A THREE-DIMENSIONAL DELTA SYMBOL METHOD AND ITS APPLICATION TO INTERSECTIONS OF THREE QUADRATIC FORMS

ESRAFIL ALI MOLLA

Abstract. Using a new three-dimensional delta symbol with a Kloosterman refinement, we prove the asymptotic formula for the number of integral points of bounded height on a non-singular intersection of three quadrics in at least 22 variables, improving on the 27 variables required by Birch and by Rydin Myerson.

hexagon:2611.01222v2

Contents

1. Introduction 1
2. Overview of the method 10
3. Notation and preliminaries 14
4. Duality of divisors 17
5. Proof of the identity 21
6. Setting up the circle method 27
7. Exponential integrals 33
8. Exponential sums 34
9. The main term 42
10. Averages over the arcs 43
11. The generic terms 49
12. The exceptional terms 51
13. Comparison with other methods 54
14. The Kloosterman refinement and the case $s \geqslant 23$ 55
15. The case $s = 22$ 94
References 100

1. Introduction

In 1962 Birch [Bir62], following Davenport, used the Hardy–Littlewood circle method to obtain an asymptotic formula for the number of integral points of bounded height on a variety defined by a system of forms of the same degree, provided that the number of variables is large in terms of the degree and the number of forms. For a single form this result has been improved in a number of important cases. For quadratic forms the decisive idea goes back to Kloosterman [Klo26], who showed how to average the complete exponential sums over the numerators a of the fractions a/q around which the arcs are placed. For cubic forms Heath-Brown [HB07] reduced

2020 *Mathematics Subject Classification.* 11D45 (11D72, 11P55, 11G35, 14G05).
Key words and phrases. circle method, delta symbol, Kloosterman refinement, systems of quadratic forms, intersections of quadrics, Hasse principle.

the number of variables required from 16 to 14. The delta symbol method of Duke, Friedlander and Iwaniec [DFI93], in the form developed by Heath-Brown [HB96], gives a convenient way of carrying out such averages over a, the Kloosterman refinement, and also over q, the double Kloosterman refinement.

For systems of forms the situation is much less satisfactory. A Kloosterman refinement needs a dissection of the unit square, or of the unit cube, with the arithmetic features of the Farey dissection of the unit interval, and for a long time nothing of the kind was available. For pairs of quadratic forms, Browning and Munshi [BM13] studied intersections of two quadrics in 9 or more variables whose singular locus is a pair of conjugate points defined over $\mathbb{Q}(i)$, such as those given by $Q_1(\mathbf{x})-x_{n+1}^2-x_{n+2}^2$ and $Q_2(\mathbf{x})$, by trading the first equation for a family of congruences. Munshi [Mun15] then applied the one-dimensional delta symbol twice, in a nested manner, and obtained the asymptotic formula for an arbitrary non-singular pair in 11 or more variables. More recently Li, Rydin Myerson and Vishe [LRV24] constructed a genuinely two-dimensional delta symbol. Its moduli are as small as Dirichlet's theorem in $\mathbb{R}^2$ allows, and its weights depend on the numerator only through a lattice. This gives the asymptotic formula for non-singular pairs for even $s \geqslant 10$ and odd $s \geqslant 11$, and for $s = 9$ under the generalised Lindelöf hypothesis, by a double Kloosterman refinement [LRV24, Theorem 1.1]. For systems of R quadratic forms the best general result is due to Rydin Myerson [RM18], who obtained the asymptotic formula for non-singular systems in at least $9R$ variables. On the qualitative side, the Hasse principle for smooth intersections of two quadrics is known in 9 variables by the work of Colliot-Thélène, Sansuc and Swinnerton-Dyer [CSS87], and Heath-Brown [HB18] has reduced this to 8 variables.

In this paper we are concerned with intersections of three quadrics. Let $F_1, F_2, F_3 \in \mathbb{Z}[x_1, \dots, x_s]$ be quadratic forms, write $\vec{F} = (F_1, F_2, F_3)$, and for $\vec{\beta} \in \mathbb{C}^3$ let $F_{\vec{\beta}} = \vec{\beta} \cdot \vec{F}$, let $M_{\vec{\beta}}$ be the symmetric matrix of $F_{\vec{\beta}}$, and let $\Delta(\vec{\beta}) = \det M_{\vec{\beta}}$. We shall always suppose that the variety $V \subset \mathbb{P}^{s-1}$ defined by $F_1 = F_2 = F_3 = 0$ is a non-singular complete intersection of codimension 3. This has a simple consequence for the net of quadrics $\{F_{\vec{\beta}}\}$. Suppose that $M_{\vec{\beta}}$ had corank at least 3 for some $\vec{\beta} \neq \vec{0}$. Choose $\vec{\beta}', \vec{\beta}''$ completing $\vec{\beta}$ to a basis of $\mathbb{C}^3$. Then $F_{\vec{\beta}'}$ and $F_{\vec{\beta}''}$ would have a common non-trivial zero $\mathbf{x}$ on $\ker M_{\vec{\beta}}$, and since also $F_{\vec{\beta}}(\mathbf{x}) = 0$, the point $[\mathbf{x}]$ would lie on V. But the rows $M_1\mathbf{x}, M_2\mathbf{x}, M_3\mathbf{x}$ of the Jacobian matrix at $\mathbf{x}$ satisfy $\sum_i \beta_i M_i \mathbf{x} = \mathbf{0}$, contradicting the non-singularity of V. Thus every member of the net has corank at most 2. The singular locus in the sense of Birch, namely the set of $\mathbf{x} \in \mathbb{C}^s$ with $\operatorname{rank}(M_1\mathbf{x}, M_2\mathbf{x}, M_3\mathbf{x}) \leqslant 2$, is the union of the kernels $\ker M_{\vec{\beta}}$. The members of corank 1 form a curve, and those of corank 2 are finite in number (Lemma 6.1). So the Birch singular locus has dimension $B \leqslant 2$, and Birch's theorem, in the form [BHB17, (1.2)], gives the asymptotic formula when $s > B + (d-1)2^{d-1}R(R+1) = B + 24$, that is, for $s \geqslant 27$. Rydin Myerson's theorem gives the same range.

Our main result goes beyond this.

Theorem 1.1. *Let F_1, F_2, F_3 be quadratic forms with integral coefficients in s variables, such that the projective variety $F_1 = F_2 = F_3 = 0$ is non-singular of codimension 3. Let $\mathbf{z}_0 \in \mathbb{R}^s \setminus \{\mathbf{0}\}$ be a real zero of F_1, F_2, F_3, and let $w \in C_c^\infty(\mathbb{R}^s)$ be supported in a sufficiently small neighbourhood of $\mathbf{z}_0$. If $s \geqslant 22$, then there is $\Delta = \Delta(s) > 0$ such*

that

$$\sum_{\substack{\mathbf{x}\in\mathbb{Z}^s\\ F_1(\mathbf{x})=F_2(\mathbf{x})=F_3(\mathbf{x})=0}} w\Big(\frac{\mathbf{x}}{P}\Big) = \mathfrak{S}\,\mathfrak{J}\,P^{s-6} + O(P^{s-6-\Delta}). \tag{1.1}$$

Here the singular series $\mathfrak{S}$, defined in (9.1), depends on F_1, F_2, F_3, and the singular integral $\mathfrak{J}$, defined in (7.2), depends on F_1, F_2, F_3 and w. The implied constant depends on F_1, F_2, F_3 and w.

The proof for $s \geqslant 23$ is given in Section 14 (Theorem 14.1), and the case $s = 22$, which needs one more ingredient, in Section 15 (Theorem 15.1). The theorem uses no unproved hypothesis. The proof gives every $\Delta < \frac{1}{88}$ for $s = 22$, every $\Delta < \frac{1}{4}$ for $s = 23$ and every $\Delta < \frac{1}{2}$ for $s = 24$, and Theorem 6.5 gives every $\Delta < (s-24)/4$ for $s \geqslant 25$. The product $\mathfrak{S}\mathfrak{J}$ is positive when V has points over every completion of $\mathbb{Q}$ and $w(\mathbf{z}_0) > 0$ (Lemma 9.1), so Theorem 1.1 gives an analytic proof of the Hasse principle for these varieties in the same range. In fact the conditions at the primes are automatic here: by the work of Birch and Lewis [BL64] and Schuur [Schu80], three quadratic forms over $\mathbb{Q}_p$ in at least 13 variables have a non-trivial common zero (see also [HB10] for systems of r quadratic forms over p-adic fields). Since V is non-singular, such a zero is a point of $V(\mathbb{Q}_p)$, and Lemma 9.1 gives $\mathfrak{S} > 0$ for every $s \geqslant 22$. So the main term in (1.1) is positive as soon as $w \geqslant 0$ and $w(\mathbf{z}_0) > 0$, and in particular $V(\mathbb{Q}) \neq \varnothing$ whenever $V(\mathbb{R}) \neq \varnothing$. (Weak approximation would follow from a version of Theorem 1.1 with congruence conditions on $\mathbf{x}$; this requires only routine changes to the local computations, but we have not carried them out.) We do not claim novelty for this consequence. Heath-Brown [HB17] has proved the Hasse principle and weak approximation for non-singular intersections of three quadrics in at least 19 variables, over any number field. He finds a suitable linear space on one of the quadrics and applies his result for pairs of quadrics; his method gives no asymptotic formula. The asymptotic formula (1.1) is new for $22 \leqslant s \leqslant 26$.

