

by definition of U, since the Swan conductor of $\mathcal{L}_{\psi(at)}$ is 1 for $a \neq 0$. This is constant, and therefore we deduce that $\mathcal{R}_{\lambda=0}$ is lisse on U.

(3) The restriction $\mathcal{R}^*_{\lambda=0}|V$ is a quotient of $\mathcal{R}|V$, and both sheaves are lisse on V; since the monodromy representation of $\mathcal{R}|V$ factors through $\pi_1(U)$, the same holds for $\mathcal{R}^*|V$. □

We can now deduce the main result of this section. We first show that $\mathcal{R}_{\lambda=0}$ is defined over $\mathbf{Z}$. Intuitively, this is because its trace function is independent of the choice of additive character ψ used in the definition of the Kloosterman sheaf. Indeed, let $\psi'(x) = \psi(\lambda x)$, for some $\lambda \in \mathbf{F}_q^\times$, be any nontrivial additive character of $\mathbf{F}_q$ and let

$$\mathrm{Kl}_{k,\psi'}(x) = q^{-\frac{k-1}{2}}(\psi' \star \cdots \star \psi')(x) = \mathrm{Kl}_k(\lambda^k x)$$

be the Kloosterman sums defined using ψ' instead of ψ. We then have, with obvious notation, the equality

$$\begin{aligned}\mathbf{R}_{\psi'}(r,0,\boldsymbol{b}) &= \sum_s \prod_{i=1}^{2} \mathrm{Kl}_{k,\psi'}(s(r+b_i))\overline{\mathrm{Kl}_{k,\psi'}(s(r+b_{i+2}))} \\ &= \sum_s \prod_{i=1}^{2} \mathrm{Kl}_k(\lambda^k s(r+b_i))\overline{\mathrm{Kl}_k(\lambda^k s(r+b_{i+2}))} = \mathbf{R}(r,0,\boldsymbol{b})\end{aligned}$$

by making the change of variable $s \mapsto \lambda^k s$, so that $(r,\boldsymbol{b}) \mapsto \mathbf{R}(r,0,\boldsymbol{b})$ does not depend on the choice of ψ.

The following lemma is a geometric incarnation of this simple computation:

LEMMA 4.27. *For any prime ℓ, there exists an ℓ-adic sheaf $\mathcal{R}^{\mathrm{univ}}$ on $\mathbf{A}^5_{\mathbf{Z}[1/\ell]}$ such that, for any prime $q \neq \ell$, we have*

$$\mathcal{R}^{\mathrm{univ}}|\mathbf{A}^5_{\mathbf{F}_q} = \mathcal{R}_{\lambda=0},$$

where $\mathcal{R}_{\lambda=0}$ is defined using the Kloosterman sheaf $\mathcal{K}\ell_{\psi,k}$ for any *nontrivial additive character ψ of $\mathbf{F}_q$.*

Proof. Let $X_1 \subset \mathbf{G}_m^{k+1}$ be the subvariety over $\mathbf{Z}$ with equation

$$x_1 \cdots x_k = t,$$

and let

$$f_1 : X_1 \longrightarrow \mathbf{A}^1$$

be the projection $(x_1,\dots,x_k,t) \mapsto t$. For any prime $q \neq \ell$ and any ψ, we then have an isomorphism

$$\mathcal{K}\ell_{k,\psi} \simeq Rf_{1,!}^{k-1}\mathcal{L}_\psi(x_1 + \cdots + x_k)$$

of sheaves on $\mathbf{A}^1_{\mathbf{F}_q}$ (up to a Tate twist). Let X_2 be the variety in $\mathbf{G}_m^{4k} \times \mathbf{A}^6$ (over $\mathbf{Z}[1/\ell]$) defined by the equations

$$\prod_{j=1}^{k} x_{i,j} = s(r + b_i), \qquad 1 \leqslant i \leqslant 4,$$

and let $f_2 : X_2 \longrightarrow \mathbf{A}^5$ be the projection

$$f_2(x_{1,1}, \dots, x_{4,k}, r, s, \boldsymbol{b}) = (r, \boldsymbol{b}).$$

By definition, it follows that for all $q \neq \ell$, we have

$$\mathcal{R}_{\lambda=0} = Rf_{2,!}^{4k-3}\mathcal{L}_\psi\Big(\sum_{j=1}^{k}(x_{1,j} + x_{2,j} - x_{3,j} - x_{4,j})\Big).$$

Let $X \subset \mathbf{G}_m^{4k-1} \times \mathbf{A}^6$ be the variety over $\mathbf{Z}[1/\ell]$ with equations

$$\begin{aligned}
\alpha_{1,2} \cdots \alpha_{1,k} &= \beta(r + b_1),\\
\alpha_{2,1} \cdots \alpha_{2,k} &= \beta(r + b_2),\\
\alpha_{3,1} \cdots \alpha_{3,k} &= \beta(r + b_3),\\
\alpha_{4,1} \cdots \alpha_{4,k} &= \beta(r + b_4).
\end{aligned}$$

The morphism $X_2 \longrightarrow \mathbf{G}_m \times X$ given by

$$(x_{1,1}, \dots, x_{4,k}, r, s, \boldsymbol{b}) \mapsto \Big(x_{1,1}, \Big(\frac{x_{1,2}}{x_{1,1}}, \dots, \frac{x_{4,k}}{x_{1,1}}, r, \frac{s}{x_{1,1}^k}, \boldsymbol{b}\Big)\Big)$$

is an isomorphism over $\mathbf{Z}[1/\ell]$. In coordinates $(x_{1,1}, x)$ of $\mathbf{G}_m \times X$, we have

$$\sum_{j=1}^{k}(x_{1,j} + x_{2,j} - x_{3,j} - x_{4,j}) = x_{1,1}g(x),$$

where $g : X \longrightarrow \mathbf{A}^1$ is the morphism

$$(\alpha_{1,2}, \dots, \alpha_{4,k}, r, s, \boldsymbol{b}) \mapsto 1 + \sum_{j=2}^{k}\alpha_{1,j} + \sum_{j=1}^{k}(\alpha_{2,j} - \alpha_{3,j} - \alpha_{4,j}).$$

Similarly, the projection f_2 is $f \circ p_2$ in the coordinates of $\mathbf{G}_m \times X$ where $f : X \longrightarrow \mathbf{A}^5$ is the projection $(\alpha_{1,2}, \dots, \alpha_{4,k}, r, s, \boldsymbol{b}) \longrightarrow (r, \boldsymbol{b})$ and p_2 is the second projection $\mathbf{G}_m \times X \longrightarrow X$. Thus we get

$$\mathcal{R}_{\lambda=0} \simeq R^{4k-3}(f \circ p_2)_!\mathcal{L}_\psi(tg(x))$$

on $\mathbf{A}^5_{\mathbf{F}_q}$ for all $q \neq \ell$.

We can now apply Lemma 4.23 to the variety X, to $Y = \mathbf{A}^5$ and to $f : X \longrightarrow Y$. We deduce the existence of a complex K on $\mathbf{A}^5_{\mathbf{Z}[1/\ell]}$ such that, for $q \neq \ell$,

$$R(f \circ p_2)_!\mathcal{L}_\psi(tg(x)) = K|\mathbf{A}^5_{\mathbf{F}_q},$$

so

$$\mathcal{R}_{\lambda=0}|\mathbf{A}^5_{\mathbf{F}_q} = R^{4k-3}(f\circ p_2)_!\mathcal{L}_\psi(tg(x)) = \mathcal{H}^{4k-3}(K)|\mathbf{A}^5_{\mathbf{F}_q},$$

and we can take $\mathcal{R}^{\mathrm{univ}} = \mathcal{H}^{4k-3}(K)$. □

PROPOSITION 4.28. *For any sufficiently large prime q, the specialized ℓ-adic sum-product sheaf $\mathcal{R}^*_{\lambda=0,\boldsymbol{b}}$ is geometrically irreducible for all $\boldsymbol{b}$ in an open dense subset of $\mathbf{A}^4_{\mathbf{F}_q}$.*

Proof. We will show that $\mathcal{R}^*_{\lambda=0,\boldsymbol{b}}$ is geometrically irreducible at the generic point. By Pink's Specialization Theorem [Kat90, Th. 8.18.2], this will imply the result on an open dense subset. We compactify $\mathbf{A}^4$ (resp. $\mathbf{A}^5$) in $\mathbf{P}^4$ (resp. $\mathbf{P}^4\times\mathbf{P}^1$), and we compactify the projection $\pi\colon \mathbf{A}^5\to\mathbf{A}^4$ using the analogue projection $\bar\pi\colon \mathbf{P}^4\times\mathbf{P}^1\to\mathbf{P}^4$.

Let W be the stalk of $\mathcal{R}^*_{\lambda=0}$ at the generic point of $\mathbf{A}^5$, and let $\varrho\colon G\longrightarrow \mathrm{GL}(W)$ be the corresponding representation of the Galois group

$$G = \mathrm{Gal}(\overline{\mathbf{F}_q(\boldsymbol{b},r)}/\overline{\mathbf{F}}_q(\boldsymbol{b},r)).$$

This representation is irreducible since the sheaf $\mathcal{R}^*_{\lambda=0}$ on $\mathbf{A}^5$ is geometrically irreducible by an application of Lemma 4.14 and Proposition 4.3.

It is then enough to prove that the restriction of ϱ to the normal subgroup

$$G_1 = \mathrm{Gal}(\overline{\mathbf{F}_q(\boldsymbol{b},r)}/\overline{\mathbf{F}_q(\boldsymbol{b})}(r))$$

is also irreducible, since this will show that the fiber of $\mathcal{R}^*_{\lambda=0}$ over the generic point of $\mathbf{A}^4$ is geometrically irreducible. Note that $G/G_1 = \mathrm{Gal}(\overline{\mathbf{F}_q(\boldsymbol{b})}/\overline{\mathbf{F}}_q(\boldsymbol{b}))$.

The quotient G/G_1 acts on the set $\mathcal{W}$ of G_1-invariant subspaces of W. Assume that the action of G_1 on W is *not* irreducible. Then there is some nonzero proper G_1-invariant subspace, which cannot be G-invariant, so the action of G/G_1 on $\mathcal{W}$ is not a trivial action.

Since the tame geometric fundamental group of $\mathbf{A}^4$ is trivial (see, e.g., [Org03, Th. 5.1] — using the fact that the tame fundamental group is independent of the choice of compactification, as explained in loc. cit., and the fact that the tame fundamental group of $\mathbf{A}^1$ is trivial), the action of G/G_1 on $\mathcal{W}$ must be ramified at some codimension 1 point of $\mathbf{A}^4$ or wildly ramified at ∞, in the sense that the inertia group (resp. wild inertia group) at such a point acts nontrivially.

Let D be divisor in $\mathbf{A}^4$ where the action of G/G_1 is ramified. Denote by I_D the corresponding inertia group and by $I_{\bar\pi^{-1}(D)}$ the inertia group of the divisor $\bar\pi^{-1}(D)$. We have the commutative diagram

$$\begin{array}{ccccc} I_{\bar\pi^{-1}(D)} & \longrightarrow & G & \longrightarrow & \mathrm{GL}(W) \\ \downarrow & & \downarrow & & \downarrow \\ I_D & \longrightarrow & G/G_1 & \longrightarrow & \mathrm{Sym}(\mathcal{W}). \end{array}$$

By Lemma 4.25, the homomorphism on the left is surjective. Since I_D acts nontrivially on $\mathcal{W}$, it follows that $I_{\bar{\pi}^{-1}(D)}$ acts nontrivially on W. Hence $\mathcal{R}^*_{\lambda=0}$ is ramified at the pullback of some codimension 1 point of $\mathbf{A}^4$, or wildly ramified at ∞. By Lemma 4.26(3), the monodromy action of $\mathcal{R}^*_{\lambda=0}$ on some dense open set V where it is lisse factors through $\pi_1(U)$. Since $\mathcal{R}^*_{\lambda=0}|V$ is a quotient of $\mathcal{R}_{\lambda=0}|V$, it follows that $\mathcal{R}_{\lambda=0}$ is either ramified at the pullback in $\mathbf{A}^5$ of some codimension 1 point of $\mathbf{A}^4$ or wildly ramified at ∞.

However, if q is sufficiently large, the sheaf $\mathcal{R}_{\lambda=0}$ is not wildly ramified at ∞ because it is defined over $\mathbf{Z}$ (by Lemma 4.27) and hence can only have wild ramification at finitely many primes (as can be seen by applying Abhyankar's Lemma [GR71, Exposé XIII, §5] as in [Kat80, Th. 4.7.1(i)]).

Furthermore, by Lemma 4.26(2), the sheaf $\mathcal{R}_{\lambda=0}$ is lisse outside the complement of the union of the subvariety $\tilde{Z}$ defined in that lemma and the divisors given by the equations $r = -b_i$ in $\mathbf{A}^5$. So the only codimension 1 points where the sheaf is ramified are the generic points of these divisors. The divisors with equation $r = -b_i$ are clearly irreducible, and the same is true of $\tilde{Z}$ by Lemma 4.26(1), so they each contain a single codimension 1 point; thus we will obtain a contradiction if we show that none of these divisors is a pullback from $\mathbf{A}^4$ under the map $(r, \boldsymbol{b}) \mapsto \boldsymbol{b}$.

It is clear that the divisors with equation $r + b_i = 0$ are not pullbacks from $\mathbf{A}^4$. Recall that the divisor $\tilde{Z}$ was defined as the (closed) projection of the subvariety with equation $x_1 + x_2 - x_3 - x_4$ of the subvariety Z of $\mathbf{A}^9$ given by (4.8), and (from Lemma 4.26 (1)) that it is irreducible. This means we will be done if we check that $\tilde{Z}$ is not a pullback from $\mathbf{A}^4$ when q is sufficiently large. For instance, note that $(r, \boldsymbol{b}) = (0, 1, 1, (-1)^k, 3^k)$ is in $\tilde{Z}$, as the image of $(1, 1, -1, 3, 0, 1, 1, (-1)^k, 3^k)$; if $\tilde{Z}$ is a pullback from $\mathbf{A}^4$, we must have also $(-1, \boldsymbol{b}) = (-1, 1, 1, (-1)^k, 3^k) \in \tilde{Z}$, but this is not the case since the corresponding equations for $(x_1, \dots, x_4)$ to be in Z impose

$$\begin{cases} x_1 = x_2 = 0, \\ x_3^k = -1 + (-1)^k, \\ x_4^k = -1 + 3^k, \end{cases}$$

and to be in $\tilde{Z}$ we should have a solution with $x_3 = -x_4$, hence

$$(-1 + (-1)^k) = (-1)^k(-1 + 3^k) \in \mathbf{F}_q.$$

This equation holds only for finitely many primes q. □

PROPOSITION 4.29. *Fix a prime ℓ. There exists a hypersurface $\mathcal{V}^{\mathrm{bad}} \subseteq \mathbf{A}^4_{\mathbf{Z}[1/\ell]}$, containing $\mathcal{V}^{\Delta}$, that is stable under the automorphism $\boldsymbol{b} \mapsto \tilde{\boldsymbol{b}} = (b_3, b_4, b_1, b_2)$ of $\mathbf{A}^4$, and such that, for any sufficiently large prime q, the specialized ℓ-adic sum-product sheaf $\mathcal{R}^*_{\lambda=0,\boldsymbol{b}}$ over $\mathbf{F}_q$ is geometrically irreducible for all $\boldsymbol{b}$ outside $\mathcal{V}^{\mathrm{bad}}$.*

Proof. First we see by Proposition 4.28 that, for a given q sufficiently large, the sheaf $\mathcal{R}^*_{\lambda=0,\boldsymbol{b}}$ is geometrically irreducible for all $\boldsymbol{b}$ outside of some subvariety of codimension $\geqslant 1$ over $\mathbf{F}_q$.