The closest work to ours is that of Pierce, Schindler and Wood [PSW16], who studied the representations of a triple $\vec{n} = (n_1, n_2, n_3) \in \mathbb{Z}^3$ by a system of three quadratic forms, that is, the inhomogeneous equations $\vec{F}(\mathbf{x}) = \vec{n}$. They developed a three-dimensional form of Kloosterman's method, based on overlapping Dirichlet boxes in $\mathbb{R}^3$, and proved that the expected asymptotic formula holds for almost every triple $\vec{n}$ in at least 10 variables, provided the discriminant curve $\det(xM_1 + yM_2 + zM_3) = 0$ is smooth. Their argument averages over $\vec{n}$: a mean square over $\vec{n}$ turns the underlying exponential sum into an absolute square, as explained in [LRV24, §1.2], and this is what allows so few variables. For an individual triple, and in particular for the homogeneous problem $\vec{n} = \vec{0}$ considered in Theorem 1.1, no such average is available, and the benchmark remains Birch's theorem with 27 variables. Our unrefined method needs no average over $\vec{n}$, because all its bounds for the error terms use only the absolute values of the exponential sums and integrals, and these do not change when $\vec{F}(\mathbf{x})$ is replaced by $\vec{F}(\mathbf{x}) - \vec{n}$. Consequently the proof of Theorem 6.5 gives, for $s \geqslant 25$, $C \geqslant 1$ and every $\Delta < (s-24)/4$,

$$\sum_{\substack{\mathbf{x}\in\mathbb{Z}^s\\ \vec{F}(\mathbf{x})=\vec{n}}} w\Big(\frac{\mathbf{x}}{P}\Big) = \mathfrak{S}(\vec{n})\,\sigma_w\Big(\frac{\vec{n}}{P^2}\Big)P^{s-6} + O\Big(P^{s-6-\Delta}\Big),$$

uniformly for all $\vec{n} \in \mathbb{Z}^3$ with $|\vec{n}| \leqslant CP^2$. Here

$$\mathfrak{S}(\vec{n}) = \sum_{q} q^{-s} \sideset{}{^*}\sum_{\vec{a} \bmod q} \mathrm{e}_q(-\vec{a}\cdot\vec{n}) S_q(\vec{a}, \mathbf{0})$$

is the singular series of the inhomogeneous problem, with S_q as in (6.2), which converges absolutely and uniformly in $\vec{n}$, and σ_w is the density of the push-forward of $w(\mathbf{x})\,d\mathbf{x}$ under $\vec{F}$ (§7). The only changes in the proof are the choice of Q in §6.2, with $M_U P^2$ replaced by $M_U P^2 + |\vec{n}|$, so that still $Q^{4/3} \asymp P^2$ and $|\vec{F}(\mathbf{x}) - \vec{n}| \leqslant \frac{1}{4} Q^{4/3}$ on the support of $w(\mathbf{x}/P)$; the factor $\mathrm{e}_q(-\vec{a}\cdot\vec{n})\mathrm{e}(-\vec{w}\cdot\vec{n})$, of modulus one, in the Poisson summation formula; and the main term, which is obtained from Proposition 5.5 applied to the shifted test functions $\vec{x} \mapsto \sigma_w(\vec{n}/P^2 + Q^{4/3}P^{-2}\vec{x})$, whose Schwartz seminorms are bounded uniformly in $\vec{n}$. We leave the details to the reader. This result for *every* triple $\vec{n}$ needs more variables than [PSW16], but it has no exceptional set and needs no condition on the discriminant curve.

We mention some other related work. For pairs of quadratic forms the representation problem was studied by Heath-Brown and Pierce [HBP17], who showed that a pair of integral quadratic forms in 5 or more variables simultaneously represents almost all pairs of integers that are not excluded by local conditions, under suitable non-singularity assumptions; there too the average over the pair of integers is essential, and their proof introduced a two-dimensional version of the Kloosterman refinement for this averaged problem. The work of Pierce, Schindler and Wood [PSW16] extends this approach to three forms. For the homogeneous problem, the general theory of systems of quadratic forms goes back to Schmidt [Sch82], who found non-trivial integral zeros of systems of rational quadratic forms under conditions on the ranks of the members of the pencil; Dietmann [Die04] improved these conditions and obtained the asymptotic formula, and Rydin Myerson [RM18] obtained it for all non-singular systems of R quadratic forms in at least $9R$ variables. For an individual target vector, Kloosterman refinements for systems of forms became available only recently. Vishe [Vis23] developed a version of the Farey dissection for systems of forms over $\mathbb{F}_q(t)$, and building on these ideas Northey and Vishe [NV24] proved the Hasse principle for smooth intersections of two cubic forms over $\mathbb{Q}$ in at least 39 variables. The two-dimensional delta symbol of Li, Rydin Myerson and Vishe [LRV24], and the three-dimensional symbol of the present paper, give such a dissection over $\mathbb{Q}$ in a smooth form, with weights that depend on the numerators only through a lattice.

Let us now explain the idea behind these results. Write $\vec{n} = \vec{F}(\mathbf{x})$ with $|\mathbf{x}| \leqslant P$, so that $|\vec{n}| \ll P^2$. If one detected each of the three equations $F_i(\mathbf{x}) = 0$ separately by the one-dimensional delta symbol of [HB96], the moduli would be of size P for each equation, and the combined modulus would be of size P^3. Munshi's nesting idea [Mun15] reduces this. One detects the first equation with a modulus q_1, writes the remaining equations as the congruences $F_i(\mathbf{x}) \equiv 0 \pmod{q_1}$ together with the equations $F_i(\mathbf{x})/q_1 = 0$ for smaller integers, and detects these. Nesting the one-dimensional symbol three times gives combined moduli of size $P^{7/4}$, and nesting it with the two-dimensional symbol of [LRV24] gives $P^{5/3}$. On the other hand, Dirichlet's theorem in $\mathbb{R}^3$ approximates every point of $[0,1]^3$ by a fraction $\vec{a}/q$ with $q \leqslant Q$ to within $(qQ^{1/3})^{-1}$, and for vectors $\vec{n}$ of size P^2 this suggests moduli of size $Q \asymp P^{3/2}$. The main technical result of this paper, Theorem 1.2 below, is a three-dimensional delta symbol which reaches this size.

Since $Q \asymp P^{3/2}$ is much smaller than the square of the length P of each variable, Poisson summation in each variable saves a factor $P/Q^{1/2} = P^{1/4}$ in the critical range $q \asymp Q$. As the main term has size P^{s-6}, we need a total saving of P^6. Without any further averaging this leads to the condition $s > 24$, that is, to the asymptotic formula for $s \geqslant 25$ (Theorem 6.5 in Section 6); the computation is made precise in §1.3. This unrefined form of the method is the first step of the proof of Theorem 1.1. As in [LRV24], the weight attached to a fraction $\vec{a}/q$ depends only on the lattice $\mathbb{Z}\vec{a}+q\mathbb{Z}^3$, and so it is the same for all the numerators $\lambda\vec{a}$ with $\gcd(\lambda, q) = 1$. This makes a Kloosterman refinement possible. For the terms attached to lines the average over λ produces Gauss sums or Ramanujan sums and saves $q^{1/2}$; for the terms attached to planes and points the numerators run over a plane or over all residues, and the complete sums save $q^{1/2}$ as well. For small moduli this is too weak, and the saving comes instead from the exponential integrals, which decay faster than in the unrefined method away from the finitely many real members of corank 2; near these members we use the theorems of Roth and Schmidt on rational approximation to algebraic points. Together the two savings give $Q^{1/2} = P^{3/4}$ on the whole critical curve, against the $P^{(24-s)/4}$ that we need. For $s = 23$ this leaves a margin of $P^{1/2}$, and for $s = 22$ a margin of $P^{1/4}$.

The margin at $s = 23$ is large enough to allow a simple treatment of everything that is not generic. At the primes dividing $\gcd(q, \gcd(\vec{\rho}))$, where $\vec{\rho}$ is the long vector of the flag, and at the primes dividing the discriminant of the member $\vec{\rho}/\gcd(\vec{\rho})$, we use no cancellation at all, only the size of the Gauss sums and the fact that the dual vectors on which they do not vanish lie in a coset. The moduli on which this loses much are rare: a large value of $\gcd(q, \gcd(\vec{\rho}))$ forces the direction of the flag to be short, and the flags with short directions carry little of the mass. For the flags whose direction is a singular member of the net the kernel of the singular member takes the place of the sum over λ; what remains is handled by an average over the numerators on a plane, based on the factorisation of a binary form over p-adic fields. At $s = 22$ the same argument leaves a single loss, at the end of the critical curve where q is close to Q. It comes from members $\vec{c}$ of the net which have corank two modulo many primes. We show that such members are sparse, by a transversality argument which says that, modulo a prime power, the reductions of corank two lie on finitely many lines (Section 15).

The construction of the delta symbol itself follows the lead of [LRV24]. In two dimensions the key observation there is a duality between the divisors of a vector $\vec{n} \in \mathbb{Z}^2$ and the vectors orthogonal to it, which reduces the vanishing of $\vec{n}$ to the vanishing of a single linear form. In three dimensions we apply this duality twice, first to reduce the vanishing of $\vec{n}$ to its lying on a line, and then to reduce this to the vanishing of a linear form. The geometry of numbers in $\mathbb{R}^3$ then forces two further kinds of terms: terms attached to rational planes and terms attached to rational lines. One might at first expect that the property needed to extract the main term would require a long pointwise analysis of the weights, as in [LRV24, §4]. In fact it follows from the exactness of the identity alone (Remark 1.4), and this argument works in any dimension.

We hope that the method will also be of interest for other problems involving three equations. Systems of $R \geqslant 4$ quadrics, however, are better treated by the methods of Rydin Myerson [RM18]. An R-dimensional version of our symbol exists, but its savings fall behind $9R$ as soon as $R \geqslant 4$; see Section 13.