To construct the exceptional subvariety over $\mathbf{Z}[1/\ell]$, we denote by $\sigma : \mathbf{A}^4_{\mathbf{Z}} \longrightarrow \mathbf{A}^4_{\mathbf{Z}}$ the automorphism $(r, \boldsymbol{b}) \mapsto (r, \tilde{\boldsymbol{b}})$. We define the ℓ-adic sheaf $\mathcal{F} = \mathcal{R}^{\mathrm{univ}} \otimes (\mathrm{Id} \times \sigma)^* \mathcal{R}^{\mathrm{univ}}$, where $\mathcal{R}^{\mathrm{univ}}$ is the sheaf on $\mathbf{Z}[1/\ell]$ constructed in Lemma 4.27. This is a constructible sheaf on $\mathbf{A}^5_{\mathbf{Z}[1/\ell]}$. Setting π to be the projection $(r, \boldsymbol{b}) \longrightarrow \boldsymbol{b}$ we define $\mathcal{E} = R^2\pi_!\mathcal{F}$, a constructible ℓ-adic sheaf on $\mathbf{A}^4_{\mathbf{Z}[1/\ell]}$.

Let $U \subset \mathbf{A}^4_{\mathbf{Z}[1/\ell]}$ be the maximal open subset where $\mathcal{E}$ is lisse. Let $H \supset \mathbf{A}^4 - U$ be any codimension 1 closed subscheme of $\mathbf{A}^4_{\mathbf{Z}[1/\ell]}$ containing the complement of U. Then let

$$\mathcal{V}^{\mathrm{bad}} = \mathcal{V}^{\Delta} \cup H \cup \sigma(H).$$

It is clear that $\mathcal{V}^{\mathrm{bad}}$ is stable under σ. We will now show that, for any $q > k$ distinct from ℓ, the specialized sheaf $\mathcal{R}^*_{\lambda=0,\boldsymbol{b}}$ over $\mathbf{F}_q$ is geometrically irreducible for $\boldsymbol{b}$ outside $\mathcal{V}^{\mathrm{bad}}$.

Let such a q be given, and fix $\boldsymbol{b} \in \mathbf{F}_q^4 \notin \mathcal{V}^{\mathrm{bad}}(\mathbf{F}_q)$. We claim that the specialized sheaf $\mathcal{R}^*_{\lambda=0,\boldsymbol{b}}$ is geometrically irreducible if and only if the weight 4 part of the stalk $H^2_c(\mathbf{A}^1_{\overline{\mathbf{F}}_q}, \mathcal{F}_{\boldsymbol{b}})$ of $\mathcal{E}|\mathbf{A}^4_{\mathbf{F}_q}$ at $\boldsymbol{b}$ is one-dimensional. If this is so, then we are done: since mixed lisse sheaves are successive extensions of pure lisse sheaves, the rank of the weight 4 part of $\mathcal{E}$ on the open set where it is lisse is constant. The first part of the argument has shown that this weight 4 part is of rank 1 on some dense open set, so we know it has rank 1 on the open set where it is lisse.

The proof of the claim is similar to the argument in the proof of Theorem 4.11 above. If $U_{\boldsymbol{b}}$ is a dense open subset on which $\mathcal{F}_{\boldsymbol{b}}$ is lisse, we have

$$H^2_c(\mathbf{A}^1_{\overline{\mathbf{F}}_q}, \mathcal{F}_{\boldsymbol{b}}) \simeq (\mathcal{F}_{\boldsymbol{b},\bar{\eta}})_{\pi_1(U_{\boldsymbol{b}})}(-1),$$

so the weight 4 part of the stalk is isomorphic to the weight 2 part of the coinvariants of $\mathcal{F}_{\boldsymbol{b},\bar{\eta}}$. This weight 2 part is isomorphic to the coinvariants of the maximal weight 2 quotient of $\mathcal{F}$, which is $\mathcal{R}^*_{\lambda=0} \otimes [\mathrm{Id} \times \sigma]^* \mathcal{R}^*_{\lambda=0}$. By Lemma 4.5 (and the geometric simplicity of the sheaves), we have a geometric isomorphism

$$\mathcal{R}^{*\vee}_{\lambda=0,\boldsymbol{b}} \simeq \mathcal{R}^*_{\lambda=0,\tilde{\boldsymbol{b}}}$$

on any dense open set where the sheaf is lisse. So the weight 4 part of $H^2_c(\mathbf{A}^1_{\overline{\mathbf{F}}_q}, \mathcal{F}_{\boldsymbol{b}})$ is the same as the coinvariants of $\mathcal{R}^*_{\lambda=0,\boldsymbol{b}} \otimes \mathcal{R}^{*\vee}_{\lambda=0,\boldsymbol{b}}$, which is just the endomorphisms of $\mathcal{R}^*_{\lambda=0}$ as a geometric monodromy representation. Since $\mathcal{R}^*_{\lambda=0}$ is geometrically semisimple, the dimension of this space is 1 if and only if the representation is geometrically irreducible. $\square$

4.5. *Irreducibility of sum-product sheaves for $\lambda \neq 0$.* This section is devoted to the study of the irreducibility of sum-product sheaves for $\lambda \neq 0$. *We always assume that $q > k \geqslant 2$.*

Using Lemma 4.12, we want to show that if $\boldsymbol{b} \notin \mathcal{V}^{\Delta}$, then $\mathcal{R}_{\lambda,\boldsymbol{b}}$ is geometrically irreducible for *all* $\lambda \neq 0$. This is the most delicate part of our argument. The strategy is as follows:

(1) We show that for $\boldsymbol{b} \notin \mathcal{V}^{\Delta}$, the sheaf $\mathcal{R}^*_{\boldsymbol{b}}$ is geometrically irreducible on $\mathbf{A}^2$; this gives the first condition in Lemma 4.12.
(2) Let $\mathbf{0} = (0,0,0,0)$; we compute explicitly the wild part of the monodromy at infinity of $\mathcal{R}_{\lambda,\mathbf{0}}$ for $\lambda \neq 0$.
(3) We show that the wild part of the monodromy at infinity of $\mathcal{R}_{\lambda,\boldsymbol{b}}$ is independent of $\boldsymbol{b}$ (for $\lambda \neq 0$) and thus is known by the previous step; this should be understood intuitively from the fact that for any $\boldsymbol{b} = (b_1, b_2, b_3, b_4)$, the map $r \mapsto (r, r, r, r)$ approximates the map $r \mapsto (r + b_1, r + b_2, r + b_3, r + b_4)$ as $r \to \infty$.
(4) We extend the computation to $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$; this leads to a verification of the second condition of Lemma 4.12.
(5) Finally, we check the last condition of this lemma.

In all of this section, we fix a tuple $\boldsymbol{b} \notin \mathcal{V}^{\Delta}(\mathbf{F}_q)$.

LEMMA 4.30. *For any $\boldsymbol{b} \notin \mathcal{V}^{\Delta}(\mathbf{F}_q)$, the sheaf $\mathcal{R}^*_{\boldsymbol{b}}$ on $\mathbf{A}^2$ is geometrically irreducible on the open subset where it is lisse.*

Proof. The result follows from Lemma 4.14 and (3.4), as in the beginning of the proof of Proposition 4.29. □

LEMMA 4.31. *The following properties hold*:

(1) *the sheaf $\mathcal{R}_{\boldsymbol{b}}$ is lisse on the complement U of the union of the divisor with equation $\lambda = 0$ and of the divisors with equations $r = -b_i$ for $1 \leqslant i \leqslant 4$;*
(2) *the generic rank of $\mathcal{R}_{\boldsymbol{b}}$ is k^4;*
(3) *the generic rank of $\mathcal{R}_{\lambda=0,\boldsymbol{b}}$ is at most k^3.*

Proof. First we prove that $\mathcal{R}_{\boldsymbol{b}}$ is lisse on U. Let

$$i : U \hookrightarrow \mathbf{A}^1 \times \mathbf{G}_m$$

and

$$j : \mathbf{A}^2 \times \mathbf{G}_m \hookrightarrow \mathbf{A}^1 \times \mathbf{P}^1 \times \mathbf{G}_m$$

be the canonical open immersions, and let $\tilde{\mathcal{K}}_{\boldsymbol{b}} = j_!\mathcal{K}_{\boldsymbol{b}}$ be the extension by 0 of $\mathcal{K}_{\boldsymbol{b}}$. Write $\tilde{\pi}$ for the projection $(r, s, \lambda) \mapsto (r, \lambda)$ on $\mathbf{A}^1 \times \mathbf{P}^1 \times \mathbf{G}_m$. Let $\pi = \tilde{\pi} \circ j$, $\tilde{W} = \tilde{\pi}^{-1}(U)$ and $W = \pi^{-1}(U)$ so that W is the preimage of $\tilde{W}$ under j. Also denote by π_W the restriction of π to W.

We note that $\mathcal{K}_{\boldsymbol{b}}$ is lisse on the complement of $s = 0$ in W and vanishes on the divisor $s = 0$. Similarly, $\tilde{\mathcal{K}}_{\boldsymbol{b}}$ is lisse on the complement of the smooth

divisor $\{s=0\}\cup\{s=\infty\}$ in $\tilde{W}$. Moreover, we have

$$\mathcal{R}_{\boldsymbol{b}}|U = R^1\pi_{W!}(\mathcal{K}_{\boldsymbol{b}}) = R^1\overline{\pi}_{W!}(\tilde{\mathcal{K}}_{\boldsymbol{b}}|W),$$

where the point is that we write the restriction of $\mathcal{R}_{\boldsymbol{b}}$ to U as a higher direct image of the restriction of a sheaf lisse outside a smooth divisor.

We next claim that the Swan conductor of $\tilde{\mathcal{K}}_{\boldsymbol{b}}$ is constant along the two divisors $s=0$ and $s=\infty$. Indeed, recall that

$$\mathcal{K}_{\boldsymbol{b}} = \mathcal{L}_{\psi(\lambda s)}\otimes\bigotimes_{i=1}^{2}(f_i^*\mathcal{K}\ell_k\otimes f_{i+2}^*\mathcal{K}\ell_k^\vee) = \mathcal{L}_{\psi(\lambda s)}\otimes\mathcal{G},$$

say. Along the divisor $s=0$, the pullbacks $f_i^*\mathcal{K}\ell_k$ and $f_{i+2}^*\mathcal{K}\ell_k^\vee$ are tamely ramified since the Kloosterman sheaf $\mathcal{K}\ell_k$ is tamely ramified at 0 and $s\mapsto(r+b_i)s$ fixes 0 and ∞. Since $\mathcal{L}_{\psi(\lambda s)}$ is unramified along $s=0$, we see that $\tilde{\mathcal{K}}_{\boldsymbol{b}}$ is tamely ramified and, in particular, has constant Swan conductor equal to zero.

On the other hand, the Kloosterman sheaf $\mathcal{K}\ell_k$ is wildly ramified at ∞ with unique break $1/k$, so the tensor product $\mathcal{G}$ above has all breaks at most $1/k$ at ∞ (again because f_i fixes ∞ as a function of s). Since $k\geqslant 2$ and the single sheaf $\mathcal{L}_{\psi(\lambda s)}$ has rank 1 and is wildly ramified at ∞ with break 1 (recall that $\lambda\neq 0$ in this argument), the sheaf $\tilde{\mathcal{K}}_{\boldsymbol{b}}$ has unique break 1 at ∞. Since the rank of $\tilde{\mathcal{K}}_{\boldsymbol{b}}$ is k^4, the Swan conductor at $s=\infty$ of $\tilde{\mathcal{K}}_{\boldsymbol{b}}$ is the constant k^4. This establishes our claim.

It follows from the above and Deligne's semicontinuity theorem [Lau81, Cor. 2.1.2] that the sheaf $R^1\pi_{W!}(\mathcal{K}_{\boldsymbol{b}})$ is lisse on U. As we observed, this is the same as the restriction of $\mathcal{R}_{\boldsymbol{b}}$ to U, and hence $\mathcal{R}_{\boldsymbol{b}}$ is lisse on U.

Now we consider the rank estimates. By the propre base change theorem, the stalk of $\mathcal{R}$ over $x=(r,\lambda,\boldsymbol{b})\in\mathbf{A}^2\times(\mathbf{A}^4\setminus\mathcal{V}^\Delta)$ is $H_c^1(\mathbf{A}^1_{\overline{\mathbf{F}}_q},\mathcal{F})$, where

$$\mathcal{F} = \mathcal{L}_{\psi(s\lambda)}\otimes\bigotimes_{i=1}^{2}[\times(r+b_i)]^*\mathcal{K}\ell_k\otimes[\times(r+b_{i+2})]^*\mathcal{K}\ell_k^\vee)).$$

We recall from Lemma 4.1 that the 0-th and 2-nd cohomology groups of $\mathcal{F}$ vanish, so that the rank of the stalk of $\mathcal{R}$ at x is minus the Euler–Poincaré characteristic of the sheaf whose cohomology we consider. The Euler–Poincaré formula for a constructible sheaf on $\mathbf{A}^1$ gives

$$\chi(\mathbf{A}^1_{\overline{\mathbf{F}}_q},\mathcal{F}) = \operatorname{rank}(\mathcal{F}) - \sum_{x\in\mathbf{P}^1}\operatorname{Swan}_x(\mathcal{F}) - \sum_{x\in\mathbf{A}^1}\operatorname{drop}_x(\mathcal{F}),$$

where $\operatorname{drop}_x(\mathcal{F})$ is the generic rank of $\mathcal{F}$ minus the dimension of the stalk at x. (See, e.g., [Kat12, p. 67] and the references there, or [Fu11, Cor. 10.2.7].)

Since we normalized the Kloosterman sheaf $\mathcal{K}\ell_k$ to have stalk 0 at 0, so does $\mathcal{F}$, and the above formula becomes

$$\chi(\mathbf{A}^1_{\overline{\mathbf{F}}_q}, \mathcal{F}) = -\sum_{x\in\mathbf{P}^1} \mathrm{Swan}_x(\mathcal{F}) - \sum_{x\in\mathbf{G}_m} \mathrm{drop}_x(\mathcal{F}).$$

(2) In the generic case $\lambda \neq 0$ and $r + b_i \neq 0$, the rank is equal to k^4 since the sheaf $\mathcal{F}$ is then lisse on $\mathbf{G}_m$, tame at 0, and has unique break 1 with multiplicity k^4 at infinity.

(3) If $\lambda = 0$ (and $\boldsymbol{b}$ generic), then we get generic rank $\leqslant \mathrm{Swan}_\infty(\mathcal{F}) \leqslant k^3$, since $\mathcal{F}$ (for $\lambda = 0$) has all breaks $\leqslant 1/k$ at ∞ and rank k^4. □

We now consider the local monodromy of $\mathcal{R}_{\boldsymbol{b}}$ in terms of the r variable for $\lambda \neq 0$. First we deal with the singularity $r = -b_i$.

LEMMA 4.32. *On the open set where $\lambda \neq 0$, the sheaf $\mathcal{R}_{\boldsymbol{b}}$ has tame ramification around the divisors $r = -b_i$ for $1 \leqslant i \leqslant 4$.*

Proof. Let $\mathcal{O}$ be the ring of integers in a finite extension of $\mathbf{Q}_\ell$ such that the sheaves $\mathcal{K}\ell_k$ and $\mathcal{L}_{\psi(\lambda s)}$ have a model over $\mathcal{O}$, in the sense of [Kat88, Remark 1.10], and let ϖ be a uniformizer of $\mathcal{O}$. Then $\mathcal{R}_{\boldsymbol{b}}$ has a model over $\mathcal{O}$ and we have

$$\mathrm{Swan}_{-b_i}(\mathcal{R}_{\boldsymbol{b}}) = \mathrm{Swan}_{-b_i}(\mathcal{R}_{\boldsymbol{b}}/\varpi)$$

for any i (see, e.g., [Kat88, Rem. 1.10]). Thus we reduce to ℓ-torsion sheaves.

We will show that the torsion sheaf $\mathcal{R}_{\boldsymbol{b}}/\varpi$ is trivialized at $-b_i$ after pullback to a covering defined by adjoining n-th roots of $r+b_i$, for some n coprime to q. This implies that $\mathcal{R}_{\boldsymbol{b}}/\varpi$ is tame at $-b_i$ and hence gives our claim.

We fix $\lambda \neq 0$, and we now view f_i as a morphism $\mathbf{A}^1 \times \mathbf{A}^1 \longrightarrow \mathbf{A}^1$ given by $(r, s) \mapsto s(r + b_i)$. Over the étale local ring at 0, the sheaf $\mathcal{K}\ell_k$ is isomorphic (by Proposition 4.6(4)) to the extension by zero of a lisse sheaf $\mathcal{U}$ on $\mathbf{G}_m$ corresponding to a principal unipotent rank k representation of the tame fundamental group

$$\varprojlim_{(n,q)=1} \boldsymbol{\mu}_n(\overline{\mathbf{F}}_q)$$

of $\mathbf{G}_m$. Hence $f_i^*\mathcal{K}\ell_k$ and $f_i^*\mathcal{U}$ are isomorphic after pullback in an étale neighborhood of the divisor D with equation $s(r + b_i) = 0$ in $\mathbf{A}^2$.

The sheaf $\mathcal{U}/\varpi$ corresponds to a representation of the monodromy group of an étale Kummer covering of $\mathbf{G}_m$, defined by adjoining the n-th root of the coordinate for some n coprime to q. Therefore $f_i^*\mathcal{U}/\varpi$ corresponds to a covering of $\mathbf{A}^2$, ramified over D, obtained by adjoining the n-th root of $s(r+b_i)$. It follows that, if we adjoin the n-th root of $r + b_i$, the cover defining $f_i^*\mathcal{U}/\varpi$ becomes isomorphic to the cover obtained by adjoining the n-th root of s of order coprime to q. Consider the map

$$g : \mathbf{A}^2 \longrightarrow \mathbf{A}^2$$

with $g(t,s) = (t^n - b_i, s)$. Then because of this isomorphism of covers, $g^* f_i^* \mathcal{U}/\varpi$ is locally isomorphic to $g^*[(r,s) \mapsto s]^* \mathcal{U}/\varpi = [(t,s) \mapsto s]^* \mathcal{U}/\varpi$. From now on, we will write $s^* \mathcal{U}/\varpi$ for $[(t,s) \mapsto s]^* \mathcal{U}/\varpi$.

The sheaf $g^* f_i^* \mathcal{K}\ell_k$ is lisse on $\mathbf{A}^2$ away from the lines $s = 0$ and $t = 0$. We claim that $g^* f_i^* \mathcal{K}\ell_k$, restricted to the open set $t \neq 0$, may be extended to $\mathbf{A}^2$ as a sheaf $\mathcal{K}\ell'_k$ in such a way that $\mathcal{K}\ell'_k$ is lisse away from the line $s = 0$ and isomorphic to $s^* \mathcal{U}/\varpi$ on the line $t = 0$. This is an étale-local condition, and it may be checked in an étale neighborhood of the line $t = 0$. In fact, since it depends only on the restriction to the open set $t \neq 0$, it may be checked on the complement of the line $t = 0$ in an étale neighborhood of itself. In such a neighborhood, we have the two aforementioned isomorphisms $g^* f_i^* \mathcal{K}\ell_k/\varpi \cong g^* f_i^* \mathcal{U}/\varpi \cong s^* \mathcal{U}/\varpi$. The existence of the desired extension is obvious for $s^* \mathcal{U}/\varpi$ and hence holds for $g^* f_i^* \mathcal{K}\ell_k$. We next denote by $\mathcal{K}\ell^0_k$ the extension by zero to $\mathbf{A}^2$ of the restriction of $\mathcal{K}\ell'_k$ to the complement of the line $s = 0$ in $\mathbf{A}^2$.

We have

$$g^* \mathcal{K}_{\boldsymbol{b}}/\varpi = g^* \mathcal{L}_{\psi(\lambda s)}/\varpi \otimes g^* f_i^* \mathcal{K}\ell_k/\varpi \otimes \bigotimes_{j \neq i} g^* f_j^* \mathcal{K}\ell_k/\varpi.$$

Let

$$\mathcal{K}^0_{\boldsymbol{b}} = g^* \mathcal{L}_{\psi(\lambda s)}/\varpi \otimes \mathcal{K}\ell^0_k \otimes \bigotimes_{j \neq i} g^* f_j^* \mathcal{K}\ell_k/\varpi$$

be the same tensor product but with the $g^* f_i^* \mathcal{K}\ell_k/\varpi$ term replaced with $\mathcal{K}\ell^0_k$. Then $\mathcal{K}^0_{\boldsymbol{b}}$ is lisse on $\mathbf{A}^2$ away from the line $s = 0$ and the lines $t^n - b_i = -b_j$ for $j \neq i$.

The sheaf $R^1 \pi_! \mathcal{K}^0_{\boldsymbol{b}}$ is lisse in an étale neighborhood of 0, by a proof similar to the proof in Lemma 4.31 that $\mathcal{K}$ is lisse. Indeed, $\mathcal{K}^0_{\boldsymbol{b}}$ is lisse near $t = 0$ away from $s = 0$ and $s = \infty$, and tamely ramified at 0, so by Deligne's semicontinuity theorem [Lau81, Cor. 2.1.2] it suffices to check that the Swan conductor of $\mathcal{K}^0_{\boldsymbol{b}}$ at ∞ is constant. The three Kloosterman sheaves all have breaks at ∞ strictly less than 1, and the same is true of $\mathcal{K}^0$ because for $t \neq 0$, it is a Kloosterman sheaf and at $t = 0$ it is unipotent and tame. Thus tensoring with $\mathcal{L}_{\psi(\lambda s)}$, all the breaks become 1 and the Swan conductor is constant.

So the local monodromy at $t = 0$ of $R^1 \pi_! \mathcal{K}^0_{\boldsymbol{b}}$ is trivial. But, by construction, the sheaf $\mathcal{K}^0_{\boldsymbol{b}}$ is isomorphic to $g^* \mathcal{K}_{\boldsymbol{b}}/\varpi$ away from $t = 0$, so the local monodromy of

$$R^1 \pi_! g^* \mathcal{K}_{\boldsymbol{b}}/\varpi = [t \mapsto t^n - b_i]^* R^1 \pi_! \mathcal{K}_{\boldsymbol{b}}/\varpi = [t \mapsto t^n - b_i]^* \mathcal{R}_{\boldsymbol{b}}/\varpi$$

around $t = 0$ is also trivial. Thus $\mathcal{R}_{\boldsymbol{b}}/\varpi$ has trivial local monodromy after adjoining the n-th roots of the uniformizer, and it is tamely ramified, as desired. □

It remains to compute the local monodromy at ∞. For this purpose, we will use the theory of nearby and vanishing cycles. Since this theory is likely to be unfamiliar to analytic number theorists, Appendix A gives a short introduction, with some explanation of its relevance for our purposes.

LEMMA 4.33. *Let $\lambda \neq 0$ be fixed in a field extension (possibly transcendental) of $\mathbf{F}_q$. Let X be the blowup of $\mathbf{P}^1 \times \mathbf{P}^1$ at the point $(r,s) = (\infty, 0)$. Consider the projection $X \longrightarrow \mathbf{P}^1$ given by $(r,s) \mapsto r$. Let $\mathcal{F}$ be the extension by zero of the sheaf $\mathcal{K}_{\lambda,\boldsymbol{b}}$ on $\mathbf{A}^2$ to X, and let $\mathcal{G}$ be the extension by zero of $\mathcal{K}_{\lambda,\mathbf{0}}$ on $\mathbf{A}^2$.*

(1) *The nearby cycles sheaves of $\mathcal{F}$ and $\mathcal{G}$ over $r = \infty$ are locally isomorphic at all $s \neq \infty$ in $\mathbf{P}^1$ and at each point of the exceptional divisor of X.*

(2) *The nearby cycles sheaves of $\mathcal{F}$ and $\mathcal{G}$ over $r = \infty$ have the property that the stalk of $R\Psi\mathcal{F}$ at $s = \infty$, as a representation of the wild inertia group, can be split into summands*

$$\varrho_1, \dots, \varrho_m,$$

and the stalk of $R\Psi\mathcal{G}$ at $s = \infty$, as a representation of the wild inertia group, can be split into summands

$$\varrho'_1, \dots, \varrho'_m$$

such that, for all i, the representations ϱ'_i and ϱ_i of the wild inertia group are isomorphic up to order 2 reparametrizations, in the sense of Definition 4.20.

Remark 4.34. We use the blowup X instead of $\mathbf{P}^1 \times \mathbf{P}^1$ because the argument below would not apply to $\mathbf{P}^1 \times \mathbf{P}^1$: for $(r,s) = (\infty, 0)$, the function $1/(rs)$ does not belong to the maximal ideal. See, e.g., [Har77, pp. 28, 29] for a quick description of blowups.

Proof. (1) Since

$$\mathcal{K}_{\lambda,\boldsymbol{b}} = \mathcal{L}_{\psi(\lambda s)} \otimes \bigotimes_{i=1}^{2} \Big([s \mapsto (r + b_i)s]^* \mathcal{K}\ell_k \Big) \otimes \Big([s \mapsto (r + b_{i+2})s]^* \mathcal{K}\ell_k^\vee \Big),$$

$$\mathcal{K}_{\lambda,\mathbf{0}} = \mathcal{L}_{\psi(\lambda s)} \otimes \bigotimes_{i=1}^{2} \Big([s \mapsto rs]^* \mathcal{K}\ell_k \Big) \otimes \Big([s \mapsto rs]^* \mathcal{K}\ell_k^\vee \Big),$$

on $\mathbf{A}^2$, the étale-local nature of nearby cycles shows that it is enough to prove that, for $1 \leqslant i \leqslant 4$, the sheaf $[s \mapsto (r + b_i)s]^* \mathcal{K}\ell_k$ is locally isomorphic to $[s \mapsto rs]^* \mathcal{K}\ell_k$ on $\mathbf{A}^1 - \{0\} \subset \mathbf{P}^1$ (with coordinate s) and on the exceptional divisor D of the blowup.

For points not on the exceptional divisor, we apply Lemma 4.17(2) to the strict henselization R of the local ring at $(\infty, s) \in X$, with $a = rs$ and $b = b_i s$,

where r and s are now viewed as elements of the field of fractions of R. Note that r^{-1} belongs then to the maximal ideal $\mathfrak{m}$ of R (since we are considering the situation at $r = \infty$) and s is a unit (since we are outside the exceptional divisor), hence a^{-1} also belongs to $\mathfrak{m}$. Moreover $b \in R$, and therefore we obtain

$$(a+b)^*\mathcal{K}\ell_k \simeq a^*\mathcal{K}\ell_k,$$

which is the desired conclusion.

The exceptional divisor D is isomorphic to $\mathbf{P}^1$ by the map $(r,s) \mapsto s/r^{-1} = rs$. Hence, for all points x on D except the point mapping to ∞ under this isomorphism, the function rs is a function in the local ring at x, and we may apply Lemma 4.17(1) to the strict henselization R of the local ring at that point, with $a = rs$ and $b = b_i/r$. The function $1/(rs)$ vanishes at the point mapping to ∞, thus is in the maximal ideal, so we may again use Lemma 4.17(2) with $a = rs$ and $b = b_i s$.

(2) We denote again by R the strict henselization of the local ring at $(\infty,\infty) \in \mathbf{P}^1 \times \mathbf{P}^1$ and by $\mathfrak{m}$ its maximal ideal. We also denote by R_1 the strict henselization of the local ring at ∞ of $\mathbf{P}^1$ (with coordinate r), and by $\mathfrak{m}_1$ its maximal ideal. Then $1/r$ is a uniformizer or R_1. Let R_0 be the extension of R_1 generated by a k-th root $1/\varrho$ of $1/r$. Let $U = \operatorname{Spec} R_0[\varrho]$. Since $1 + b_i/r \equiv 1 \pmod{\mathfrak{m}}$, there exists $y_i \in R_1 \subset R$ with $y_i^k = 1 + b_i/r$ and $y_i \equiv 1 \pmod{\mathfrak{m}_1}$. We can apply Lemma 4.16 to U, where f is the projection to $\operatorname{Spec} R_1 - \{\infty\}$ composed with the inclusion $\operatorname{Spec} R_1 - \{\infty\} \to \mathbf{A}^1 - \{-b_1, \dots, b_4\}$ and $r_i = \varrho y_i$. We observe that $r_1 r_2 r_3 r_4 = \varrho^4 y_1 y_2 y_3 y_4$ is a perfect square, as y_1, y_2, y_3, y_4 are all units in R_1, hence squares in R_1 and thus squares in R_0. Hence, by Lemma 4.16, we have an isomorphism of local monodromy representations

$$[f \times \mathrm{Id}]^*\Big(\mathcal{L}_{\psi(\lambda s)} \otimes \bigotimes_{i=1}^{2} f_i^*\mathcal{K}\ell_k \otimes f_{i+2}^*\mathcal{K}\ell_k^\vee\Big) \simeq \bigoplus_{\zeta_2,\zeta_3,\zeta_4 \in \boldsymbol{\mu}_k} \mathcal{L}_{\psi(\lambda s)} \otimes \mathcal{L}_{\tilde{\psi}}\Big(s^{1/k}\varrho\Big(y_1 + \zeta_2 y_2 - \zeta_3 y_3 - \zeta_4 y_4\Big)\Big),$$

where $\tilde{\psi}(x) = \psi(kx)$ as before.

The complex of nearby cycles is preserved by this pullback to a k-th power covering, as is the action of the wild inertia subgroup (because the action of the full inertia group is restricted to the inertia group of the covering, which contains the wild inertia group).

Since the nearby cycle functor is additive, we have a local isomorphism

$$[f \times \mathrm{Id}]^* R\Psi\mathcal{K}_{\lambda,\boldsymbol{b}} \simeq \bigoplus_{\zeta_2,\zeta_3,\zeta_4 \in \boldsymbol{\mu}_k} R\Psi\Big(\mathcal{L}_{\psi(\lambda s)} \otimes \mathcal{L}_{\tilde{\psi}}\Big(s^{1/k}\varrho\Big(y_1 + \zeta_2 y_2 - \zeta_3 y_3 - \zeta_4 y_4\Big)\Big)\Big),$$

and we handle each term in the sum separately. We will show that, for each $(\zeta_2, \zeta_3, \zeta_4)$, either the corresponding component has no nearby cycles for *any*

$\boldsymbol{b} \in \mathbf{A}^4$ (not only for $\boldsymbol{b} \notin \mathcal{V}^\Delta$), or that its nearby cycles, with the action of the wild inertia group, are independent of $\boldsymbol{b} \in \mathbf{A}^4$, up to reparametrizations of order 2. We consider two cases.

Case 1. Assume that $1 + \zeta_2 = \zeta_3 + \zeta_4$. In that case, the element

$$y_1 + \zeta_2 y_2 - \zeta_3 y_3 - \zeta_4 y_4$$

of R belongs to the maximal ideal. Since ϱ^{-1} is a uniformizer of R_0, the element

$$\varrho(y_1 + \zeta_2 y_2 - \zeta_3 y_3 - \zeta_4 y_4)$$

belongs to R_0. Thus the sheaves

$$\mathcal{L}_{\psi(\lambda s)}, \qquad \mathcal{L}_{\tilde{\psi}}\left(s^{1/k}\varrho\,(y_1 + \zeta_2 y_2 - \zeta_3 y_3 - \zeta_4 y_4)\right)$$

both extend to lisse sheaves in an étale neighborhood of (∞, ∞) away from the line $s = \infty$.

To check that their tensor product has no vanishing cycles, it suffices (by Deligne's semicontinuity theorem once more [Lau81, Th. 5.1.1]) to check that the Swan conductor is constant. But the breaks at infinity (in terms of s) of

$$\mathcal{L}_{\tilde{\psi}}\left(s^{1/k}\varrho\,(y_1 + \zeta_2 y_2 - \zeta_3 y_3 - \zeta_4 y_4)\right)$$

are all $\leqslant 1/k$, while $\mathcal{L}_{\psi(\lambda s)}$ has break 1, so the tensor product has all breaks equal to 1, and we are done.

Case 2. Assume that $1 + \zeta_2 \neq \zeta_3 + \zeta_4$. Then we have

$$y_1 + y_2\zeta_2 - y_3\zeta_3 - y_4\zeta_4 = (1 + \zeta_2 - \zeta_3 - \zeta_4)d,$$

where $d \in R_1$ satisfies $d \equiv 1 \,(\mathrm{mod}\, \mathfrak{m}_1)$. Let $\mu = \varrho d$ and $u = rd^k = \mu^k$. Then we have

$$\varrho(y_1 + y_2\zeta_2 - y_3\zeta_3 - y_4\zeta_4) = \mu(1 + \zeta_2 - \zeta_3 - \zeta_4).$$

So, after pulling back to U (which is also the cover defined by adjoining μ), we are dealing with the sheaf

$$\mathcal{L}_{\psi(\lambda s)} \otimes \mathcal{L}_{\tilde{\psi}}\Big(s^{1/k}\mu(1 + \zeta_2 - \zeta_3 - \zeta_4)\Big).$$

The wild inertia action on the nearby cycles of this sheaf, in terms of the variable u, can be computed on the pullback to the cover defined by μ with $\mu^k = u$. Thus it is independent of $\boldsymbol{b} \in \mathbf{A}^4$, because this formula for the pullback is independent of $\boldsymbol{b}$ and the cover is also independent of $\boldsymbol{b}$.