1.1. **A three-dimensional delta symbol.** We write

$$\delta_{\vec{n}} = \begin{cases} 1 & \text{if } \vec{n} = \vec{0}, \\ 0 & \text{if } \vec{n} \in \mathbb{Z}^R \setminus \{\vec{0}\}. \end{cases} \tag{1.2}$$

Throughout, $|\cdot|$ is the Euclidean norm. We write $\mathrm{e}(x) = \exp(2\pi i x)$ and $\mathrm{e}_q(x) = \exp(2\pi i x/q)$, and an asterisk on a sum over $\vec{a} \bmod q$ indicates that $\gcd(\vec{a}, q) = 1$. For $q \in \mathbb{N}$ and $\vec{a} \in \mathbb{Z}^3$ with $\gcd(\vec{a}, q) = 1$ let

$$\Lambda(\vec{a}, q) = \{k\vec{a} + q\vec{y} : \ k \in \mathbb{Z}, \ \vec{y} \in \mathbb{Z}^3\}. \tag{1.3}$$

We fix smooth non-negative functions ω and ω_0 on $\mathbb{R}$ as in [LRV24, §1.3]: ω is supported on $(\frac{1}{2}, 1)$ with $\int \omega = 1$, and ω_0 is even, supported on $(-\frac{1}{2}, \frac{1}{2})$, and satisfies $\omega_0(0) = 1$. For a unit vector $\hat{\vec{\nu}}$ and $\vec{w} \in \mathbb{R}^3$ we note that $|\vec{w} \times \hat{\vec{\nu}}|$ is the length of the component of $\vec{w}$ orthogonal to $\hat{\vec{\nu}}$.

Theorem 1.2. *Let $\vec{n} \in \mathbb{Z}^3$ and let $Q \geqslant 1$ be a large parameter. For each $q \in \mathbb{N}$ and $\vec{a} \in \mathbb{Z}^3$ with $\gcd(\vec{a}, q) = 1$ there exists a function $p_{\Lambda(\vec{a},q)}$ on $\mathbb{R}^3$, defined in (5.1) below, which depends only on the lattice $\Lambda(\vec{a}, q)$ and the choice of the functions ω and ω_0, such that*

$$\delta_{\vec{n}} = \sum_{1 \leqslant q \leqslant Q} \sideset{}{^*}\sum_{\vec{a} \bmod q} \int_{\mathbb{R}^3} p_{\Lambda(\vec{a},q)}(\vec{w}) \, \mathrm{e}\Big((\vec{a}/q + \vec{w}) \cdot \vec{n}\Big) \, d\vec{w} + O_N(Q^{-N}) \tag{1.4}$$

for any $N > 0$, with an implicit constant independent of $\vec{n}$. The functions $p_{\Lambda(\vec{a},q)}$ satisfy

$$\int_{\mathbb{R}^3} p_{\Lambda(\vec{a},q)}(\vec{w}) \, Q^4 \hat{g}(-Q^{4/3}\vec{w}) \, d\vec{w} = g(\vec{0}) + O_{N,\delta,g}(Q^{-N}) \qquad \text{if } q < Q^{1/3-\delta}, \tag{1.5}$$

for every Schwartz function g on $\mathbb{R}^3$ and every $\delta > 0$, where $\hat{g}(\vec{\xi}) = \int g(\vec{x}) \mathrm{e}(-\vec{\xi} \cdot \vec{x}) \, d\vec{x}$. More explicitly, we have the decomposition

$$\begin{aligned} p_{\Lambda(\vec{a},q)}(\vec{w}) = p_{1,q}(\vec{w}) &+ \sum_{\substack{\vec{\nu} \in \Lambda^\perp(\vec{a},q) \\ k = \gcd(\vec{\nu})/\gcd(\vec{\nu},q)}} \omega\Big(\frac{|\vec{\nu}|}{Q^{1/3}}\Big) p_{2,\vec{\nu},k,q}(\vec{w}) \\ &+ \sum_{\substack{(\vec{\nu},\vec{\rho}) \ \textit{flag}, \ \vec{\nu} \in \Lambda^\perp(\vec{a},q), \ \vec{\rho} \in \Lambda(\vec{a},q) \\ k = \gcd(\vec{\rho})/\gcd(\vec{\rho},q)}} \omega\Big(\frac{|\vec{\nu}|}{Q^{1/3}}\Big) \omega\Big(\frac{|\vec{\rho}|}{Q^{2/3}}\Big) p_{3,\vec{\nu},\vec{\rho},k,q}(\vec{w}), \end{aligned} \tag{1.6}$$

where $\Lambda^\perp(\vec{a}, q) = \{\vec{\nu} \in \mathbb{Z}^3 : \ q \mid \vec{\nu} \cdot \vec{a}\}$, a flag is a pair of non-zero vectors $\vec{\nu}, \vec{\rho} \in \mathbb{Z}^3$ with $\vec{\nu} \cdot \vec{\rho} = 0$ and $\gcd(\vec{\nu}) \mid \vec{\rho}$, and the functions $p_{1,q}$, $p_{2,\vec{\nu},k,q}$ and $p_{3,\vec{\nu},\vec{\rho},k,q}$, defined in (4.6)–(4.8), satisfy

$$p_{1,q}(\vec{w}) \ll_N \frac{Q}{q(1 + |\vec{w}|Q^{4/3})} \Big(1 + |\vec{w}| q Q^{1/3}\Big)^{-N}, \tag{1.7}$$

$$p_{2,\vec{\nu},k,q}(\vec{w}) \ll_N \frac{Q}{kq(1 + |\vec{w} \times \hat{\vec{\nu}}|Q^{4/3})} \Big(1 + |\vec{w} \times \hat{\vec{\nu}}| kq Q^{1/3} + |\vec{w} \cdot \hat{\vec{\nu}}| Q^{4/3}\Big)^{-N}, \tag{1.8}$$

$$p_{3,\vec{\nu},\vec{\rho},k,q}(\vec{w}) \ll_{\varepsilon,N} Q^{\varepsilon} \Big(1 + |\vec{w} \cdot \hat{\vec{\rho}}| kq Q^{1/3} + |\vec{w} \times \hat{\vec{\rho}}| Q^{4/3}\Big)^{-N}, \tag{1.9}$$

for $\frac{1}{2}Q^{1/3} < |\vec{\nu}| < Q^{1/3}$ and $\frac{1}{2}Q^{2/3} < |\vec{\rho}| < Q^{2/3}$, where $\hat{\vec{\nu}} = \vec{\nu}/|\vec{\nu}|$ and $\hat{\vec{\rho}} = \vec{\rho}/|\vec{\rho}|$.

We will also need a bound for the total mass of the weights.

Proposition 1.3. *Let $1 \leqslant q \leqslant Q$, $W > 0$ and $r_q = (qQ^{1/3})^{-1}$. Then*

$$\sum_{\vec{a} \bmod q}^{*} \int_{|\vec{w}| \leqslant W} |p_{\Lambda(\vec{a},q)}(\vec{w})|\, d\vec{w}$$
$$\ll_\varepsilon Q^\varepsilon \Big(q^2 Q \min(W, r_q)^2 \min(W, Q^{-4/3}) + qQ^2 \min(W, r_q) \min(W, Q^{-4/3})^2 \Big).$$

In particular

$$\sum_{q \leqslant Q} \sum_{\vec{a}}^{*} \int_{\mathbb{R}^3} |p_{\Lambda(\vec{a},q)}(\vec{w})|\, d\vec{w} \ll_\varepsilon Q^\varepsilon.$$

We make some comments on Theorem 1.2. Since $\delta_{\vec{n}} = \int_{[0,1]^3} \mathrm{e}(\vec{w} \cdot \vec{n})\, d\vec{w}$, the identity (1.4) may be viewed as a smooth partition of unity on $[0,1]^3$, with weights placed around the reduced fractions $\vec{a}/q$ with $q \leqslant Q$. By (1.7)–(1.9), the weight $p_{\Lambda(\vec{a},q)}(\vec{w} - \vec{a}/q)$ is essentially supported on the ball $|\vec{w} - \vec{a}/q| \ll q^{-1}Q^{-1/3+\varepsilon}$, and by Dirichlet's theorem in $\mathbb{R}^3$ these balls could not be made smaller and still cover $[0,1]^3$. For $|\vec{n}| \leqslant \frac{1}{2}Q^{4/3}$ the function $\mathrm{e}(\vec{w} \cdot \vec{n})$ hardly oscillates on the ball when $q \asymp Q$, and so we take $Q \asymp |\vec{n}|^{3/4}$ in applications.

The weight attached to $\vec{a}/q$ depends only on the lattice $\Lambda(\vec{a}, q)$, and $\Lambda(\vec{a}, q) = \Lambda(\lambda\vec{a}, q)$ whenever $\gcd(\lambda, q) = 1$, see Lemma 5.1. As in [LRV24], this is what makes a Kloosterman refinement possible. We do not use it in the unrefined argument of Sections 6–12; it is the basis of the Kloosterman refinement in Sections 14 and 15.

Remark 1.4. The property (1.5) says that $p_{\Lambda(\vec{a},q)}$ is equal to 1 when tested against functions concentrated at scale $Q^{-4/3}$, which is the scale on which the singular integral is extracted. The two-dimensional symbol satisfies the stronger pointwise statement [LRV24, (1.6)], whose proof occupies [LRV24, §4]. We do not prove a pointwise analogue of [LRV24, (1.6)] here. Our proof of (1.5) in Section 5.4 is different: we apply the identity (1.4) itself to the function $\vec{n} \mapsto g(\vec{n}/Q^{4/3})\mathrm{e}(-\vec{a} \cdot \vec{n}/q)$, and use only (1.7)–(1.9) and Poisson summation. The argument works for delta symbols of this shape in any dimension. For the one-dimensional symbol it gives the property of Heath-Brown's function h in [HB96, Lemma 9].