Since $1/r$ and $1/u$ are uniformizers of R_1, there is a unique automorphism σ of R_1 sending r to u. Since $d \equiv 1 \,(\mathrm{mod}\, \mathfrak{m}_1)$, it follows that

$$\frac{1}{u} \equiv \frac{1}{r} \,(\mathrm{mod}\, (1/r)^2),$$

and hence σ is a reparametrization of order 2 (see Definition 4.20). This is the desired result. □

We will describe the wild part of the local monodromy at $r = \infty$ of $\mathcal{R}_{\lambda,\boldsymbol{b}}$ using the following data.

Definition 4.35. Let $k \geqslant 2$, and let q be a prime with $q \nmid k$. We denote by S_k the multiset of *nonzero* elements of $\overline{\mathbf{F}}_q$ of the form

$$(1 + \zeta_2 - \zeta_3 - \zeta_4)^k,$$

where ζ_2, ζ_3 and ζ_4 range over $\boldsymbol{\mu}_k(\overline{\mathbf{F}}_q)$.

We first use this definition to treat the local monodromy for $\mathcal{R}_{\lambda,\mathbf{0}}$.

LEMMA 4.36. *Let $\lambda \neq 0$ be fixed in a field extension (possibly transcendental) of $\mathbf{F}_q$. The local monodromy representation of $\mathcal{R}_{\lambda,\mathbf{0}}$ at $r = \infty$ is isomorphic to that of the sheaf*

$$\bigoplus_{\alpha \in S_k} [\times \alpha\lambda^{-1}]^* \mathcal{H}_{k-1},$$

where $\mathcal{H}_{k-1}$ is the sheaf defined in Definition 4.18, *plus a tamely ramified representation.*

The meaning of the direct sum over the multiset S_k is

$$\bigoplus_{\substack{\zeta_2,\zeta_3,\zeta_4 \in \boldsymbol{\mu}_k \\ 1+\zeta_2-\zeta_3-\zeta_4 \neq 0}} [\times((1 + \zeta_2 - \zeta_3 - \zeta_4)^k \lambda)^{-1}]^* \mathcal{H}_{k-1},$$

and similarly below.

Proof. Note that every representation of the inertia group is a sum of a wildly ramified representation and a tamely ramified representation, as the wild part is a q-group, so has semisimple ℓ-adic representation theory, hence every representation of the wild inertia group splits canonically into trivial and nontrivial parts. Thus, because $\mathcal{H}_{k-1}$ is totally wild at ∞, we concern ourselves only with the wild summand.

The change of variable

$$(r, s) \mapsto (\lambda/r, rs)$$

is an isomorphism $\mathbf{G}_m \times \mathbf{A}^1 \longrightarrow \mathbf{G}_m \times \mathbf{A}^1$ (with inverse $(\xi, x) \mapsto (\lambda/\xi, x\xi/\lambda)$). In terms of the variables (ξ, x), the sheaf $\mathcal{R}_{\lambda,\mathbf{0}}$ becomes the Fourier transform with respect to ψ of the sheaf

$$\mathcal{F} = \bigotimes_{i=1}^{2} \mathcal{K}\ell_k \otimes \overline{\mathcal{K}\ell_k},$$

on $\mathbf{A}^1$ with coordinate x, reflecting the trace function identity

$$\sum_{s \in \mathbf{F}_q} \psi(\lambda s) \prod_{i=1}^{2} \mathrm{Kl}_k(rs)\overline{\mathrm{Kl}_k(rs)} = \sum_{x \in \mathbf{F}_q} \psi(x\xi) \prod_{i=1}^{2} \mathrm{Kl}_k(x)\overline{\mathrm{Kl}_k(x)}.$$

We now need to compute the local monodromy at $\xi = 0$ of this Fourier transform, which we can do using Laumon's local Fourier transform functors. Laumon's results (see, e.g., [Kat90, Th. 7.4.3 and Cor. 7.4.3.1]) give an isomorphism

$$\mathcal{R}_{\lambda,\mathbf{0}}/(\mathcal{R}_{\lambda,\mathbf{0}})_0 \simeq \mathrm{FT}_\psi \operatorname{loc}(\infty, 0)(\mathcal{F}(\infty))$$

of representations of the inertia group at 0, where $(\mathcal{R}_{\lambda,\mathbf{0}})_0$ is the stalk at 0 and $\mathcal{F}(\infty)$ is the local monodromy representation of $\mathcal{F}$ at ∞. Since the stalk at 0 is a trivial representation of the inertia group, this implies that the wild summand of the local monodromy is the same as that of $\mathrm{FT}_\psi \operatorname{loc}(\infty, 0)(\mathcal{F}(\infty))$.

Using Lemma 4.9 as in Lemma 4.16, the local monodromy at ∞ of $\mathcal{F}$ is isomorphic to that of

$$\bigoplus_{\zeta_2,\zeta_3,\zeta_4 \in \boldsymbol{\mu}_k} \mathcal{L}_{\tilde{\psi}}\left(x^{1/k}(1+\zeta_2-\zeta_3-\zeta_4)\right) = \bigoplus_{\zeta_2,\zeta_3,\zeta_4 \in \boldsymbol{\mu}_k} \mathcal{L}_{\tilde{\psi}}\Big(\Big((1+\zeta_2-\zeta_3-\zeta_4)^k x\Big)^{1/k}\Big),$$

where $\tilde{\psi}(x) = \psi(kx)$. All triples $(\zeta_2, \zeta_3, \zeta_4)$ with $1+\zeta_2-\zeta_3-\zeta_4 = 0$ give tamely ramified local monodromy, whose local Fourier transform at 0 is also tamely ramified (see, e.g., [Kat90, Th. 7.4.4 (3)]), so they do not contribute to the wild part of the local monodromy.

Otherwise, if $\alpha = (1+\zeta_2-\zeta_3-\zeta_4)^k \neq 0$ is an element of S_k, then we have the following isomorphisms of local monodromy representations at 0 in $\mathbf{A}^1$ with (Fourier) coordinate ξ, using the definition of the sheaf $\mathcal{H}_{k-1}$:

$$\begin{aligned}
\mathrm{FT}_\psi \operatorname{loc}(\infty, 0)(\mathcal{L}_{\tilde{\psi}}((\alpha x)^{1/k})) &= [\times\alpha^{-1}]^* R\Phi_{\eta_0} \mathrm{FT}_\psi(\mathcal{L}_{\tilde{\psi}}(x^{1/k})) \\
&\simeq [\times\alpha^{-1}]^* R\Phi_{\eta_0}([\xi \mapsto \xi^{-1}]^* \mathcal{H}_{k-1}) \\
&\simeq [\times\alpha^{-1}]^* [\xi \mapsto \xi^{-1}]^* \mathcal{H}_{k-1,\eta_\infty} \\
&\simeq [\xi \mapsto (\alpha/\xi)]^* \mathcal{H}_{k-1,\eta_\infty}.
\end{aligned}$$

(It is important to note that when composing pullbacks, one applies the leftmost functions first, since this is the opposite order from the usual composition of functions, where the rightmost is applied first. So ξ is sent to $\alpha^{-1}\xi$, which is sent to $(\alpha^{-1}\xi)^{-1} = \alpha/\xi$.) Since $\xi = \lambda/r$, this concludes the proof. □

We can finally conclude

Corollary 4.37. *Let $\lambda \neq 0$ be fixed in a field extension (possibly transcendental) of $\mathbf{F}_q$. The wild inertia representation of $\mathcal{R}_{\lambda,\boldsymbol{b}}$ at $r = \infty$ is the same as that of the sheaf*

$$\bigoplus_{\alpha \in S_k} [\times\alpha\lambda^{-1}]^* \mathcal{H}_{k-1}$$

plus a trivial representation.

For the proof, we use the same notation as in Lemma 4.33. Thus, X denotes the blowup of $\mathbf{P}^1 \times \mathbf{P}^1$ at $(\infty, 0)$ and $\mathcal{F}$ and $\mathcal{G}$ on X are the extensions by 0 from $\mathbf{A}^1 \times \mathbf{A}^1$ to X of $\mathcal{K}_{\lambda,\boldsymbol{b}}$ and $\mathcal{K}_{\lambda,\mathbf{0}}$ respectively. Let π be the proper map

$$X \longrightarrow \mathbf{P}^1 \times \mathbf{P}^1 \longrightarrow \mathbf{P}^1,$$

where the second map is the proper projection $(r, s) \mapsto r$.

We need to compute the wild inertia representations at ∞ of $R\pi_*\mathcal{F}$ and $R\pi_*\mathcal{G}$. To do that, we use the nearby cycles $R\Psi\mathcal{F}$ and $R\Psi\mathcal{G}$ relative to π. These are complexes of sheaves with an inertia group action on the fiber over ∞ over X. We know by Lemma 4.33 that $R\Psi\mathcal{F}$ and $R\Psi\mathcal{G}$ are locally isomorphic away from the point (∞, ∞).

The key step is the following sub-lemma:

LEMMA 4.38. *Away from* (∞, ∞), *the wild inertia group acts trivially on* $R\Psi\mathcal{G}$.

Proof. Let $f_1(r, s) = rs$. By definition, we have

$$\mathcal{K}_{\lambda,\mathbf{0}} = \mathcal{L}_{\psi(\lambda s)} \otimes f_1^*\mathcal{K}\ell_k^{\otimes 2} \otimes (f_1^*\overline{\mathcal{K}\ell_k}^{\vee})^{\otimes 2}.$$

Because we are verifying a local condition away from the line $s = \infty$, we may ignore the factor $\mathcal{L}_{\psi(\lambda s)}$ and consider only the nearby cycles of

$$\mathcal{K} = f_1^*\mathcal{K}\ell_k^{\otimes 2} \otimes (f_1^*\overline{\mathcal{K}\ell_k}^{\vee})^{\otimes 2}.$$

For any $\alpha \in \mathbf{G}_m$, let s_α be the map $(r, s) \mapsto (\alpha r, \alpha^{-1}s)$. We have $f_1 \circ s_\alpha = f_1$, hence $s_\alpha^*\mathcal{K} \simeq \mathcal{K}$. The action of s_α extends to the blowup X and to the fiber of X over ∞, so it extends by functoriality to the nearby cycles complex $R\Psi\mathcal{K}$. Since s_α acts by scaling on the coordinate r of the base local ring, the induced isomorphism $s_\alpha^* R\Psi\mathcal{K} \simeq R\Psi\mathcal{K}$ sends the Galois action on the nearby cycles complex to its multiplicative translate by α. Since the nearby cycles sheaf is constructible [Del77, Th. Finitude, Th. 3.2], only finitely many different irreducible representations of the inertia group can appear in the stalks of $R\Psi\mathcal{K}$ as Jordan–Hölder factors anywhere on the fiber over ∞. (On each open set where $R\Psi\mathcal{K}$ is lisse, there is a single representation with finitely many Jordan–Hölder factors, and at each other point there is another representation, again with finitely many Jordan–Hölder factors.) By symmetry, if any irreducible inertia representation appears in the stalks, its multiplicative translates by α must also appear. But by [Kat88, 4.1.6], any nontrivial wildly ramified representation has infinitely many nonisomorphic multiplicative translates as α varies, so the wild inertia group must act trivially on the stalks.

Let I_1 be the wild inertia group. There is an I_1-invariants functor from ℓ-adic sheaves with an action of I_1 to ℓ-adic sheaves, and an adjoint functor that views ℓ-adic sheaves as ℓ-adic sheaves with a trivial action of I_1, giving

a natural adjunction map $(R\Psi\mathcal{G})^{I_1} \to R\Psi\mathcal{G}$. Because I_1 is a pro-q group, the I_1-invariants functor on ℓ-adic sheaves has no higher cohomology. Because the stalks are I_1-invariant, this map is an isomorphism on stalks away from (∞,∞), hence an isomorphism away from (∞,∞), so the wild inertia group acts trivially on $R\Psi\mathcal{G}$ away from (∞,∞). □

Proof of Corollary 4.37. It follows from Lemmas 4.38 and 4.33(1) that the wild inertia group acts trivially on $R\Psi\mathcal{F}$ away from (∞,∞).

Let $Z = \{(\infty,\infty)\}$, and let U be the open complement. Let i be the closed immersion of Z and j the open immersion of U. We have distinguished triangles

$$R\pi_* j_! R\Psi\mathcal{F}|U \to R\pi_* R\Psi\mathcal{F} \to R\pi_* i_* R\Psi\mathcal{F}|Z \to$$

and

$$R\pi_* j_! R\Psi\mathcal{G}|U \to R\pi_* R\Psi\mathcal{G} \to R\pi_* i_* R\Psi\mathcal{G}|Z \to .$$

The middle terms are the local monodromy representations of $R\pi_*\mathcal{F}$ and $R\pi_*\mathcal{G}$ at ∞, which we want to compute. The third terms are the stalks of $R\Psi\mathcal{F}$ and $R\Psi\mathcal{G}$ at (∞,∞). The left-hand terms, by the above, have trivial wild inertia action at ∞.

Since the representations of the wild inertia group are semisimple (as it is a pro-q-group acting on an ℓ-adic vector space), this implies that the nontrivial part of the wild inertia representation on the local monodromy of $R\pi_*\mathcal{F}$ and of $R\pi_*\mathcal{G}$ are each equal to the nontrivial parts of the wild inertia representation on the stalks of $R\Psi\mathcal{F}$ and $R\Psi\mathcal{G}$ at (∞,∞). By Lemma 4.33(2), the stalks of $R\Psi\mathcal{F}$ and $R\Psi\mathcal{G}$ at (∞,∞) can be split into summands that are isomorphic as representations of the wild inertia group up to order 2 reparametrizations, so the nontrivial parts of the wild inertia representations on $R\pi_*\mathcal{F}$ and $R\pi_*\mathcal{G}$ can be split into summands that are equal up to order 2 reparametrizations.

Finally, Lemma 4.36 shows that the wild inertia representation at ∞ of $R\pi_*\mathcal{G}$ is exactly as claimed in the statement. Since, by Lemma 4.21, any summand of the local monodromy at ∞ of $R\pi_*\mathcal{G}$ (i.e., of $\mathcal{R}_{\lambda,\mathbf{0}}$) is preserved by reparametrizations of order 2, we obtain in fact the same decomposition for $\mathcal{R}_{\lambda,\boldsymbol{b}}$ also. □

Corollary 4.39. *Let $\lambda \neq 0$ be fixed in a field extension (possibly transcendental) of $\mathbf{F}_q$. The wild inertia representation of $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ at $r = \infty$ is isomorphic to that of*

$$\bigoplus_{\alpha \in S_k} [\times \alpha/\lambda]^* \mathcal{H}_{k-1}$$

plus a trivial representation.

Proof. In view of Corollary 4.37 and of the definition of $\mathcal{R}^*$, it suffices to prove that the weight < 1 part of $\mathcal{R}_{\boldsymbol{b}}$ is tamely ramified at $r = \infty$. To do this

we will study the action of the decomposition group at ∞ on the stalk of the weight < 1 part of $\mathcal{R}_{\lambda \boldsymbol{b}}$ at a generic point of the r-line.

We apply Lemma 4.22(2) to $C = \mathbf{A}^1 \times \mathbf{P}^1 \times \mathbf{G}_m$, with coordinates (r, s, λ), with its dense open subset $U = \mathbf{A}^1 \times \mathbf{A}^1 \times \mathbf{G}_m$ (with open embedding j), to the morphism $\pi : C \longrightarrow \mathbf{A}^1 \times \mathbf{G}_m$ given by $\pi(r, s, \lambda) = (r, \lambda)$ and to the sheaf $\mathcal{F} = j_!\mathcal{K}_{\boldsymbol{b}}$ on C. The assumptions of Lemma 4.22 are easily verified using Lemma 4.1(2) and (3).

Taking $x = (r, \lambda)$ for a generic value of r, the lemma implies that the part of weight < 1 of

$$(R^1\pi_*\mathcal{F})_x = H^1(\pi^{-1}(\bar{x}), \mathcal{F}) = (\mathcal{R}_{\boldsymbol{b}})_x$$

is isomorphic to

$$\mathcal{K}_{x,\boldsymbol{b}}^{I(0)}/(\mathcal{K}_{x,\boldsymbol{b}})_0 \oplus \mathcal{K}_{x,\boldsymbol{b}}^{I(\infty)}/(\mathcal{K}_{x,\boldsymbol{b}})_\infty = \mathcal{K}_{x,\boldsymbol{b}}^{I(0)},$$

since $\mathcal{K}_{x,\boldsymbol{b}}$ is totally wildly ramified at $s = \infty$ and has stalk 0 at $s = 0$.

Recall that the local monodromy representation of $\mathcal{K}\ell_k$ at 0 is unipotent. Let K be an algebraically closed field extension of $\mathbf{F}_q$ containing λ, so that over K the decomposition group representation of $\mathcal{K}\ell_k$ at 0 is unipotent. Hence the decomposition group representation of

$$[(r, s) \mapsto s(r + b_i)]^*\mathcal{K}\ell_k$$

at a point where $s = 0$ is unipotent (still over K). The decomposition group representation of $\mathcal{L}_{\psi(\lambda s)}$ is trivial at a point where $s = 0$. Hence we conclude that the decomposition group over K also acts unipotently on the tensor product $\mathcal{K}_{x,\boldsymbol{b}}$. Hence the space of inertia invariants $\mathcal{K}_{x,\boldsymbol{b}}^{I(0)}$ is a unipotent representation of the Galois group of the residue field of the generic point x. In particular, the inertia group at $r = \infty$ acts unipotently. Because it is unipotent, it must factor through a pro-ℓ group and hence be tame. □

We need some last elementary geometric considerations to isolate features of the local monodromy at ∞ that will allow us to deduce the irreducibility and disjointness of the sheaves $\mathcal{R}_{\lambda,\boldsymbol{b}}$.

LEMMA 4.40. *Let $k \geqslant 2$ be given.*

(1) *If q is sufficiently large, then the multiset S_k contains an element with multiplicity* 1.
(2) *If q is sufficiently large, then the group of $\mu \in \overline{\mathbf{F}}_q^\times$ such that $\mu S_k = S_k$ is trivial if k is even and is reduced to $\{\pm 1\}$ if k is odd.*

Proof. We denote by $\tilde{S}_k \subset \mathbf{C}$ the analogue of S_k defined using $\boldsymbol{\mu}_k(\bar{\mathbf{Q}})$. We observe that the set of nonzero numbers $\zeta_1 + \zeta_2 - \zeta_3 - \zeta_4$, where ζ_i runs over $\boldsymbol{\mu}_k(\overline{\mathbf{F}}_q)$, is the set of k-th roots of the elements of S_k, and similarly for $\tilde{S}_k$ and $\zeta_i \in \boldsymbol{\mu}_k(\bar{\mathbf{Q}})$. Moreover, the nonzero element of this form has the same

multiplicity as its k-th power as an element of $\tilde{S}_k$. Indeed, there is a bijection from the set of representations

$$\alpha = \zeta_1 + \zeta_2 - \zeta_3 - \zeta_4$$

to those of

$$\alpha^k = (1 + \zeta_2' - \zeta_3' - \zeta_4')^k,$$

given by

$$(\zeta_1, \dots, \zeta_4) \mapsto (\zeta_2/\zeta_1, \zeta_3/\zeta_1, \zeta_4/\zeta_1)$$

with inverse

$$(\zeta_2', \zeta_3', \zeta_4') \mapsto (\zeta, \zeta\zeta_2', \zeta\zeta_3', \zeta\zeta_4'),$$

where ζ is such that $\alpha = \zeta(1 + \zeta_2' - \zeta_3' - \zeta_4')$.

(1) Since any two distinct elements of $\tilde{S}_k$ are equal modulo q for finitely many primes q, it is enough to check that the set $\tilde{S}_k$ contains an element of multiplicity 1 in $\mathbf{C}$. To find an element of $\tilde{S}_k$ with multiplicity 1, it is sufficient to find an $\mathbf{R}$-linear map $\mathbf{C} \longrightarrow \mathbf{R}$ with a unique maximum and minimum on $\boldsymbol{\mu}_k$. Clearly a generic linear function has this property; e.g., if k is even, we may take the real part.

(2) We first show the corresponding property for $\tilde{S}_k$. Let T_k be the multiset of numbers $\zeta_1 + \zeta_2 - \zeta_3 - \zeta_4$. By the description above, it is enough to show that the group of complex numbers μ such that $\mu T_k = T_k$ is equal to $\boldsymbol{\mu}_k$ if k is even and to $\boldsymbol{\mu}_{2k}$ if k is odd.

Consider the convex hull of T_k. It is the difference of two copies of twice the convex hull of the k-th roots of unity. Since the convex hull of $\boldsymbol{\mu}_k$ in $\mathbf{C}$ is a k-sided regular polygon, the convex hull of T_k is a k-sided regular polygon if k is even, and a $2k$-sided regular polygon if k is odd. The result is then clear.

To reduce the case of S_k to the complex case, we note that an arbitrary nonempty finite set $S \subset \mathbf{C}$ or $S \subset \overline{\mathbf{F}}_q$ may only be equal to its multiplicative translate by μ if μ is a root of unity. Moreover, $\mu S = S$, where μ is a primitive n-th roots of unity, if and only if the coefficients of a monic polynomial whose roots are S vanish in degrees coprime to n. When reducing a polynomial with algebraic coefficients modulo a prime q large enough, the only degrees that are zero modulo q are those that are zero in $\mathbf{C}$. Hence, for q large enough, the same roots of unity stabilize S_k as $\tilde{S}_k$. □

Finally we can conclude the basic irreducibility statement for sum-product sheaves when λ is nonzero:

PROPOSITION 4.41. *For q large enough in terms of k, the sheaf $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ is geometrically irreducible whenever $\lambda \neq 0$ is fixed in a field extension (possibly transcendental) of $\mathbf{F}_q$ and $\boldsymbol{b} \notin \mathcal{V}^\Delta$.*

Proof. We apply Lemma 4.12(b) to $Y = \mathbf{G}_m \times \mathbf{P}^1$ where the coordinate of $\mathbf{G}_m$ is λ and the coordinate of $\mathbf{P}^1$ is r and to the first projection $f : Y \longrightarrow X = \mathbf{G}_m$. We consider the sheaf on Y that is the extension by zero of the sheaf $\mathcal{R}^*_{\boldsymbol{b}}$ on $\mathbf{G}_m \times \mathbf{A}^1$. The divisor D is the union of the divisors $\{r = -b_i\}$ and $\{r = \infty\}$. If the three conditions of Lemma 4.12 hold, then we obtain our desired conclusion.

By Lemma 4.14 and (3.4), the sheaf $\mathcal{R}^*_{\boldsymbol{b}}$ is geometrically irreducible on $Y - D$, so that the first condition holds. It is also pure on $Y - D$ by definition.

Next, we will show that the second condition holds by showing that there exists an irreducible component of multiplicity 1 in the local monodromy at ∞ of the restriction of $\mathcal{R}^*_{\boldsymbol{b}}$ to the fiber of f over a geometric generic point of $\mathbf{G}_m$ whose isomorphism class is Galois-invariant.

By Corollary 4.39, the wild inertia representation of $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ at $r = \infty$ is isomorphic to that of $\bigoplus_{\alpha \in S_k} [\times \alpha/\lambda]^* \mathcal{H}_{k-1}$ plus a trivial representation. By Lemma 4.40(1), assuming q is large enough, some α appears with multiplicity 1 in S_k. Take such an α. Let V be the subspace of that local monodromy representation that is sent to $[\times \alpha/\lambda]^* \mathcal{H}_{k-1}$ under this isomorphism.

The irreducible components of the summands $[\times \alpha/\lambda]^* \mathcal{H}_{k-1}$ as representations of the wild inertia group are disjoint by Lemma 4.21(3). So we may characterize V as the subspace generated by all representations of the wild inertia group that are isomorphic to wild inertia representations that appear in $[\times \alpha/\lambda]^* \mathcal{H}_{k-1}$. Because $[\times \alpha/\lambda]^* \mathcal{H}_{k-1}$ is a representation of the full decomposition group, that set of isomorphism classes is stable under the action of the decomposition group, so V is a subrepresentation of the local monodromy representation as a representation of the full decomposition group. (Here we work over a large enough finite field so that all of S_k, including α, is contained in the base field.)

We will show that V, restricted to the inertia group, is irreducible. Restricted to the wild inertia group, it is isomorphic to $[\times \alpha/\lambda]^* \mathcal{H}_{k-1}$. By Lemma 4.21(2), the action by conjugation of the tame inertia group on the irreducible wild inertia subrepresentations of $[\times \alpha/\lambda]^* \mathcal{H}_{k-1}$ is transitive. Thus any subspace would be a sum of wild inertia characters and would be invariant under the tame inertia subgroup. So it must contain all the characters or none, and therefore V is indeed irreducible.

Then the irreducible representation V occurs with multiplicity 1 because each wild inertia component in it occurs with multiplicity 1, and its isomorphism class is invariant under conjugation by the Galois group because it extends to a representation of the full decomposition group.

For the third condition of Lemma 4.12, it is enough to show that the functions

$$\lambda \mapsto \mathrm{Swan}_r(\mathcal{R}^*_{\lambda,\boldsymbol{b}} \otimes \mathcal{R}^{*\vee}_{\lambda,\boldsymbol{b}})$$

are locally constant on the divisors $r = -b_i$ and $r = \infty$. By Lemma 4.32, this function is constant (equal to 0) on the divisors $r = -b_i$ for $1 \leqslant i \leqslant 4$. The Swan conductor is determined by the restriction to the wild inertia subgroup. By Corollary 4.37, the restriction of $\mathcal{R}_{\lambda,\boldsymbol{b}}$ to the wild inertia subgroup is a sum of terms of the form $[\times\alpha/\lambda]^*\mathcal{H}_{k-1}$ plus a trivial representation. Hence the restriction of $\mathcal{R}^*_{\lambda,\boldsymbol{b}} \otimes \mathcal{R}^{*\vee}_{\lambda,\boldsymbol{b}}$ to the wild inertia subgroup is a sum of representations of the form $[\times\alpha/\lambda]^*\mathcal{H}_{k-1} \otimes [\times\beta/\lambda]^*\mathcal{H}^\vee_{k-1}$, representations of the forms $[\times\alpha/\lambda]^*\mathcal{H}_{k-1}$ and $[\times\beta/\lambda]^*\mathcal{H}^\vee_{k-1}$, and a trivial representation.

Therefore, on the divisor $r = \infty$, it suffices to check that the Swan conductor of

$$[\times\alpha/\lambda]^*\mathcal{H}_{k-1} \otimes [\times\beta/\lambda]^*\mathcal{H}^\vee_{k-1} = [\times\alpha/\lambda]^*(\mathcal{H}_{k-1} \otimes [\times\beta/\alpha]^*\mathcal{H}^\vee_{k-1})$$

depends only on (α, β) but is independent of $\lambda \in \mathbf{G}_m$, and the same property for a single hypergeometric sheaf $[\times\alpha/\lambda]^*\mathcal{H}_{k-1}$. But scalar multiplication does not affect Swan conductors (since it is just an automorphism of the local field and hence preserves the wild ramification filtration), and hence these Swan conductors are equal to the Swan conductors of $\mathcal{H}_{k-1} \otimes [\times\beta/\alpha]^*\mathcal{H}^\vee_{k-1}$ and $\mathcal{H}_{k-1}$ respectively and thus are independent of λ. □

4.6. *Final steps.* In this final section, we compare different specialized sum-product sheaves $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$.

We now show distinctness of specialized sum-product sheaves for distinct λ. We recall that the subvariety $\mathcal{V}^{\mathrm{bad}}$ has been defined in Proposition 4.29. It is defined over $\mathbf{Z}[1/\ell]$ and stable under $\boldsymbol{b} \mapsto \tilde{\boldsymbol{b}} = (b_3, b_4, b_1, b_2)$.

LEMMA 4.42. *For $\boldsymbol{b}$ not contained in $\mathcal{V}^{\mathrm{bad}}(\mathbf{F}_q)$, and for $\lambda_1 \neq \lambda_2$ in $\mathbf{F}_q$, the sheaves $\mathcal{R}^*_{\lambda_1,\boldsymbol{b}}$ and $\mathcal{R}^*_{\lambda_2,\boldsymbol{b}}$ are not geometrically isomorphic.*

Proof. Let us recall first that, by definition, $\mathcal{V}^{\mathrm{bad}}$ contains $\mathcal{V}^{\Delta}$, and therefore the sheaves $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ are geometrically irreducible for $\boldsymbol{b} \notin \mathcal{V}^{\mathrm{bad}}$.

First assume that $\lambda_1 = 0$ and $\lambda_2 \neq 0$. (The case $\lambda_2 = 0$ and $\lambda_1 \neq 0$ is of course similar.) We will show that the generic ranks of the two sheaves $\mathcal{R}^*_{0,\boldsymbol{b}}$ and $\mathcal{R}^*_{\lambda_2,\boldsymbol{b}}$ are different, which of course implies that they are not geometrically isomorphic. By Lemma 4.31(2) and (3), we have

$$\operatorname{rank} \mathcal{R}_{\lambda_2,\boldsymbol{b}} = k^4 > k^3 = \operatorname{rank} \mathcal{R}_{0,\boldsymbol{b}}.$$

Applying Lemma 4.22(2) exactly as in the proof of Corollary 4.39, we see that the part of weight < 1 of $\mathcal{R}_{\lambda_2,\boldsymbol{b}}$ has rank

$$\dim \mathcal{K}^{I(0)}_{\lambda=\lambda_2,\eta,\boldsymbol{b}},$$

while (by the same argument) the part of weight < 1 of $\mathcal{R}_{0,\boldsymbol{b}}$ has rank

$$\dim \mathcal{K}^{I(0)}_{\lambda=0,\eta,\boldsymbol{b}} + \dim \mathcal{K}^{I(\infty)}_{\lambda=0,\eta,\boldsymbol{b}} \geqslant \dim \mathcal{K}^{I(0)}_{\lambda=0,\eta,\boldsymbol{b}}.$$

However the local monodromy representation of $\mathcal{K}_{\lambda,\eta,\boldsymbol{b}}$ at 0 is independent of λ because $\mathcal{K}$ is a tensor product of Kloosterman sheaves defined independently of λ with $\mathcal{L}_{\psi(\lambda s)}$, which is lisse at 0. So the rank of the inertia invariants is independent of λ also.

Hence there is a larger "drop" in the generic rank when passing from $\mathcal{R}_{\lambda,\boldsymbol{b}}$ to $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ when λ is 0, and we deduce that

$$\operatorname{rank}\mathcal{R}^*_{\lambda_2,\boldsymbol{b}} > \operatorname{rank}\mathcal{R}^*_{0,\boldsymbol{b}}.$$

Now assume that λ_1 and λ_2 are nonzero and distinct. Corollary 4.37 shows that the wild inertia representation of $\mathcal{R}^*_{\lambda_1,\boldsymbol{b}}$ at ∞ is the multiplicative translate by λ_2/λ_1 of the wild inertia representation of $\mathcal{R}^*_{\lambda_2,\boldsymbol{b}}$, which is itself isomorphic to the wild inertia representation of

$$\bigoplus_{\alpha\in S_k}[\times\alpha\lambda_2^{-1}]^*\mathcal{H}_{k-1}.$$

Since the wild inertia representation of $\mathcal{H}_{k-1}$ is not isomorphic to any nontrivial multiplicative translate of itself by Lemma 4.21(3), these local monodromy representations are therefore isomorphic only if $S_k = (\lambda_2/\lambda_1)S_k$. By Lemma 4.40(2), this is only possible if $\lambda_2 = \lambda_1$ or if $\lambda_2 = -\lambda_1$, and that second case occurs only if k is odd.