Remark 1.5. As in [LRV24, Remark 1.4], one might expect the weights to be independent of $\vec{a}$, as $p_{1,q}$ is. Because of the geometry of numbers in $\mathbb{R}^3$, the arcs of $p_{1,q}$ alone accumulate around fractions which lie on rational planes and rational lines of small height. The terms p_2 and p_3 correct this; they are supported on discs around fractions on rational planes and on rods around fractions on rational lines.

1.2. **Description of the method.** We now describe how Theorem 1.2 is proved, and how it is used for Theorem 1.1.

The two-dimensional symbol of [LRV24] rests on a duality between the divisors of a non-zero vector $\vec{n} \in \mathbb{Z}^2$ and the vectors $\vec{r}$ with $\vec{n} \in \mathbb{Z}\vec{r}^{\perp}$. It reduces the vanishing of $\vec{n}$ to the vanishing of the single linear form $\vec{r} \cdot \vec{n}$, which is detected by the one-dimensional symbol. In three dimensions we use this duality twice. A non-zero vector $\vec{n} \in \mathbb{Z}^3$ lies on a unique line $\mathbb{Z}\vec{\nu}_0$ with $\vec{\nu}_0$ primitive, and the divisors of $\vec{n}$ are dual to the vectors $\vec{\nu}$ with $\vec{n}/\gcd(\vec{\nu}) \in \mathbb{Z}\vec{\nu}_0$ (Lemma 4.1(i)). This reduces the vanishing of $\vec{n}$ to its lying on a line. The condition that $\vec{n}$ lies on the line through $\vec{\nu}$ is a pair of linear equations, that is, the vanishing of a vector in the two-dimensional lattice $\mathbb{Z}^3/\mathbb{Z}\vec{\nu}_0$, and to it we apply the same duality in the plane $\vec{\nu}^{\perp}$ (Lemma 4.1(ii)). This reduces the problem to the vanishing of a single linear form $\vec{c} \cdot \vec{n}$ with $\vec{c} \perp \vec{\nu}$, which we detect with the

one-dimensional symbol. At each step there is a divisor term, which accounts for the vectors $\vec{n}$ that have already been detected.

So a vector $\vec{n}$ is detected by a *flag* $(\vec{\nu}, \vec{\rho})$: a vector $\vec{\nu}$ of height about $Q^{1/3}$, a vector $\vec{\rho} \perp \vec{\nu}$ of height about $Q^{2/3}$, and a one-dimensional symbol with moduli up to $Q/\gcd(\vec{\rho})$. The heights are chosen so that every modulus that occurs is at most Q, for vectors $\vec{n}$ of size up to $Q^{4/3}$ (Proposition 4.2). The result is an exact identity with three kinds of terms: flag terms, plane terms (the divisor terms of the inner step) and point terms (the divisor terms of the outer step). Writing the periodic part of each term in lowest terms (Lemmas 5.2 and 5.3) and applying Fourier inversion gives (1.4) and (1.6). The point terms give $p_{1,q}$, supported on balls of radius $r_q = (qQ^{1/3})^{-1}$ around every fraction $\vec{a}/q$; the plane terms give $p_{2,\vec{\nu},k,q}$, supported on discs of radius r_q/k and thickness $Q^{-4/3}$ around the fractions on the plane $q \mid \vec{\nu} \cdot \vec{a}$; and the flag terms give $p_{3,\vec{\nu},\vec{\rho},k,q}$, supported on rods of length r_q/k and thickness $Q^{-4/3}$ around the fractions on the line $q \mid \vec{\rho} \times \vec{a}$ inside that plane.

For the application we take $\vec{n} = \vec{F}(\mathbf{x})$, where $\vec{F} = (F_1, F_2, F_3)$ and $|\mathbf{x}| \leqslant P$, so that $Q \asymp P^{3/2}$. We apply Poisson summation to the sum over $\mathbf{x}$ modulo q. The dual vector $\mathbf{u} = \mathbf{0}$ and the moduli $q < Q^{1/3-\delta}$ give the main term, by (1.5). Everything else is estimated with absolute values. For a fixed $\vec{a}$ the complete exponential sum is attached to the single quadratic form $\vec{a} \cdot \vec{F}$, and its absolute value is given exactly by the number $K(\vec{a}, q)$ of solutions of $M_{\vec{a}}\mathbf{h} \equiv \mathbf{0} \pmod{q}$, where $M_{\vec{a}}$ is the matrix of $\vec{a} \cdot \vec{F}$ (Lemma 8.1). The average of $K(\vec{a}, q)^{1/2}$ over the numerators and the moduli that occur in each kind of term costs only Q^{ε} (Proposition 10.7), except for flags whose direction $\vec{\rho}/\gcd(\vec{\rho})$ is a singular member of the net of quadrics, and planes which are contained in the discriminant curve of the net. For those the exponential sum vanishes unless the dual vector satisfies extra congruences (Lemma 8.3), and these compensate the larger value of $K(\vec{a}, q)$ (Proposition 12.1). The geometry of the net that we need is collected in Section 6.1: every member has corank at most 2, the members of corank 1 are non-singular points of the discriminant curve, and there are only finitely many members of corank 2.

1.3. **The savings.** We now explain where the condition $s > 24$ comes from. Trivially, the left side of (1.1) is $O(P^s)$, and the main term has size P^{s-6}, so we must save P^6 on all error terms.

Apply Theorem 1.2 with $Q \asymp P^{3/2}$, so that $Q^{4/3} \asymp P^2$, and consider the moduli $q \asymp Y$ and the range $|\vec{w}| \asymp W$. Put $X = 1 + P^2W$. By Proposition 1.3, the delta symbol costs nothing in L^1: the total mass of the weights is Q^{ε}. For fixed q, $\vec{a}$ and $\vec{w}$, Poisson summation modulo q gives

$$\sum_{\mathbf{x}} w(\mathbf{x}/P)\, \mathrm{e}\Big((\vec{a}/q + \vec{w}) \cdot \vec{F}(\mathbf{x})\Big) = q^{-s} \sum_{\mathbf{u} \in \mathbb{Z}^s} S_q(\vec{a}, \mathbf{u})\, I_q(\vec{w}, \mathbf{u}),$$

where S_q is a complete exponential sum modulo q and I_q an oscillatory integral. The factors have the following sizes:

- the integral restricts the dual vectors to $|\mathbf{u}| \ll P^{\varepsilon} qX/P$, so there are about $(qX/P)^s$ of them, and none except $\mathbf{u} = \mathbf{0}$ if $qX < P^{1-\varepsilon}$ (Lemma 7.1(i));
- the integral satisfies $|I_q(\vec{w}, \mathbf{u})| \ll P^s X^{-(s-2)/2}$ (Lemma 7.1(ii));
- the exponential sum satisfies $|S_q(\vec{a}, \mathbf{u})| = q^{s/2} K(\vec{a}, q)^{1/2}$ or 0, and $K(\vec{a}, q)$ is bounded on average (Lemma 8.1 and Proposition 10.7).

So the dual vectors $\mathbf{u} \neq \mathbf{0}$ contribute about

$$q^{-s} \cdot \left(\frac{qX}{P}\right)^s \cdot q^{s/2} \cdot P^s X^{-(s-2)/2} = q^{s/2} X^{s/2+1}$$

for each $(q, \vec{a}, \vec{w})$. We now weight this by the mass of $p_{\Lambda(\vec{a},q)}$ in the range, which by Proposition 1.3 is $Q^{\varepsilon} Y \mathfrak{M}(Y, W)$ with $\mathfrak{M}$ as in Proposition 10.7. The product is computed in Lemma 11.1:

$$Y^{s/2+1} X^{s/2+1} \mathfrak{M}(Y, W) \ll \left(\frac{YX}{Q}\right)^{s/2+2} Q^{s/2}.$$

Every range on the curve $YX \asymp Q$ contributes $Q^{s/2}$, and the ranges below it contribute less. The total error is therefore about

$$Q^{s/2} = P^{3s/4},$$

which is a saving of $P^{s/4}$ over the trivial bound. This exceeds the required P^6 exactly when $s > 24$.

Consider the extreme case $q \asymp Q$, $|\vec{w}| \leqslant P^{-2}$. There the modulus $q \asymp P^{3/2}$ is larger than the length P of the sum over each variable. Poisson summation replaces P^s terms by about $(q/P)^s = P^{s/2}$ dual terms, each of size $q^{-s} \cdot q^{s/2} \cdot P^s = P^{s/4}$, so the saving is $P^{s/4}$, or $P/Q^{1/2} = P^{1/4}$ for each variable. The mass of the weights in this range is $\asymp 1$. With the two-dimensional symbol for pairs of quadrics one has $Q \asymp P^{4/3}$, and the same computation saves $P/Q^{1/2} = P^{1/3}$ for each variable against the required P^4, that is, $s > 12$; with the one-dimensional symbol for one quadric one has $Q \asymp P$, a saving of $P^{1/2}$ for each variable, and $s > 4$.

	Q	saving per variable	needed	pointwise	λ-average
one quadric [HB96]	P	$P^{1/2}$	P^2	$s \geqslant 5$	$s \geqslant 4$
two quadrics [LRV24]	$P^{4/3}$	$P^{1/3}$	P^4	$s \geqslant 13$	$s \geqslant 11$
three quadrics	$P^{3/2}$	$P^{1/4}$	P^6	$s \geqslant 25$	$s \geqslant 22$

The last column refers to the Kloosterman refinement. The flag terms allow an average over the numerators $\lambda\vec{a}$ with $\gcd(\lambda, q) = 1$, a set of size about q, and square-root cancellation in this average saves a further $Q^{1/2} = P^{3/4}$ in the extreme case $q \asymp Q$. The error becomes $P^{3s/4-3/4}$, which is smaller than P^{s-6} for $s > 21$. We carry out the average over λ in Section 14. Square-root cancellation is available in it at the primes not dividing $2D_F\Delta(\vec{a})$, but it saves only $q^{1/2}$, which is too little when q is small. In that range the missing factor comes from the exponential integral, whose decay is better than in Lemma 7.1 away from the finitely many directions of corank 2. The two savings together give $Q^{1/2}$ at every point of the critical curve $q(1 + P^2|\vec{w}|) \asymp Q$, and after the losses described above they lead to Theorem 1.1.