Thus it only remains to deal with the case when k is odd, $\lambda_1 = -\lambda_2$, and both are nonzero. We assume that we have a geometric isomorphism

$$\mathcal{R}^*_{\lambda_1,\boldsymbol{b}} \simeq \mathcal{R}^*_{-\lambda_1,\boldsymbol{b}} \tag{4.9}$$

for some $\lambda_1 \neq 0$, and we proceed to derive a contradiction. This isomorphism, and the fact that $\mathcal{R}^*_{\lambda_1,\boldsymbol{b}}$ and $\mathcal{R}^*_{-\lambda_1,\boldsymbol{b}}$ are geometrically irreducible, implies that $H^2_c(\mathbf{A}^1_{\overline{\mathbf{F}}_q} - \{-\boldsymbol{b}\}, \mathcal{R}^*_{\lambda_1,\boldsymbol{b}} \otimes \mathcal{R}^{*\vee}_{-\lambda_1,\boldsymbol{b}})$ is one-dimensional, where we use $\{-\boldsymbol{b}\}$ to denote the closed set $\{-b_1, -b_2, -b_3, -b_4\}$. This cohomology group is the stalk at λ_1 of the constructible ℓ-adic sheaf

$$\mathcal{G} = R^2p_!(\mathcal{R}^*_{\boldsymbol{b}} \otimes g^*\mathcal{R}^{*\vee}_{\boldsymbol{b}})(1),$$

where $p : (\mathbf{A}^1 - \{-\boldsymbol{b}\}) \times \mathbf{G}_m \to \mathbf{G}_m$ is the projection $(r,\lambda) \mapsto \lambda$ and g is the automorphism $(r,\lambda) \mapsto (r,-\lambda)$ of $(\mathbf{A}^1 - \{-\boldsymbol{b}\}) \times \mathbf{G}_m$.

By Deligne's semicontinuity theorem [Lau81], the sheaf $\mathcal{G}$ is lisse on $\mathbf{G}_m$; indeed, the Swan conductors are constant functions of λ on the ramification divisors, by an argument similar to that at the end of the proof of Proposition 4.41. Hence, since the stalk of $\mathcal{G}$ at $\lambda_1 \in \mathbf{G}_m$ is one-dimensional, the sheaf $\mathcal{G}$ is lisse of rank 1 on $\mathbf{G}_m$.

By Verdier duality (see, e.g., [KL85, §1, (1.1.3)] and the references there, and the fact that the dual of a (shifted) lisse sheaf is the shifted dual lisse sheaf), the dual of the sheaf $\mathcal{G}$ is isomorphic to

$$R^0p_*(\mathcal{R}^{*,\vee}_{\boldsymbol{b}} \otimes g^*\mathcal{R}^*_{\boldsymbol{b}}) \simeq p_*(\mathcal{H}\mathrm{om}(\mathcal{R}^*_{\boldsymbol{b}}, g^*\mathcal{R}^*_{\boldsymbol{b}})),$$

and the latter is therefore lisse on $\mathbf{G}_m$. We have a natural adjunction morphism

$$p^*p_*\mathcal{H}\mathrm{om}(\mathcal{R}^*_{\boldsymbol{b}}, g^*\mathcal{R}^*_{\boldsymbol{b}}) \to \mathcal{H}\mathrm{om}(\mathcal{R}^*_{\boldsymbol{b}}, g^*\mathcal{R}^*_{\boldsymbol{b}}).$$

We tensor with $\mathcal{R}^*_{\boldsymbol{b}}$ and compose with the canonical morphism

$$\mathcal{R}^*_{\boldsymbol{b}} \otimes \mathcal{H}\mathrm{om}(\mathcal{R}^*_{\boldsymbol{b}}, g^*\mathcal{R}^*_{\boldsymbol{b}}) \to g^*\mathcal{R}^*_{\boldsymbol{b}}$$

to deduce a morphism

$$\phi\colon \mathcal{R}^*_{\boldsymbol{b}} \otimes p^*\mathcal{G}^\vee \to g^*\mathcal{R}^*_{\boldsymbol{b}}.$$

The restriction to the geometric generic fiber $p^{-1}(\bar\eta)$ of $\mathbf{A}^1 - \{-\boldsymbol{b}\} \times \mathbf{G}_m$ of $p^*\mathcal{G}^\vee$ is

$$(p^*p_*\mathcal{H}\mathrm{om}(\mathcal{R}^*_{\boldsymbol{b}}, g^*\mathcal{R}^*_{\boldsymbol{b}}))|p^{-1}(\bar\eta) = p^*(p_*\mathcal{H}\mathrm{om}(\mathcal{R}^*_{\boldsymbol{b}}, g^*\mathcal{R}^*_{\boldsymbol{b}})|\bar\eta),$$

which is

$$(p_*\mathcal{H}\mathrm{om}(\mathcal{R}^*_{\boldsymbol{b}}, g^*\mathcal{R}^*_{\boldsymbol{b}}))_{\bar\eta} = \Gamma(p^{-1}(\bar\eta), \mathcal{H}\mathrm{om}(R^*_{\boldsymbol{b}}, g^*R^*_{\boldsymbol{b}})|p^{-1}(\bar\eta))$$

viewed as a constant sheaf.

The restriction to the geometric generic fiber of the previously described morphism is a natural homomorphism

$$\mathcal{R}^*_{\boldsymbol{b}}|p^{-1}(\bar\eta) \otimes \Gamma(p^{-1}(\bar\eta), \mathcal{H}\mathrm{om}(R^*_{\boldsymbol{b}}, g^*R^*_{\boldsymbol{b}})|p^{-1}(\bar\eta)) \to g^*\mathcal{R}^*_{\boldsymbol{b}}|p^{-1}(\bar\eta).$$

Specifically, we can describe this as the map that sends a section (say s) of $\mathcal{R}^*_{\boldsymbol{b}}$ over an open subset of $p^{-1}(\bar\eta)$ and a global section (say f) over $p^{-1}(\bar\eta)$ of the sheaf of homomorphisms from $\mathcal{R}^*_{\boldsymbol{b}}$ to $g^*\mathcal{R}^*_{\boldsymbol{b}}$ to the image $f(s)$.

This morphism is nontrivial as long as $\Gamma(p^{-1}(\bar\eta), \mathcal{H}\mathrm{om}(\mathcal{R}^*_{\boldsymbol{b}}, g^*\mathcal{R}^*_{\boldsymbol{b}})|p^{-1}(\bar\eta))$ is nonzero, as any nonzero section must correspond to a homomorphism that is nontrivial on some open set. The space of global sections is indeed nontrivial because we saw it is isomorphic to the stalk of $\mathcal{G}$ at $\bar\eta$, which is one-dimensional.

Hence $\phi_{\bar\eta}$ is nonzero on the geometric generic fiber. Because $\mathcal{R}^*_{\boldsymbol{b}}$ and $g^*\mathcal{R}^*_{\boldsymbol{b}}$ are geometrically irreducible lisse sheaves, and $p^*\mathcal{G}^\vee$ is one-dimensional, this implies that $\phi_{\bar\eta}$ is an isomorphism. Hence ϕ is a geometric isomorphism on any open dense set U on which $g^*\mathcal{R}^*_{\boldsymbol{b}}$, $p^*\mathcal{G}$, and $\mathcal{R}^*_{\boldsymbol{b}}$ are lisse.

We have seen that $\mathcal{G}$ is lisse on $\mathbf{G}_m$, and we know that $\mathcal{R}^*_{\boldsymbol{b}}$ is lisse on the complement of the divisors $\lambda = 0$ and $r = -b_i$, and the same holds for $g^*\mathcal{R}^*_{\boldsymbol{b}}$. So the homomorphism ϕ is a geometric isomorphism on the complement U of these divisors.

Our next goal is to prove that $\mathcal{G}$ is in fact geometrically trivial. For this, we now specialize the r variable. For r fixed but generic, we deduce from the above that $\mathcal{R}^*_{r,\boldsymbol{b}}$ is geometrically isomorphic to $(g^*\mathcal{R}^*)_{r,\boldsymbol{b}} \otimes \mathcal{G}$. However, $\mathcal{R}_{r,\boldsymbol{b}}$ is the restriction to $\mathbf{G}_m$ of the Fourier transform with respect to ψ of the sheaf

$$\mathcal{F} = \bigotimes_{1\leqslant i\leqslant 2} [s \mapsto (r+b_i)s]^*\mathcal{K}\ell_k \otimes [s \mapsto (r+b_{i+2})s]^*\mathcal{K}\ell_k^\vee$$

on $\mathbf{A}^1$ with variable s. The sheaf $\mathcal{F}$ is lisse on $\mathbf{G}_m$, with unipotent tame local monodromy at 0, and with all breaks $\leqslant 1/k$ at ∞. By Fourier transform theory it follows that $\mathcal{R}_{r,\boldsymbol{b}}$ is lisse on $\mathbf{G}_m$ (see [Kat90, Lemma 7.3.9(3)]), with unipotent tame local monodromy at ∞ ([Kat90, Th. 7.4.1(1), Th. 7.4.4(3)]) and with all breaks $\leqslant 1/(k-1)$ at 0. (See [Kat90, Th. 7.5.4(5)]; note the integers c, d in the assumption of that reference are not necessarily coprime.)

Pulling-back by g, we see that the sheaf $g^*\mathcal{R}_{r,\boldsymbol{b}}$ has the same ramification properties, and hence also $g^*\mathcal{R}^*_{r,\boldsymbol{b}}$. From this and the isomorphism $\mathcal{R}^*_{r,\boldsymbol{b}} \simeq (g^*\mathcal{R}^*)_{r,\boldsymbol{b}} \otimes \mathcal{G}$, it follows that $\mathcal{G}$ must also be lisse on $\mathbf{G}_m$, tame with unipotent monodromy at ∞, and with (unique) break $\leqslant 1/(k-1)$ at 0. But since a rank 1 sheaf has an integral break, this means that $\mathcal{G}$ is also tame at 0, and since unipotent monodromy in rank 1 is trivial, this means that $\mathcal{G}$ is lisse at ∞. However, a sheaf on $\mathbf{P}^1$ that is lisse on $\mathbf{P}^1 - \{0\}$ and tamely ramified at 0 is geometrically trivial, so $\mathcal{G}$ is geometrically trivial.

We have therefore proved that $\mathcal{R}^*_{\boldsymbol{b}}$ and $g^*\mathcal{R}^*_{\boldsymbol{b}}$ are geometrically isomorphic. But this is impossible, since this would imply that

$$\limsup_{d\to+\infty}\left|\frac{1}{q^d}\frac{1}{q^{2d}}\sum_{\lambda,r\in\mathbf{F}_{q^d}}\sum \mathbf{R}(r,\lambda,\boldsymbol{b},\mathbf{F}_{q^d})\overline{\mathbf{R}(r,-\lambda,\boldsymbol{b};\mathbf{F}_{q^d})}\right| > 0$$

(since $\mathbf{R}(r,\lambda,\boldsymbol{b};\mathbf{F}_{q^d}) = t_{\mathcal{R}^*}(r,\lambda,\boldsymbol{b};\mathbf{F}_{q^d}) + O(1)$ where $\mathcal{R}^*_{\boldsymbol{b}}$ is lisse) and this contradicts the estimate (3.6) for odd-rank Kloosterman sheaves. □

We can now finally recapitulate and prove Theorem 4.10.

Proof of Theorem 4.10. Let $\mathcal{V}^{\mathrm{bad}}$ be the subvariety in Proposition 4.29. It is defined over $\mathbf{Z}[1/\ell]$, and hence its degree is bounded independently of q. It is also stable under $\boldsymbol{b} \mapsto \tilde{\boldsymbol{b}}$ by construction. For q large enough, let $\boldsymbol{b} \notin \mathcal{V}^{\mathrm{bad}}(\mathbf{F}_q)$. Then $\mathcal{R}^*_{0,\boldsymbol{b}}$ is geometrically irreducible (by Proposition 4.29), and $\mathcal{R}^*_{\lambda,\boldsymbol{b}}$ is geometrically irreducible for all $\lambda \neq 0$ if q is large enough by Proposition 4.41 since $\mathcal{V}^{\mathrm{bad}}$ is defined to contain $\mathcal{V}^{\Delta}$.

The second part of Theorem 4.10 is given by Lemma 4.42, and the third by Proposition 4.24. □

5. Functions of triple divisor type in arithmetic progressions to large moduli

In this section, we prove Theorem 1.7. Let f be a holomorphic primitive cusp form of level 1 and weight k. We denote by $\lambda_f(n)$ the Hecke eigenvalues, which are normalized so that we have $|\lambda_f(n)| \leqslant d_2(n)$. The method will be very similar to that used in [FKM15b], and some technical details will be handled rather quickly as they follow very closely the corresponding steps for the triple divisor function.

For any prime q and integer a coprime to q, we denote

$$E(\lambda_f \star 1, x; q, a) := \sum_{\substack{n\leqslant x\\ n\equiv a \,(\mathrm{mod}\, q)}} (\lambda_f \star 1)(n) - \frac{1}{\varphi(q)} \sum_{\substack{n\leqslant x\\ (n,q)=1}} (\lambda_f \star 1)(n).$$

5.1. *Preliminaries.* We first recall several useful results. We begin by stating the estimates for linear and bilinear forms involving the hyper-Kloosterman sums $\mathrm{Kl}_3(a;q)$.

PROPOSITION 5.1. *Let q a prime number, $M, N \in [1,q]$, $\mathcal{N}$ an interval of length N, and $(\alpha_m)_m$, $(\beta_n)_n$ two sequences supported respectively on $[1,M]$ and $\mathcal{N}$. Let a be an integer coprime to q.*

Let V and W be smooth functions compactly supported in the interval $[1,2]$ and satisfying

$$V^{(j)}(x), W^{(j)}(x) \ll_j Q^j \tag{5.1}$$

for some $Q \geqslant 1$ and for all $j \geqslant 0$.

Let $\varepsilon > 0$ be given.

(1) *There exists an absolute constant $C_1 \geqslant 0$ such that we have*

$$\sum_{m,n\geqslant 1}\sum \lambda_f(m) V\Big(\frac{m}{M}\Big) W\Big(\frac{n}{N}\Big) \mathrm{Kl}_3(amn;q) \ll q^{\varepsilon} Q^{C_1} MN\Big(\frac{1}{q} + \frac{q^{1/2}}{N}\Big) \tag{5.2}$$

and

$$\sum_{m} \lambda_f(m) V\Big(\frac{m}{M}\Big) \mathrm{Kl}_3(am;q) \ll q^{\varepsilon} Q^{C_1} M\Big(\frac{1}{q^{1/8}} + \frac{q^{3/8}}{M^{1/2}}\Big). \tag{5.3}$$

(2) *We have*

$$\sum_{m\leqslant M,\, n\in\mathcal{N}}\sum \alpha_m \beta_n \mathrm{Kl}_3(amn;q) \ll q^{\varepsilon} \|\alpha\|_2 \|\beta\|_2 (MN)^{1/2}\Big(\frac{1}{M^{1/2}} + \frac{q^{1/4}}{N^{1/2}}\Big). \tag{5.4}$$

(3) *If*

$$1 \leqslant M \leqslant N^2, \quad N < q, \quad MN \leqslant q^{3/2},$$

we have

$$\sum_{m,n\geqslant 1}\sum \lambda_f(m) V\Big(\frac{m}{M}\Big) W\Big(\frac{n}{N}\Big) \mathrm{Kl}_3(amn;q) \ll q^{\varepsilon} Q^{C_1} MN\Big(\frac{q^{1/4}}{M^{1/6}N^{5/12}}\Big). \tag{5.5}$$

In all estimates, the implied constant depends only on ε.

Proof. The bound (5.2) is an instance of the completion method and follows from an application of the Poisson summation formula to the sum over n, using the fact that

$$\widehat{\mathrm{Kl}_3}(u) \ll 1, \quad \widehat{\mathrm{Kl}_3}(0) = -\frac{1}{p^{3/2}}$$

(the former because $\mathrm{Kl}_3(\cdot;q)$ is the trace function of a Fourier sheaf modulo q, and the latter by direct computation).

The bounds (5.3) and (5.4) are special cases of [FKM15a, Th. 1.2] and [FKM14, Th. 1.17]. The bound (5.5) is a consequence of Theorem 1.3 (for $c=1$) after summation by parts. □

PROPOSITION 5.2. *Let q be a prime number, and let V, W be two smooth functions compactly supported on $]0,+\infty[$. Let $K : \mathbf{Z} \longrightarrow \mathbf{C}$ be any q-periodic arithmetic function.*

We have

$$\begin{aligned}\sum_{m,n\geqslant 1}\sum K(mn)\lambda_f(m)V(m)W(n)&\\ = \frac{\widehat{K}(0)}{q^{1/2}}\sum_{m,n\geqslant 1}\sum \lambda_f(m)V(m)W(n)&\\ + \Big(K(0)-\frac{\widehat{K}(0)}{q^{1/2}}\Big)\sum_{m,n\geqslant 1}\sum \lambda_f(m)V(m)W(qn)&\\ + \frac{1}{q^{3/2}}\sum_{m,n\geqslant 1}\tilde{K}(mn)\lambda_f(m)\check{V}\Big(\frac{m}{q^2}\Big)\hat{W}\Big(\frac{n}{q}\Big),&\end{aligned}$$

where $\hat{W}$ denotes the Fourier transform of W, $\check{V}$ is the weight k Bessel transform given by

$$\check{V}(x) = 2\pi i^k \int_0^\infty V(t)J_{k-1}(4\pi\sqrt{xt})dt, \tag{5.6}$$

and

$$\widetilde{K}(m) = \frac{1}{q^{1/2}}\sum_{(u,q)=1} K(u)\,\mathrm{Kl}_3(mu;q).$$

In particular, if a is an integer coprime with q and $K(n) = \delta_{n=a \,(\mathrm{mod}\, q)}$, then we have

$$K(0)=0, \quad \hat{K}(0) = \frac{1}{q^{1/2}}, \quad \widetilde{K}(m) = \frac{1}{q^{1/2}}\,\mathrm{Kl}_3(am;q)$$

by direct computations.

Proof. We split the sum into

$$\sum_{q|n}\Big(\sum_m \cdots\Big) + \sum_{(n,q)=1}\Big(\sum_m \cdots\Big). \tag{5.7}$$

The contribution of those n divisible by q is

$$K(0)\sum_{m,n\geqslant 1}\lambda_f(m)V(m)W(qn).$$

For those n coprime to q, we apply the Fourier inversion formula

$$K(mn) = \frac{\widehat{K}(0)}{q^{1/2}} + \frac{1}{q^{1/2}} \sum_{\substack{u \,(\mathrm{mod}\, q)\\(u,q)=1}} \widehat{K}(u) e\Big(-\frac{umn}{q}\Big).$$

The contribution of the first term is

$$\frac{\widehat{K}(0)}{q^{1/2}} \sum_{\substack{m,n\geqslant 1\\(n,q)=1}} \lambda_f(m)V(m)W(n) = \frac{\widehat{K}(0)}{q^{1/2}}\Big(\sum_{m,n\geqslant 1} \lambda_f(m)V(m)W(n) - \sum_{m,n\geqslant 1} \lambda_f(m)V(m)W(qn)\Big).$$

For the last term, we apply the Voronoi summation formula to the sum over m: we have

$$\sum_{m\geqslant 1} \lambda_f(m)V(m)e\Big(-\frac{mnu}{q}\Big) = \frac{1}{q}\sum_{m\geqslant 1} \lambda_f(m)\check{V}\Big(\frac{m}{q^2}\Big)e\Big(\frac{\overline{nu}m}{q}\Big)$$

for each u (see, e.g., [FGKM14, Lemma 2.2]). Therefore, the total contribution of the second term in (5.7) equals

$$\frac{1}{q}\sum\sum_{\substack{m,n\geqslant 1\\(n,q)=1}} \lambda_f(m)\check{V}\Big(\frac{m}{q^2}\Big)W(n)\tilde{K}(m,n)$$

with

$$\tilde{K}(m,n) = \frac{1}{q^{1/2}} \sum_{(u,q)=1} \widehat{K}(u) e\Big(\frac{m\overline{nu}}{q}\Big).$$

We finish by applying the Poisson summation formula to the sum over n: we have

$$\begin{aligned}\sum_{(n,q)=1} W(n)\tilde{K}(m,n) &= \frac{1}{q}\sum_{n} \hat{W}\Big(\frac{n}{q}\Big) \sum_{(v,q)=1} e\Big(\frac{m\overline{uv}+nv}{q}\Big)\\ &= \frac{1}{q^{1/2}}\sum_{n} \hat{W}\Big(\frac{n}{q}\Big)\,\mathrm{Kl}_2(mn\overline{u};q)\end{aligned}$$

for each m, so that the total contribution becomes

$$\frac{1}{q^{3/2}}\sum\sum_{\substack{m,n\\(n,q)=1}} \tilde{K}(mn)\lambda_f(m)\check{V}\Big(\frac{m}{q^2}\Big)\hat{W}\Big(\frac{n}{q}\Big),$$

where

$$\tilde{K}(m) = \frac{1}{q^{1/2}} \sum_{(u,q)=1} \hat{K}(u)\,\mathrm{Kl}_2(m\overline{u};q) = \frac{1}{q^{1/2}} \sum_{(u,q)=1} K(u)\,\mathrm{Kl}_3(mu;q)$$

for any m. This gives the formula we stated. □

5.2. *Decomposition of $E(\lambda_f \star 1, x; q, a)$.* Given any $A \geqslant 1$ as in Theorem 1.7, we fix some $B \geqslant 1$ sufficiently large (to depend on A). Given $x \geqslant 2$, we set

$$\mathcal{L} := \log x, \quad \Delta = 1 + \mathcal{L}^{-B}.$$

Arguing as in [FKM15b], we perform a partition of unity on the m and n variables and decompose $E(\lambda_f \star 1, x; q, a)$ into $O(\log^2 x)$ terms of the form

$$\tilde{E}(V, W; q, a) = \sum_{mn = a \,(\mathrm{mod}\, q)} \lambda_f(m) V(m) W(n) - \frac{1}{q} \sum_{(mn,q)=1} \lambda_f(m) V(m) W(n),$$

where V, W are smooth functions satisfying

$$\operatorname{supp} V \subset [M, \Delta M], \ \operatorname{supp} W \subset [N, \Delta N],$$
$$x^j V^{(j)}(x), \ x^j W^{(j)}(x) \ll_j \mathcal{L}^{Bj},$$

and where

$$x\mathcal{L}^{-C} \leqslant MN \leqslant x$$

for some $C \geqslant 0$ large enough, depending on the value of the parameter A in Theorem 1.7.

Applying Proposition 5.2 to the first term, we obtain

$$\tilde{E}(V, W; q, a) = \frac{1}{q^{1/2}} \sum_{m,n \geqslant 1} \mathrm{Kl}_3(mn; q) \lambda_f(m) \check{V}\Big(\frac{m}{q^2}\Big) \hat{W}\Big(\frac{n}{q}\Big),$$

and hence it only remains to prove that

$$\frac{1}{q^{1/2}} \sum_{m,n \geqslant 1} \mathrm{Kl}_3(mn; q) \lambda_f(m) \check{V}\Big(\frac{m}{q^2}\Big) \hat{W}\Big(\frac{n}{q}\Big) \ll_A \frac{MN}{q} \mathcal{L}^{-A}. \tag{5.8}$$

The following standard lemma describes the decay of the Fourier and Bessel transforms of V and W:

Lemma 5.3. *Let V, W be as above, and let $\check{W}$, $\hat{W}$ be their Bessel and Fourier transforms as defined in* (5.6). *There exists a constant $D \geqslant 0$ such that for any $x > 0$, any $E \geqslant 0$ and any $j \geqslant 0$, we have*

$$x^j \check{V}^{(j)}(x) \ll_{E,f,j} M\mathcal{L}^{Bj} \Big(\frac{\mathcal{L}^{Dj}}{1 + xM}\Big)^E, \tag{5.9}$$

$$x^j \hat{W}^{(j)}(x) \ll_{E,j} N\mathcal{L}^{Bj} \Big(\frac{\mathcal{L}^{Dj}}{1 + xN}\Big)^E. \tag{5.10}$$

Proof. By the change of variable $u = 4\pi\sqrt{xt}$, we find that

$$\begin{aligned} \check{V}(x) &= \frac{i^k}{8\pi} \int_0^\infty \frac{u^2}{x} V\Big(\frac{u^2}{16\pi^2 x}\Big) J_{k-1}(u) \frac{du}{u} \\ &= \frac{i^k}{8\pi\sqrt{x}} \int_0^\infty \Big(\frac{u^2}{x}\Big)^{1/2} V\Big(\frac{u^2}{16\pi^2 x}\Big) J_{k-1}(u) du. \end{aligned}$$

Since $J_{k-1}(u) \ll_f (1+u)^{-1/2}$, we have

$$\check{V}(x) \ll \frac{M}{(1+xM)^{1/2}}.$$

On the other hand, applying [KMV02, Lem. 6.1] we obtain the bound

$$\check{V}(x) \ll_{f,j} \frac{M^{1/2}}{x^{1/2}} \frac{(1+|\log xM|)\mathcal{L}^{O(j)}}{(xM)^{\frac{j-1}{2}}}(xM)^{1/4}.$$

In particular, if $xM \geqslant 1$, then by taking j large enough, we see that $\check{V}(x) \ll_{f,E} M\mathcal{L}^{O(E)}(xM)^{-E}$, which concludes the proof of (5.9) when $j = 0$. The general case is similar, and the proof of (5.10) follows similar lines (using easier standard properties of the Fourier transform). □

Set

$$M^* = q^2/M \text{ and } N^* = q/N.$$

Then this lemma shows that, if $\eta > 0$ is arbitrarily small, the contribution to the sum (5.8) of the (m,n) such that

$$m \geqslant x^{\eta/2}M^* \text{ or } n \geqslant x^{\eta/2}N^*$$

is negligible. Therefore, by (5.9) and (5.10), and a smooth dyadic partition of unity, we are reduced to estimating sums of the type

$$S(M',N') = \sum_{m,n\geqslant 1}\sum \lambda_f(m)\mathrm{Kl}_3(amn;q)V^*(m)W^*(n),$$

where

$$1/2 \leqslant M' \leqslant M^*x^{\eta/2}, \quad 1/2 \leqslant N' \leqslant N^*x^{\eta/2},$$

and V^*, W^* are smooth compactly supported functions with

$$\mathrm{supp}(V^*) \subset [M', 2M'], \quad \mathrm{supp}(W^*) \subset [N', 2N']$$
$$u^jV^*(j)(u), \quad u^jW^*(j)(u) \ll \mathcal{L}^{O(j)}$$

for any $j \geqslant 0$. Precisely, it is enough to prove that

$$S(M',N') \ll_A q\mathcal{L}^{-A}.$$

Since the trivial bound for $S(M',N')$ is

$$S(M',N') \ll M'N'\mathcal{L},$$

we may assume that

$$q\mathcal{L}^{-A-1} \leqslant M'N' \leqslant q^3x^{-1+\eta}.$$

Let us write

$$x = q^{2-\delta}, \quad M = q^\mu, \quad N = q^\nu, \quad M' = q^{\mu'}, \quad N' = q^{\nu'}$$

so that

$$M^* = q^{\mu^*}, \quad N^* = q^{\nu^*}$$

with

$$\mu^* = 2 - \mu, \quad \nu^* = 1 - \nu, \quad \mu' \leqslant \mu^* + \eta/2, \quad \nu' \leqslant \nu^* + \eta/2$$

and

$$\mu + \nu = 2 - \delta + o(1).$$

Let us write

$$S(M', N') = q^{\sigma(\mu', \nu')}.$$

Then Proposition 5.1 translates to the estimates

$$\sigma(\mu', \nu') \leqslant \tau(\mu', \nu') + o(1),$$

where

$$\tau(\mu', \nu') \leqslant \mu' + \nu' + \max(-1, 1/2 - \nu') \quad \text{(by (5.2))}, \tag{5.11}$$

$$\tau(\mu', \nu') \leqslant \mu' + \nu' + \max(-1/8, 3/8 - \mu'/2) \quad \text{(by (5.3))}, \tag{5.12}$$

$$\tau(\mu', \nu') \leqslant \mu' + \nu' + \max(-\mu'/2, 1/4 - \nu'/2) \quad \text{(by (5.4))}, \tag{5.13}$$

$$\tau(\mu', \nu') \leqslant \mu' + \nu' + \max(-\nu'/2, 1/4 - \mu'/2) \quad \text{(by (5.4) with } M, N \text{ interchanged)}, \tag{5.14}$$

$$\tau(\mu', \nu') \leqslant \mu' + \nu' + 1/4 - \mu'/6 - 5\nu'/12 \quad \text{(by (5.5), if } 0 \leqslant \mu' \leqslant 2\nu'). \tag{5.15}$$

(Indeed, note that the conditions $\nu' \leqslant 1$ and $\mu' + \nu' \leqslant 3/2$ also required in (5.5) are always satisfied for η small enough, since $\mu' + \nu' \leqslant 3 + (2 - \delta)(-1 + \eta) < \frac{3}{2}$ and $\nu' = 1 - \nu \leqslant 1$.)

We will prove that if $\delta < \frac{1}{26}$ and η is small enough, then we have $\sigma(\mu', \nu') \leqslant 1 - \kappa$, where $\kappa > 0$ depends only on δ and η. This implies the desired estimate. In the argument, we denote by $o(1)$ quantities tending to 0 as η tends to 0 or q tends to infinity.

First, since

$$\mu' + \nu' - 1 \leqslant \mu' + \nu' - \frac{1}{8} \leqslant 1 + \delta - \frac{1}{8} + o(1) < 1,$$

we may replace (5.11) and (5.12) by

$$\tau(\mu', \nu') \leqslant \mu' + \frac{1}{2}, \tag{5.16}$$

$$\tau(\mu', \nu') \leqslant \frac{\mu' + \nu'}{2} + \frac{\nu'}{2} + \frac{3}{8}. \tag{5.17}$$

We now distinguish various cases:

– if $\mu' \leqslant \frac{1}{2} - \kappa$, then we obtain the bound by (5.16);

– if

$$\nu' > 2(\delta + \kappa) \text{ and } \mu' > \frac{1}{2} + (2\delta + \kappa),$$

then we obtain the bound by (5.14);

– if $\nu' \leqslant 2(\delta+\kappa)$, we obtain a suitable bound, provided κ is small enough, by (5.17) since then

$$\frac{\mu'+\nu'}{2}+\frac{\nu'}{2}+\frac{3}{8} \leqslant \frac{1+\delta+o(1)}{2}+\delta+\kappa+\frac{3}{8} \leqslant \frac{7}{8}+\frac{3\delta}{2}+\kappa+o(1);$$

– finally, if $\mu' \leqslant (2\delta+\kappa)+\frac{1}{2}$, then from $\mu'+\nu' \geqslant 1$, we deduce that

$$2\nu' \geqslant 1-4\delta-2\kappa \geqslant \frac{1}{2}+2\delta-4\kappa \geqslant \mu'$$

provided κ is small enough, and so (5.15) is applicable and gives the desired bound since

$$\begin{aligned}\mu'+\nu'+\frac{1}{4}-\frac{\mu'}{6}-\frac{5\nu'}{12} &= \frac{7}{12}(\mu'+\nu')+\frac{1}{4}+\frac{\mu'}{4}+o(1)\\ &\leqslant \frac{7}{12}(1+\delta)+\frac{1}{4}+\frac{1}{4}\Big(\frac{1}{2}+2\delta+\kappa\Big)+o(1)\\ &= 1-\frac{13}{12}(1/26-\delta)+\frac{\kappa}{4}+o(1).\end{aligned}$$

Appendix A. Nearby and vanishing cycles

Let R be a Henselian discrete valuation ring R with fraction field K. Let S be the spectrum of R, and denote its generic point by η and its special point by s. Let $\bar{\eta}$ be a geometric point over η and $\bar{s}$ a geometric point over s.