For comparison, the two-dimensional symbol with these refinements gives $s \geqslant 11$ for odd s and $s \geqslant 10$ for even s, which is the unconditional theorem of [LRV24]. There the difference between even and odd s comes from the Ramanujan sums, which for even s save q rather than $q^{1/2}$. In our setting this extra saving does not help at $s = 22$, because the remaining loss comes from moduli which share most of their prime factors with fixed integers, and for these there is no sum over λ to exploit.

1.4. **Plan of the paper.** Section 2 is an informal account of the method, which explains how the delta symbol is built up from the one-dimensional case and why it saves what it saves; it also contains a table of notation. In Section 3 we set notation

and collect the properties of Heath-Brown's function h that we need. Section 4 contains the duality Lemma 4.1, the exact flag identity, and the definition and analytic properties of the p-functions. In Section 5 we prove Theorem 1.2 using properties of the lattice $\Lambda(\vec{a}, q)$, we prove the main-term property (1.5), and we prove Proposition 1.3.

The rest of the paper is devoted to the application. In Section 6 we apply Theorem 1.2, reduce Theorem 6.5 to Lemma 6.7, and collect the geometry of nets of quadrics. Sections 7 and 8 treat the exponential integrals and the exponential sums, including the local averages of $K(\vec{a}, q)$. The main term is obtained in Section 9. In Section 10 we average the exponential sums over the arcs, and in Sections 11 and 12 we bound the remaining contributions; this proves the unrefined Theorem 6.5. Section 13 compares delta symbol methods for three quadrics. Section 14 contains the Kloosterman refinement and the proof for $s \geqslant 23$, and Section 15 the additional argument for $s = 22$.

How to read the paper. The paper can be read at three levels. A reader who wants only the ideas can read the introduction and Section 2. A reader who wants to understand the method in full, but not the refinements, can read Sections 3–12, which give a complete proof of the asymptotic formula for $s \geqslant 25$; this part is largely self-contained and uses only standard tools. Sections 14 and 15 are more technical, and a guide to Section 14 is given at its beginning. Throughout the paper, most proofs begin with a few sentences explaining the idea, and the longer arguments are preceded by a heuristic account of what has to be shown and why it should be true; the reader may find it useful to read these passages first, and to return to the details later.

Declaration on the use of AI. During the preparation of this work the author used Claude, an AI assistant developed by Anthropic, to help draft and revise parts of the text and of the arguments, to check proofs, computations and references, and to prepare the LaTeX manuscript. The author has reviewed and edited all the material, and takes full responsibility for the content of the paper.

2. Overview of the method

This section explains, with as little notation as possible, how a delta symbol is built and where its saving comes from. It may be skipped by readers familiar with [HB96] and [LRV24]. Nothing in the later sections depends on it, except the notation collected in §2.7.

2.1. **Detecting zero with divisors.** Let n be an integer with $|n| \leqslant N$, and let ω be a smooth function supported in $(\frac{1}{2}, 1)$ with $\int \omega = 1$. The starting point of Duke, Friedlander and Iwaniec [DFI93] is the observation that, for $n \neq 0$, the map $d \mapsto |n|/d$ permutes the positive divisors of n, so that

$$\sum_{d \mid n} \Big(\omega\Big(\frac{d}{Q}\Big) - \omega\Big(\frac{|n|}{dQ}\Big) \Big) = 0 \qquad (n \neq 0). \tag{2.1}$$

For $n = 0$ every d divides n, the second term is $\omega(0) = 0$, and the sum is $\sum_d \omega(d/Q) \approx Q$. Thus the left side of (2.1), divided by $\sum_d \omega(d/Q)$, is exactly δ_n. If we take $Q = (2N)^{1/2}$, then in both terms only divisors $d \leqslant Q$ occur: in the first term because

of the support of ω, and in the second because $|n|/(dQ) > \frac{1}{2}$ forces $d < 2N/Q = Q$. Finally we write the condition $d \mid n$ with additive characters,

$$\mathbf{1}_{d|n} = \frac{1}{d} \sum_{a \bmod d} \mathrm{e}\Big(\frac{an}{d}\Big) = \frac{1}{d} \sum_{q|d} \sideset{}{^*}\sum_{a \bmod q} \mathrm{e}\Big(\frac{an}{q}\Big),$$

and collect the terms with the same reduced fraction a/q. The result is Heath-Brown's identity (3.3):

$$\delta_n = \frac{c_Q}{Q^2} \sum_{q \leqslant Q} \sideset{}{^*}\sum_{a \bmod q} \mathrm{e}\Big(\frac{an}{q}\Big) h\Big(\frac{q}{Q}, \frac{n}{Q^2}\Big).$$

Two features of this identity are important. First, only moduli $q \leqslant Q \asymp N^{1/2}$ occur. This is the same size as in Dirichlet's approximation theorem: every real number lies within $1/(qQ)$ of a fraction a/q with $q \leqslant Q$. Secondly, the weight $h(q/Q, n/Q^2)$ attached to a/q does not depend on a. So after the sum over n has been evaluated, one can sum the resulting complete exponential sums over a, which is Kloosterman's refinement, and even over q.

2.2. **Why a delta symbol saves.** Let F be a non-singular quadratic form in s variables, and suppose we want to count $\mathbf{x} \in \mathbb{Z}^s$ with $|\mathbf{x}| \leqslant P$ and $F(\mathbf{x}) = 0$. Then $n = F(\mathbf{x})$ has size $N \asymp P^2$, so $Q \asymp P$. For fixed q and a we apply Poisson summation to the sum over $\mathbf{x}$ modulo q:

$$\sum_{|\mathbf{x}| \leqslant P} \mathrm{e}\Big(\frac{aF(\mathbf{x})}{q}\Big) \cdots = q^{-s} \sum_{\mathbf{u} \in \mathbb{Z}^s} S_q(a, \mathbf{u}) I_q(\mathbf{u}),$$

where S_q is a complete exponential sum modulo q and I_q is an oscillatory integral. When $q \asymp Q \asymp P$, the dual vectors $\mathbf{u}$ effectively run over $|\mathbf{u}| \ll q/P \asymp 1$, the Gauss sum S_q has size $q^{s/2}$, and I_q has size P^s. So the sum over $\mathbf{x}$, which has P^s terms, is replaced by $O(1)$ terms of size $q^{-s} q^{s/2} P^s = P^{s/2}$. The saving is $P^{1/2}$ for each variable. Since we need to save P^2 to beat the trivial bound against the main term P^{s-2}, the method works when $s/2 > 2$, that is, $s \geqslant 5$. For one quadric this rough calculation already contains the main idea: *the smaller the moduli compared with P^2, the more Poisson summation saves.*

2.3. **Two equations.** For two quadratic forms the vector $\vec{n} = (F_1(\mathbf{x}), F_2(\mathbf{x})) \in \mathbb{Z}^2$ has size $N \asymp P^2$. Applying the one-dimensional symbol to each coordinate separately gives moduli $q_1, q_2 \leqslant P$, so the combined modulus $q_1 q_2$ can be as large as P^2. Poisson summation then saves only $P/(q_1 q_2)^{1/2}$ per variable, that is, nothing at all in the worst case.

Munshi's nested method [Mun15] does better. Detect $F_1(\mathbf{x}) = 0$ with a modulus $q_1 \leqslant P$. Split $F_2(\mathbf{x}) = 0$ into the congruence $F_2(\mathbf{x}) \equiv 0 \pmod{q_1}$ and the equation $F_2(\mathbf{x})/q_1 = 0$, and detect the latter, whose size is now P^2/q_1, with a modulus $q_2 \leqslant P/q_1^{1/2}$. The combined modulus is at most $P^{3/2}$.

Dirichlet's theorem in $\mathbb{R}^2$ suggests that one can go further. Every point of $[0,1]^2$ lies within $(qQ^{1/2})^{-1}$ of a fraction $\vec{a}/q$ with $\vec{a} \in \mathbb{Z}^2$ and $q \leqslant Q$. For vectors $\vec{n}$ of size N the natural choice is then $Q \asymp N^{2/3} = P^{4/3}$, and the saving per variable becomes $P/Q^{1/2} = P^{1/3}$, which gives $s > 12$ without any refinement. Li, Rydin Myerson and Vishe [LRV24] found a delta symbol with exactly these moduli. Two new features appear, and both persist in three dimensions.

The weights must depend on $\vec{a}$. In two dimensions the arcs of radius $(qQ^{1/2})^{-1}$ around all fractions $\vec{a}/q$ do not cover the square evenly. Fractions on a rational line

of small height are much denser than elsewhere, because on such a line the problem is one-dimensional, and there the arcs overlap heavily. A smooth partition of unity must therefore give less weight to these fractions. In [LRV24] the weight attached to $\vec{a}/q$ depends only on the lattice $\mathbb{Z}\vec{a} + q\mathbb{Z}^2$, which records how $\vec{a}/q$ sits with respect to the rational lines. Since $\mathbb{Z}\lambda\vec{a} + q\mathbb{Z}^2 = \mathbb{Z}\vec{a} + q\mathbb{Z}^2$ when $\gcd(\lambda, q) = 1$, a Kloosterman refinement over the numerators $\lambda\vec{a}$ is still possible.

A duality replaces (2.1). For $\vec{n} \in \mathbb{Z}^2 \setminus \{\vec{0}\}$ there are exactly two primitive vectors $\pm\vec{c}$ orthogonal to $\vec{n}$, and the divisors e of $\gcd(\vec{n})$ are in bijection, up to this sign, with the vectors $\vec{r} = e\vec{c}$. Comparing a weighted sum over these $\vec{r}$ with a weighted sum over the divisors of $\gcd(\vec{n})$, as in (2.1), gives an identity which vanishes for $\vec{n} \neq \vec{0}$. Its terms involve only the single linear form $\vec{r} \cdot \vec{n}$, whose vanishing is detected by the one-dimensional symbol.

2.4. **Three equations.** For three quadratic forms, $\vec{n} = \vec{F}(\mathbf{x}) \in \mathbb{Z}^3$ has size P^2, and Dirichlet's theorem in $\mathbb{R}^3$ suggests moduli $Q \asymp N^{3/4} = P^{3/2}$. We apply the duality twice. First, a non-zero $\vec{n} \in \mathbb{Z}^3$ lies on a unique rational line, and the divisors of $\gcd(\vec{n})$ are dual to the vectors $\vec{\nu}$ with $\vec{n}/\gcd(\vec{\nu})$ on the line $\mathbb{Z}\vec{\nu}/\gcd(\vec{\nu})$ (Lemma 4.1(i)). This reduces the vanishing of $\vec{n}$ to the statement that $\vec{n}$ lies on a given line, which is a pair of linear conditions. Secondly, these two conditions describe the vanishing of a vector in a two-dimensional lattice, and the two-dimensional duality (Lemma 4.1(ii)) reduces it to the vanishing of a single linear form $\vec{c} \cdot \vec{n}$ with $\vec{c} \perp \vec{\nu}$. Finally the one-dimensional symbol detects $\vec{c} \cdot \vec{n} = 0$.

So each vector is detected by a *flag*: a vector $\vec{\nu}$ of height about $Q^{1/3}$, a vector $\vec{\rho} = G\vec{c} \perp \vec{\nu}$ of height about $Q^{2/3}$, and a one-dimensional symbol with moduli up to Q/G. Each step of the duality also produces a "divisor term", which accounts for the vectors that have already been detected. In the end there are three kinds of terms, corresponding to the three kinds of special positions a fraction $\vec{a}/q$ can have:

term	fractions $\vec{a}/q$ involved	support of the weight	weight
point	all fractions	ball of radius $r_q = (qQ^{1/3})^{-1}$	$p_{1,q}$
plane	fractions on a plane $q \mid \vec{\nu} \cdot \vec{a}$	disc of radius r_q/k, thickness $Q^{-4/3}$	$p_{2,\vec{\nu},k,q}$
flag	fractions on a line $q \mid \vec{\rho} \times \vec{a}$	rod of length r_q/k, thickness $Q^{-4/3}$	$p_{3,\vec{\nu},\vec{\rho},k,q}$

Here r_q is the radius suggested by Dirichlet's theorem. To get a feeling for the sizes, take $P = 10^4$. Then $Q = P^{3/2} = 10^6$, the vectors $\vec{\nu}$ have length about 10^2, the vectors $\vec{\rho}$ about 10^4, the arcs have radius $r_q \approx 10^{-2}/q$, and Poisson summation saves a factor $P^{1/4} = 10$ for each variable in the critical range $q \asymp Q$. With $s = 25$ variables this is a total saving of 10^{25}, against the $P^6 = 10^{24}$ that is required: a margin of $10 = P^{1/4}$.

2.5. **Extracting the main term.** The main term comes from the dual vector $\mathbf{u} = \mathbf{0}$ and from small moduli q. To extract it one needs to know that the weight $p_{\Lambda(\vec{a},q)}$ behaves like the indicator function of its arc when tested against slowly varying functions. In two dimensions this is proved in [LRV24, §4] by a careful pointwise analysis of the weights. We use a shortcut (Proposition 5.5). We apply the exact identity (1.4) itself to the test function $\vec{n} \mapsto g(\vec{n}/Q^{4/3})\mathrm{e}(-\vec{a} \cdot \vec{n}/q)$, and use Poisson summation on the sum over $\vec{n}$. The term we want appears as the zero frequency, and all other frequencies are negligible by the decay of the weights. The argument uses only the exactness of the identity and the bounds (1.7)–(1.9), and it works for symbols of this shape in any dimension.

2.6. **The structure of the proof of Theorem 1.1.** The proof proceeds in seven steps, which correspond to the sections of the paper. Steps 1–6 give the unrefined Theorem 6.5, and step 7 completes the proof of Theorem 1.1.

1. *The identity* (Sections 4 and 5). We construct the delta symbol, prove the bounds for the weights and their total mass, and prove the main-term property.
2. *Poisson summation* (Section 6). We insert the identity with $\vec{n} = \vec{F}(\mathbf{x})$, apply Poisson summation, and split the result into the main term N_0, the generic terms N_1 and the exceptional terms N_2. We also collect the geometry of the net of quadrics.
3. *Integrals and sums* (Sections 7 and 8). The integrals I_q restrict the dual vectors to $|\mathbf{u}| \ll qX/P$ and have size $P^s X^{-(s-2)/2}$. The sums $S_q(\vec{a}, \mathbf{u})$ are Gauss sums for the single quadratic form $\vec{a} \cdot \vec{F}$, and their absolute value is given exactly by the number $K(\vec{a}, q)$ of solutions of a linear congruence.
4. *The main term* (Section 9). It gives the singular series and the singular integral.
5. *Averages of* $K(\vec{a}, q)$ (Sections 8 and 10). The quantity $K(\vec{a}, q)$ is large when the form $\vec{a} \cdot \vec{F}$ is close to a singular member of the net, and we show that this is rare on average over the fractions that occur in each kind of term.
6. *The error terms* (Sections 11 and 12). We add up the bounds over dyadic ranges of q and $|\vec{w}|$. All ranges on the curve $q(1 + P^2|\vec{w}|) \asymp Q$ contribute the same amount $Q^{s/2} = P^{3s/4}$, and this gives the condition $s > 24$.
7. *The Kloosterman refinement* (Sections 14 and 15). We use that the weights are the same for all numerators $\lambda\vec{a}$, and sum the exponential sums over λ before taking absolute values. Together with a sharper treatment of the integrals for small moduli, this saves a further factor $Q^{1/2}$ on the critical curve. At the primes shared by q and the flag, and at the primes dividing the discriminant of its direction, we use only the size of the Gauss sums; this costs a factor which is large only for rare moduli, and gives $s \geqslant 23$. For $s = 22$ we show in addition that the members of the net which have corank two modulo many primes are sparse.

2.7. **Notation.** For the convenience of the reader we list the main notation.

P, s	size of the variables, number of variables
$\vec{F} = (F_1, F_2, F_3)$	the quadratic forms; $F_{\vec{\beta}} = \vec{\beta} \cdot \vec{F}$, with matrix $M_{\vec{\beta}}$
$\Delta(\vec{\beta})$	$\det M_{\vec{\beta}}$, the discriminant of the net, a ternary form of degree s
Q	the parameter of the delta symbol, $Q \asymp P^{3/2}$, so that $Q^{4/3} \asymp P^2$
q, $\vec{a}$	modulus and numerator of a fraction $\vec{a}/q$, $\vec{a} \in \mathbb{Z}^3$, $\gcd(\vec{a}, q) = 1$
$\Lambda(\vec{a}, q)$, $\Lambda^{\perp}(\vec{a}, q)$	the lattice $\mathbb{Z}\vec{a} + q\mathbb{Z}^3$, and $\{\vec{\nu} :\ q \mid \vec{\nu} \cdot \vec{a}\}$
flag $(\vec{\nu}, \vec{\rho})$	$\vec{\nu} \perp \vec{\rho}$, $\gcd(\vec{\nu}) \mid \vec{\rho}$; $\vec{\rho} = G\vec{c}$ with $\vec{c}$ primitive
p_1, p_2, p_3	the weights of the point, plane and flag terms
$\vec{w}$, X	distance from $\vec{a}/q$, and $X = 1 + P^2\lvert\vec{w}\rvert$
$\mathbf{u}$, L	dual vector after Poisson summation, and its effective range $L \asymp qX/P$
$S_q(\vec{a}, \mathbf{u})$, $I_q(\vec{w}, \mathbf{u})$	complete exponential sum and oscillatory integral, (6.2) and (6.3)
$K(\vec{a}, q)$	number of $\mathbf{h} \bmod q$ with $M_{\vec{a}}\mathbf{h} \equiv \mathbf{0}$; $\lvert S_q\rvert = q^{s/2}K^{1/2}$ or 0
$\mathcal{Z}_2$, $\mathcal{C}_F$, $\mathcal{N}_F$	points of corank 2; singular primitive vectors; exceptional plane directions
D_F, $b(q)$	product of bad primes, and the part of q composed of them
$\mathfrak{M}(Y, W)$	normalised mass of the weights with $q \asymp Y$, $\lvert\vec{w}\rvert \leqslant W$, (10.1)

3. Notation and preliminaries

We first set some notation which will be used throughout. The notation $|\cdot|$, $\delta_{\vec{n}}$, $\mathrm{e}(x)$, $\mathrm{e}_q(x)$, $\sideset{}{^*}\sum_{\vec{a} \bmod q}$ and $\Lambda(\vec{a}, q)$ will always be as defined before Theorem 1.2. We use $\vec{a}, \vec{b}, \vec{c}, \vec{n}, \vec{r}, \vec{\nu}, \vec{\rho}$ to denote vectors in $\mathbb{Z}^3$, $\vec{w}, \vec{x}, \vec{y}, \vec{\xi}$ to denote vectors in $\mathbb{R}^3$, and $\mathbf{x}, \mathbf{u}, \mathbf{h}, \dots$ to denote vectors in $\mathbb{Z}^s$ or $\mathbb{R}^s$, where s is the number of variables in Theorem 6.5. For a non-zero integer vector $\vec{x}$ we write $\gcd(\vec{x})$ for the greatest common divisor of its coordinates and $\gcd(\vec{x}, q) = \gcd(\gcd(\vec{x}), q)$; for $g \in \mathbb{N}$, $g \mid \vec{x}$ means that g divides every coordinate of $\vec{x}$. We set $\mathbf{1}_S = 1$ if S is true and 0 otherwise. We write $A \ll B$ or $A = O(B)$ if $|A| \leqslant C|B|$ for an implicit constant C, and $A \asymp B$ if $A \ll B \ll A$. From Section 6 onward we fix the forms F_1, F_2, F_3 and the weight w, we use ε for an arbitrarily small positive constant whose value may change from line to line, and all implicit constants may depend on s, w, F_1, F_2, F_3 and ε.