For any proper scheme $f : X \longrightarrow S$, and any prime ℓ invertible on S, the nearby cycles function $R\Psi$ is a functor from ℓ-adic sheaves on X_η to the derived category of ℓ-adic sheaves on $X_{\bar{s}}$ equipped with an action of the absolute Galois group G of K (see, e.g., [DK73, Exp. XIII] for the definition and further references):

$$\begin{array}{ccccc} X_s & \xrightarrow{i} & X & \xleftarrow{j} & X_{\overline{\eta}} \\ \downarrow & & \downarrow{\scriptstyle f} & & \downarrow \\ s & \xrightarrow{i_0} & S & \xleftarrow{j_0} & \overline{\eta}. \end{array} \tag{A.1}$$

Given $\mathcal{F}$ a sheaf on X and $\mathcal{F}_s := i^*\mathcal{F}$ and $\mathcal{F}_{\overline{\eta}} := j^*\mathcal{F}$, the complex $R\Psi\mathcal{F}$ is defined as

$$R\Psi\mathcal{F} = i^* R j_* \mathcal{F}_{\overline{\eta}}.$$

The mapping cone of the adjunction map $i^*\mathcal{F} \to R\Psi\mathcal{F}$ is noted $R\Phi\mathcal{F}$ and is called the complex of vanishing cycles; one then has a cohomology exact sequence arising from the corresponding distinguished triangle

$$\cdots \to H^i(X_{\bar{s}}, \mathcal{F}_s) \to H^i(X_{\bar{s}}, R\Psi\mathcal{F}) \to H^i(X_{\bar{s}}, R\Phi\mathcal{F}) \to \cdots . \tag{A.2}$$

The functor $R\Psi$ has several key properties that we use in this paper:

(1) (See [Lau81, (1.3.3.1)], [DK73, (2.1.8.3)]) For any $i \geqslant 0$, there is a natural isomorphism of G-representations

$$H^i(X_{\overline{\eta}}, \mathcal{F}) = H^i(X_{\overline{s}}, R\Psi\mathcal{F}). \tag{A.3}$$

Since the left-hand side of (A.3) is, together with its Galois action, the local monodromy representation of the higher-direct image sheaf $R^i f_*\mathcal{F}$ at s, the nearby cycle complex will enable us to compute the local monodromy representation at specific points of some global sheaves obtained by pushforward on curves.

(2) (See [Lau81, Th 1.3.1.3], [Del77, Th. Finitude, Prop. 3.7]) The functor $R\Psi$ is defined étale-locally: if two pairs $(X \longrightarrow S, \mathcal{F})$ and $(X' \longrightarrow S, \mathcal{F}')$ are given that are isomorphic in an étale neighborhood of a point $x \in X$, i.e., if there exist a scheme U over S, a point $\tilde{x} \in U$ and étale morphisms making the diagram

$$\begin{array}{ccc} U & \xrightarrow{g'} & X' \\ g \downarrow & & \downarrow f' \\ X & \xrightarrow{f} & S \end{array}$$

commute with $g(\tilde{x}) = x$, $g'(\tilde{x}) = x'$ (say), and if $g^*\mathcal{F} \simeq (g')^*\mathcal{F}'$, then we have

$$g^* R\Psi\mathcal{F} \simeq (g')^* R\Psi\mathcal{G};$$

i.e., the nearby cycles complexes are isomorphic in the same étale neighborhood.

This will be useful to compare the local monodromy of a given sheaf on a given curve to possibly simpler ones on other (also possibly simpler) curves, which are étale-locally isomorphic and take advantage of some existing computations of nearby cycles: for instance, the local acyclicity of smooth morphisms (which handles the case of a lisse sheaf on a smooth scheme) and Laumon's local Fourier transform, which describes the nearby cycles that arise when computing the Fourier transform of a sheaf (in other words, the stationary phase formula).

References

[Blo13] V. Blomer, Applications of the Kuznetsov formula on GL(3), *Invent. Math.* **194** no. 3 (2013), 673–729. MR 3127065. Zbl 1292.11064. https://doi.org/10.1007/s00222-013-0454-3.

[BFK+17] V. Blomer, É. Fouvry, E. Kowalski, P. Michel, and D. Milićević, On moments of twisted L-functions, *Amer. J. Math.* **139** no. 3 (2017), 707–768. https://doi.org/10.1353/ajm.2017/0019.

[BFK+] V. Blomer, É. Fouvry, E. Kowalski, P. Michel, D. Milićević, and W. Sawin, On the non-vanishing of twisted L-functions, in preparation.

[BM15] V. Blomer and D. Milićević, The second moment of twisted modular L-functions, *Geom. Funct. Anal.* **25** no. 2 (2015), 453–516. MR 3334233. Zbl 06438343. https://doi.org/10.1007/s00039-015-0318-7.

[BFG88] D. Bump, S. Friedberg, and D. Goldfeld, Poincaré series and Kloosterman sums for $\mathrm{SL}(3,\mathbf{Z})$, *Acta Arith.* **50** no. 1 (1988), 31–89. MR 0945275. Zbl 0647.10020.

[Del77] P. Deligne, *Cohomologie Étale*, *Lecture Notes in Math.* **569**, Springer-Verlag, New York, 1977, Séminaire de Géométrie Algébrique du Bois-Marie SGA $4\frac{1}{2}$; avec la collaboration de J. F. Boutot, A. Grothendieck, L. Illusie et J. L. Verdier. MR 0463174. Zbl 0345.00010. https://doi.org/10.1007/BFb0091516.

[Del80] P. Deligne, La conjecture de Weil. II, *Inst. Hautes Études Sci. Publ. Math.* no. 52 (1980), 137–252. MR 0601520. Zbl 0456.14014. https://doi.org/10.1007/BF02684780.

[DK73] P. Deligne and N. Katz, *Groupes de Monodromie en Géométrie Algébrique. II*, *Lecture Notes in Math.* **340**, Springer-Verlag, New York, 1973, Séminaire de Géométrie Algébrique du Bois-Marie 1967–1969 (SGA 7 II). MR 0354657. https://doi.org/10.1007/BFb0060505.

[DI83] J.-M. Deshouillers and H. Iwaniec, Kloosterman sums and Fourier coefficients of cusp forms, *Invent. Math.* **70** no. 2 (1982/83), 219–288. MR 0684172. Zbl 0502.10021. https://doi.org/10.1007/BF01390728.

[FGKM14] E. Fouvry, S. Ganguly, E. Kowalski, and P. Michel, Gaussian distribution for the divisor function and Hecke eigenvalues in arithmetic progressions, *Comment. Math. Helv.* **89** no. 4 (2014), 979–1014. MR 3284303. Zbl 1306.11079. https://doi.org/10.4171/CMH/342.

[FKM14] E. Fouvry, E. Kowalski, and P. Michel, Algebraic trace functions over the primes, *Duke Math. J.* **163** no. 9 (2014), 1683–1736. MR 3217765. Zbl 1318.11103. https://doi.org/10.1215/00127094-2690587.

[FKM15a] E. Fouvry, E. Kowalski, and P. Michel, Algebraic twists of modular forms and Hecke orbits, *Geom. Funct. Anal.* **25** no. 2 (2015), 580–657. MR 3334236. Zbl 1344.11036. https://doi.org/10.1007/s00039-015-0310-2.

[FKM15b] E. Fouvry, E. Kowalski, and P. Michel, On the exponent of distribution of the ternary divisor function, *Mathematika* **61** no. 1 (2015), 121–144. MR 3333965. Zbl 1317.11080. https://doi.org/10.1112/S0025579314000096.

[FKM15c] E. Fouvry, E. Kowalski, and P. Michel, A study in sums of products, *Philos. Trans. Roy. Soc. A* **373** no. 2040 (2015), 20140309, 26. MR 3338119. Zbl 06674711. https://doi.org/10.1098/rsta.2014.0309.

[FM98] E. Fouvry and P. Michel, Sur certaines sommes d'exponentielles sur les nombres premiers, *Ann. Sci. École Norm. Sup.* **31** no. 1 (1998), 93–130. MR 1604298. Zbl 0915.11045. https://doi.org/10.1016/S0012-9593(98)80019-0.

[FI85] J. B. Friedlander and H. Iwaniec, Incomplete Kloosterman sums and a divisor problem, *Ann. of Math.* **121** no. 2 (1985), 319–350, with an appendix by B. J. Birch and E. Bombieri. MR 0786351. Zbl 0572.10029. https://doi.org/10.2307/1971175.

[Fu10] L. Fu, Calculation of ℓ-adic local Fourier transformations, *Manuscripta Math.* **133** no. 3-4 (2010), 409–464. MR 2729262. Zbl 1206.14035. https://doi.org/10.1007/s00229-010-0377-x.

[Fu11] L. Fu, *Etale Cohomology Theory*, *Nankai Tracts in Math.* **13**, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2011. MR 2791606. Zbl 1228.14001. https://doi.org/10.1142/9789814307734.

[Fu16] L. Fu, ℓ-adic GKZ hypergeometric sheaves and exponential sums, *Adv. Math.* **298** (2016), 51–88. MR 3505737. Zbl 06591692. https://doi.org/10.1016/j.aim.2016.04.021.

[GKR09] P. Gao, R. Khan, and G. Ricotta, The second moment of Dirichlet twists of Hecke L-functions, *Acta Arith.* **140** no. 1 (2009), 57–65. MR 2557853. Zbl 1242.11035. https://doi.org/10.4064/aa140-1-4.

[GR71] A. Grothendieck and M. Raynaud, *Revêtements Étales et Groupe Fondamental*, *Lecture Notes in Math.* **224**, Springer-Verlag, New York, 1971, Séminaire de Géométrie Algébrique du Bois Marie 1960–1961 (SGA 1); dirigé par Alexandre Grothendieck.; augmenté de deux exposés de M. Raynaud. MR 0354651. https://doi.org/10.1007/BFb0058656.

[Har77] R. Hartshorne, *Algebraic Geometry*, *Graduate Texts in Math.* **52**, Springer-Verlag, New York, 1977. MR 0463157. Zbl 0367.14001. https://doi.org/10.1007/978-1-4757-3849-0.

[HB81] D. R. Heath-Brown, The fourth power mean of Dirichlet's L-functions, *Analysis* **1** no. 1 (1981), 25–32. MR 0623640. Zbl 0479.10027. https://doi.org/10.1524/anly.1981.1.1.25.

[HB86] D. R. Heath-Brown, The divisor function $d_3(n)$ in arithmetic progressions, *Acta Arith.* **47** no. 1 (1986), 29–56. MR 0866901. Zbl 0549.10034.

[HL] J. Hoffstein and M. Lee, Second moments and simultaneous non-vanishing of GL(2) automorphic L-series. arXiv 1308.5980.

[IK04] H. Iwaniec and E. Kowalski, *Analytic Number Theory*, *Amer. Math. Soc. Colloq. Publ.* **53**, Amer. Math. Soc., Providence, RI, 2004. MR 2061214. Zbl 1059.11001. https://doi.org/10.1090/coll/053.

[Kat80] N. M. Katz, *Sommes Sxponentielles*, *Astérisque* **79**, Soc. Math. France, Paris, 1980. MR 0617009. Zbl 0469.12007.

[Kat88] N. M. Katz, *Gauss sums*, *Kloosterman Sums*, *and Monodromy Groups*, *Ann. of Math. Stud.* **116**, Princeton Univ. Press, Princeton, NJ, 1988. MR 0955052. Zbl 0675.14004. https://doi.org/10.1515/9781400882120.

[Kat90] N. M. Katz, *Exponential Sums and Differential Equations*, *Ann. of Math. Stud.* **124**, Princeton Univ. Press, Princeton, NJ, 1990. MR 1081536. Zbl 0731.14008. https://doi.org/10.1515/9781400882434.

[Kat96] N. M. Katz, *Rigid Local Systems*, *Ann. of Math. Stud.* **139**, Princeton Univ. Press, Princeton, NJ, 1996. MR 1366651. Zbl 0864.14013. https://doi.org/10.1515/9781400882595.

[Kat01] N. M. Katz, Sums of Betti numbers in arbitrary characteristic, *Finite Fields Appl.* **7** no. 1 (2001), 29–44. MR 1803934. Zbl 1068.14501. https://doi.org/10.1006/ffta.2000.0303.

[Kat12] N. M. Katz, *Convolution and Equidistribution*: *Sato-Tate Theorems for Finite-Field Mellin Transforms*, *Ann. Math. Stud.* **180**, Princeton Univ. Press, Princeton, NJ, 2012. MR 2850079. Zbl 1261.11084.

[KL85] N. M. Katz and G. Laumon, Transformation de Fourier et majoration de sommes exponentielles, *Inst. Hautes Études Sci. Publ. Math.* no. 62 (1985), 361–418. MR 0823177. Zbl 0603.14015. Available at http://www.numdam.org/item?id=PMIHES_1985__62__145_0.

[Kow14] E. Kowalski, *An Introduction to the Representation Theory of Groups*, *Grad. Stud. in Math.* **155**, Amer. Math. Soc., 2014. MR 3236265. Zbl 1320.20008.

[KMV02] E. Kowalski, P. Michel, and J. VanderKam, Rankin-Selberg L-functions in the level aspect, *Duke Math. J.* **114** no. 1 (2002), 123–191. MR 1915038. Zbl 1035.11018. https://doi.org/10.1215/S0012-7094-02-11416-1.

[Lau81] G. Laumon, Semi-continuité du conducteur de Swan (d'après P. Deligne), in *The Euler-Poincaré Characteristic* (*French*), *Astérisque* **83**, Soc. Math. France, Paris, 1981, pp. 173–219. MR 0629128. Zbl 0504.14013.

[Lau87] G. Laumon, Transformation de Fourier, constantes d'équations fonctionnelles et conjecture de Weil, *Inst. Hautes Études Sci. Publ. Math.* no. 65 (1987), 131–210. MR 0908218. Zbl 0641.14009. Available at http://www.numdam.org/item?id=PMIHES_1987__65__131_0.

[Lau03] G. Laumon, Transformation de Fourier homogène, *Bull. Soc. Math. France* **131** no. 4 (2003), 527–551. MR 2044494. Zbl 1088.11044.

[LRS95] W. Luo, Z. Rudnick, and P. Sarnak, On Selberg's eigenvalue conjecture, *Geom. Funct. Anal.* **5** no. 2 (1995), 387–401. MR 1334872. Zbl 0844.11038. https://doi.org/10.1007/BF01895672.

[Mil80] J. S. Milne, *Étale Cohomology*, *Princeton Math. Ser.* **33**, Princeton Univ. Press, Princeton, N.J., 1980. MR 0559531. Zbl 0433.14012.

[Mun13a] R. Munshi, Shifted convolution of divisor function d_3 and Ramanujan τ function, in *The Legacy of Srinivasa Ramanujan*, *Ramanujan Math. Soc. Lect. Notes Ser.* **20**, Ramanujan Math. Soc., Mysore, 2013, pp. 251–260. MR 3221314. Zbl 06389649.

[Mun13b] R. Munshi, Shifted convolution sums for GL(3) × GL(2), *Duke Math. J.* **162** no. 13 (2013), 2345–2362. MR 3127803. Zbl 1330.11033. https://doi.org/10.1215/00127094-2371416.

[Nun] R. M. Nunes, Squarefree integers in large arithmetic progressions. arXiv 1602.00311.

[Org03] F. Orgogozo, Altérations et groupe fondamental premier à p, *Bull. Soc. Math. France* **131** no. 1 (2003), 123–147. MR 1975807. Zbl 1083.14506.

[Sou07] K. Soundararajan, The fourth moment of Dirichlet L-functions, in *Analytic Number Theory*, *Clay Math. Proc.* **7**, Amer. Math. Soc., Providence, RI, 2007, pp. 239–246. MR 2362204. Zbl 1208.11102.

[Top16] B. Topacogullari, The shifted convolution of divisor functions, *Q. J. Math.* **67** no. 2 (2016), 331–363. MR 3509996. Zbl 1356.11070. https://doi.org/10.1093/qmath/haw010.

[You11] M. P. Young, The fourth moment of Dirichlet L-functions, *Ann. of Math.* **173** no. 1 (2011), 1–50. MR 2753598. Zbl 1296.11112. https://doi.org/10.4007/annals.2011.173.1.1.

(Received: December 3, 2015)
(Revised: April 3, 2017)

ETH Zürich, Departement Mathematik, Zürich, Switzerland,
Rämistrasse 101, CH-8092 Zürich, Switzerland
E-mail: kowalski@math.ethz.ch

Ecole Polytechnique Fédérale de Lausanne, Institute of Mathematics,
EPFL, Station 8, CH-1015 Lausanne, Switzerland
E-mail: philippe.michel@epfl.ch

ETH Zürich, Institute for Theoretical Studies, CH-8092 Zürich,
Switzerland
E-mail: william.sawin@math.ethz.ch