This section collects the one-dimensional tools that the rest of the paper is built on: the function h from Heath-Brown's delta symbol, with the bounds we need for it, and a form of Poisson summation over lattices. The reader who knows [HB96] can skim the statements and move on; the proofs are short and are included to keep the paper self-contained.

For $m \in \mathbb{N}$ we define

$$c_m = \left(\int_{\mathbb{R}^m} \omega(|\vec{x}|)\, d\vec{x} \right)^{-1}. \tag{3.1}$$

Since ω_0 is even and smooth, $\vec{x} \mapsto \omega_0(|\vec{x}|)$ is a smooth function on $\mathbb{R}^m$.

As in [HB96] and [LRV24, (2.2)], we define for $y > 0$ and $z \in \mathbb{R}$

$$h(y,z) = \sum_{j\in\mathbb{N}} \frac{1}{yj}\Big(\omega(yj) - \omega\Big(\frac{|z|}{yj}\Big)\Big). \tag{3.2}$$

This function appears in the one-dimensional delta symbol: by [HB96, Theorem 1], for $Q_1 \geqslant 1$ and $t \in \mathbb{Z}$,

$$\delta_t = \frac{c_{Q_1}}{Q_1^2} \sum_{q\in\mathbb{N}} \sideset{}{^*}\sum_{a \bmod q} \mathrm{e}_q(at)\, h\Big(\frac{q}{Q_1}, \frac{t}{Q_1^2}\Big), \qquad c_{Q_1} = Q_1\Big(\sum_{m\in\mathbb{N}} \omega(m/Q_1)\Big)^{-1}, \tag{3.3}$$

and $c_{Q_1} = 1 + O_N(Q_1^{-N})$ by Poisson summation. The identity (3.3) is exact. It is worth keeping a picture of h in mind. For a fixed small $y = q/Q_1$, the function $z \mapsto h(y,z)$ is essentially constant, of size at most $1/y$, on the interval $|z| < y/2$; it then decays, and becomes negligible as soon as $|z|$ is a little larger than y. So $h(q/Q_1, t/Q_1^2)$ behaves like a smoothed version of Q_1/q times the indicator function of $|t| \lesssim qQ_1$. Lemma 3.1 makes this precise: part (1) gives the support and the size, parts (2) and (3) give the derivatives in the two regions, and part (4) says that, although h can be as large as $1/y$, its integral over z is essentially bounded. We need the following properties of h; parts (1) and (2) are [HB96, Lemma 4], and we include short proofs of the remaining parts.

Lemma 3.1. *Let $y > 0$ and $z \in \mathbb{R}$. Part (1) holds for all $y > 0$; in parts (2)–(4) we assume $0 < y \leqslant 1$.*

(1) The value of $h(y,z)$ is non-zero only if $y \leqslant \max(1, 2|z|)$. Moreover $h(y,z) \ll 1/y$ for all z, and $h(y,z)$ does not depend on z for $|z| < y/2$.
(2) For $i \geqslant 1$, $\partial_z^i h(y,z)$ vanishes for $|z| < y/2$, and $\partial_z^i h(y,z) \ll_i y^{-1}|z|^{-i}$.
(3) For any $\varepsilon, N > 0$ and $i \geqslant 0$, we have $\partial_z^i h(y,z) \ll_{\varepsilon,N,i} y^N$ for $y^{1-\varepsilon} \leqslant |z| \leqslant 1$.
(4) For any $\varepsilon > 0$, $\int_{-1}^1 |h(y,z)|\, dz \ll_\varepsilon y^{-\varepsilon}$, and $\int_{-1}^1 |\partial_z^i h(y,z)|\, dz \ll_{\varepsilon,i} y^{-i-\varepsilon}$ for $i \geqslant 1$.

Proof. Everything follows from the definition (3.2), in which h is the difference of two sums over j. The first sum does not depend on z, and the second one is only present when $|z|$ is comparable to yj. Parts (1) and (2) count the terms which can be non-zero, and part (3) uses Poisson summation in j to see a cancellation between the two sums.

(1) If $y \geqslant 1$ then $\omega(yj) = 0$ for all $j \in \mathbb{N}$, and if moreover $y \geqslant 2|z|$ then $|z|/(yj) \leqslant \frac{1}{2}$, so $\omega(|z|/(yj)) = 0$. The first sum in (3.2) has at most $1/(2y) + 1$ non-zero terms, each $\ll 1$. A term of the second sum is non-zero only if $|z| < yj < 2|z|$, which forces $|z| \geqslant y/2$; there are at most $3|z|/y$ such j, and each term is $\ll 1/|z|$.

(2) We differentiate the second sum term by term. Up to sign, the j-th term contributes

$$(yj)^{-1-i}\, \omega^{(i)}\Big(|z|/(yj)\Big) \ll |z|^{-1-i},$$

and it vanishes unless $|z| < yj < 2|z|$. There are at most $3|z|/y$ such j.

Part (3) is the only place where something more than counting is needed. In the range $y^{1-\varepsilon} \leqslant |z| \leqslant 1$ each of the two sums in (3.2) is of size about $1/y$, and the point is that they cancel almost completely. The reason is that both sums are Riemann sums, with step y, of functions with the same integral, and Riemann sums of smooth functions are extremely accurate.

(3) Fix z with $y^{1-\varepsilon} \leqslant |z| \leqslant 1$, and let $f(t) = t^{-1}(\omega(t) - \omega(|z|/t))$ for $t > 0$ and $f(t) = 0$ for $t \leqslant 0$. Then f is smooth, since $\omega(|z|/t)$ vanishes for $t \leqslant |z|$, and

it is supported in $[\min(|z|, \frac{1}{2}), 2]$; moreover $|f^{(M)}(t)| \ll_M |z|^{-M-1}$ on $[|z|, 2|z|]$ and $|f^{(M)}(t)| \ll_M 1$ elsewhere, and $h(y,z) = \sum_{j\in\mathbb{Z}} f(yj) = y^{-1}\sum_{k\in\mathbb{Z}} \hat{f}(k/y)$ by Poisson summation. The substitution $t \mapsto |z|/t$ shows that $\hat{f}(0) = \int f = 0$. For $k \neq 0$, integration by parts gives $\hat{f}(k/y) \ll_M (y/|k|)^M \int |f^{(M)}| \ll_M (y/|k|)^M |z|^{-M}$, since $|z| \leqslant 1$. Hence $h(y,z) \ll_M y^{-1}(y/|z|)^M \leqslant y^{\varepsilon M - 1}$. Choosing M large in terms of ε and N gives the bound y^N. The same argument applies to $\partial_z^i h$, since $\partial_z^i f$ also has integral zero and the same support.

(4) This follows from (1)–(3): split the interval $[-1,1]$ into the range $|z| < y^{1-\varepsilon}$, where we use the trivial bounds of (1) and (2), and the range $|z| \geqslant y^{1-\varepsilon}$, where (3) shows that h is negligible. Thus the integral of $|h|$ is $\ll y^{-1}\cdot y^{1-\varepsilon} + y^N$, and for $i \geqslant 1$ the integral of $|\partial_z^i h|$ is

$$\ll y^{-1}\int_{y/2}^{y^{1-\varepsilon}} z^{-i}\,dz + y^N \ll y^{-i-\varepsilon}.$$

□

We shall also use Poisson summation on lattices in the following form.

The lemma below says that a smooth count of the points of a lattice in a ball of radius T is equal to the volume of the ball divided by the covolume, with an error which is negligible as soon as T is large compared with the inverse of the shortest dual vector. In other words, no lattice point counting error appears once the lattice is "well spread" at scale T.

Lemma 3.2. *Let L be a lattice of rank m in a Euclidean space V, and suppose that the dual lattice L^* has no non-zero vector of length less than $\lambda > 0$. Then for $T \geqslant 1$ with $\lambda T \geqslant 1$ and any $N > 0$*

$$\sum_{\vec{r}\in L} \omega\Big(\frac{|\vec{r}|}{T}\Big) = \frac{T^m}{c_m \operatorname{covol}(L)}\Big(1 + O_N\big((\lambda T)^{-N}\big)\Big).$$

Proof. The idea is that Poisson summation turns the sum over L into a sum over the dual lattice L^*, in which the zero vector gives the volume term and every other vector is far out in the tail of a Schwartz function. By Poisson summation the sum is $\operatorname{covol}(L)^{-1}T^m\sum_{\vec{\xi}\in L^*}\hat{f}(T\vec{\xi})$ with $f(\vec{y}) = \omega(|\vec{y}|)$. The term $\vec{\xi} = \vec{0}$ gives the main term. The balls of radius $\lambda/2$ around the points of L^* are disjoint, so the number of $\vec{\xi} \in L^*$ with $|\vec{\xi}| \leqslant u$ is $\leqslant (1 + 2u/\lambda)^m \ll (u/\lambda)^m$ for $u \geqslant \lambda$. Since $\hat{f}$ is a Schwartz function, $|\hat{f}(T\vec{\xi})| \ll_{N'} (T|\vec{\xi}|)^{-N'}$, and splitting the $\vec{\xi} \neq \vec{0}$ into the shells $2^i\lambda \leqslant |\vec{\xi}| < 2^{i+1}\lambda$ we obtain

$$\sum_{\vec{\xi}\neq\vec{0}} |\hat{f}(T\vec{\xi})| \ll_{N'} \sum_{i\geqslant 0} 2^{(i+1)m}(2^i\lambda T)^{-N'} \ll (\lambda T)^{-N'}$$

for $N' > m$. So the other terms contribute $\ll_N (\lambda T)^{-N}T^m/\operatorname{covol}(L)$. □

For a non-zero $\vec{\nu} \in \mathbb{Z}^3$ we write $g = \gcd(\vec{\nu})$, $\vec{\nu}_0 = \vec{\nu}/g$, and

$$L_{\vec{\nu}} = \{\vec{r} \in \mathbb{Z}^3 :\ \vec{r}\cdot\vec{\nu} = 0\}.$$

This is a primitive lattice of rank 2 and covolume $|\vec{\nu}_0|$. We fix a 3×2 integer matrix $B_{\vec{\nu}}$ whose columns form a basis of $L_{\vec{\nu}}$; then $\vec{x} \mapsto B_{\vec{\nu}}^T\vec{x}$ maps $\mathbb{Z}^3$ onto $\mathbb{Z}^2$ with kernel $\mathbb{Z}\vec{\nu}_0$. The dual lattice of $L_{\vec{\nu}}$ in the plane $\vec{\nu}^\perp$ is the orthogonal projection of $\mathbb{Z}^3$ to that plane, and its non-zero vectors have length at least $1/|\vec{\nu}_0|$, since the projection of $\vec{x}$ has length $|\vec{x}\times\vec{\nu}_0|/|\vec{\nu}_0|$. In the language of Lemma 3.2, the lattice $L_{\vec{\nu}}$ may therefore

be counted in balls of radius T with a negligible error as soon as T is a little larger than $|\vec{\nu}_0|$. This is how Lemma 3.2 will be used in the construction of the delta symbol.

4. Duality of divisors

In this section we make the preparations for the proof of Theorem 1.2. We begin with the duality of divisors on which the construction rests. It is the three-dimensional counterpart of [LRV24, Lemma 3.1], which is itself a higher-dimensional analogue of the identity used by Duke, Friedlander and Iwaniec [DFI93, (2.3)].

Before stating the lemma we explain the idea in its simplest form. In one dimension, a non-zero integer n has the divisors d and $|n|/d$ in pairs, and the identity (2.1) compares a sum over d with the same sum over $|n|/d$; the two agree when $n \neq 0$, and the difference only survives at $n = 0$. In three dimensions a non-zero vector $\vec{n}$ determines a rational line, namely the line through $\vec{n}$, and on this line $\vec{n}$ is λ times a primitive vector, with $\lambda = \gcd(\vec{n})$. Part (i) of the lemma is the one-dimensional identity applied to λ, written in terms of the vectors $\vec{\nu} = g\vec{\nu}_0$ on this line. Part (ii) is the same statement one dimension lower: inside a plane $L_{\vec{\nu}_0}$, a non-zero vector $\vec{m}$ determines a unique line, now described by the primitive vectors $\pm\vec{c}$ orthogonal to it. In both parts the right-hand side involves only the size of $\vec{n}$ (or of $\vec{x} \times \vec{\nu}_0$), and this is what will allow us to subtract it off.

4.1. **A duality between divisors and flags.**

Lemma 4.1. *Let $W : [0, \infty) \to \mathbb{R}$ be any function.*

(i) For $\vec{n} \in \mathbb{Z}^3 \setminus \{\vec{0}\}$ we have

$$\sum_{\vec{\nu} \in \mathbb{Z}^3 \setminus \{\vec{0}\}} \mathbf{1}_{g|\vec{n}}\, \mathbf{1}_{\vec{n}/g \in \mathbb{Z}\vec{\nu}_0}\, W(|\vec{\nu}|) = 2 \sum_{g' | \vec{n}} W\Big(|\vec{n}|/g'\Big),$$

where $g = \gcd(\vec{\nu})$ and $\vec{\nu}_0 = \vec{\nu}/g$.

(ii) Let $\vec{\nu}_0$ be primitive and let $\vec{x} \in \mathbb{Z}^3$ with $\vec{m} = B_{\vec{\nu}_0}^T \vec{x} \neq \vec{0}$. Then

$$\sum_{\substack{\vec{r} \in L_{\vec{\nu}_0} \setminus \{\vec{0}\} \\ \vec{r} = e\vec{c},\ \vec{c} \text{ primitive}}} \mathbf{1}_{e|\vec{m}}\, \mathbf{1}_{\vec{c}\cdot\vec{x}=0}\, W(|\vec{r}|) = 2 \sum_{e' | \vec{m}} W\Big(|\vec{x} \times \vec{\nu}_0|/e'\Big).$$

Proof. (i) There are exactly two primitive vectors parallel to $\vec{n}$. The conditions on the left force $\vec{\nu}$ to be parallel to $\vec{n}$, so the only question is which multiples of these two primitive vectors occur, and with what weight. If $\vec{n} = \lambda\vec{\nu}_0$ with $\lambda \in \mathbb{N}$, the conditions say that $g \mid \lambda$, and $g \mapsto g' = \lambda/g$ permutes the divisors of λ, with $|\vec{\nu}| = g|\vec{\nu}_0| = |\vec{n}|/g'$. The same holds for $-\vec{\nu}_0$. So each divisor g' of λ is counted twice on the left, once for $\vec{\nu}_0$ and once for $-\vec{\nu}_0$, and with the weight $W(|\vec{n}|/g')$; this is the factor 2 on the right.

(ii) Write $B = B_{\vec{\nu}_0}$ and $\vec{c} = B\vec{s}$ with $\vec{s} \in \mathbb{Z}^2$ primitive. Then $\vec{c}\cdot\vec{x} = \vec{s}\cdot\vec{m}$, so $\vec{c}\cdot\vec{x} = 0$ if and only if $\vec{m} = \pm\lambda\vec{s}^{\perp}$, where $\lambda = \gcd(\vec{m})$ and $\vec{s}^{\perp} = (-s_2, s_1)$. This determines $\vec{s}$ up to sign, and the condition $e \mid \vec{m}$ says that $e \mid \lambda$. So, exactly as in part (i), the vectors $\vec{r}$ which occur are $e\vec{c}$ for the two choices $\pm\vec{c}$ and for the divisors e of λ; it only remains to express $|\vec{r}|$ in terms of the divisor $e' = \lambda/e$. This is a computation with the Gram matrix of the basis B. Let $G = B^T B$, so that $\det G = |\vec{\nu}_0|^2$, and let $\pi(\vec{x}) = BG^{-1}\vec{m}$ be the orthogonal projection of $\vec{x}$ to $\vec{\nu}_0^{\perp}$. For a symmetric 2×2 matrix we have $(\vec{s}^{\perp})^T \operatorname{adj}(G)\vec{s}^{\perp} = \vec{s}^T G\vec{s}$, hence

$$|\pi(\vec{x})|^2 = \vec{m}^T G^{-1}\vec{m} = \frac{\lambda^2\, \vec{s}^T G\vec{s}}{|\vec{\nu}_0|^2} = \frac{\lambda^2 |B\vec{s}|^2}{|\vec{\nu}_0|^2}.$$

Therefore $|\vec{r}| = e|\vec{c}| = e|\vec{\nu}_0||\pi(\vec{x})|/\lambda = |\vec{x}\times\vec{\nu}_0|/e'$ with $e' = \lambda/e$. Here we used $|B\vec{s}| = |\vec{c}|$ and $|\vec{\nu}_0||\pi(\vec{x})| = |\vec{x}\times\vec{\nu}_0|$, the latter because the projection of $\vec{x}$ to $\vec{\nu}_0^{\perp}$ has length $|\vec{x}\times\vec{\nu}_0|/|\vec{\nu}_0|$. As in (i), the map $e \mapsto e'$ is a bijection of the divisors of λ, and the factor 2 comes from the two signs of $\vec{c}$. □

We now combine the two parts of the lemma with the one-dimensional delta symbol. Put $R_1 = Q^{1/3}$, and for a non-zero $\vec{\nu}\in\mathbb{Z}^3$ with $g = \gcd(\vec{\nu})$ put $\varrho_{\vec{\nu}} = Q^{2/3}/g$.

The proposition below is the delta symbol before Fourier analysis. It expresses $\delta_{\vec{n}}$ as a sum of three kinds of terms. The *flag terms* (4.1) come from detecting $\vec{n} = \vec{0}$ in three steps: $\vec{n}$ lies on the line $\mathbb{R}\vec{\nu}$, then inside the plane $L_{\vec{\nu}}$ the projected vector vanishes along the direction $\vec{c}$, and finally the single integer $\vec{c}\cdot\vec{n}/(ge)$ vanishes, which is detected by Heath-Brown's one-dimensional symbol. The *plane terms* (4.2) and the *point terms* (4.3) are the "divisor sums" on the right of Lemma 4.1(ii) and (i), which have to be subtracted at each step. The parameters are chosen so that every vector $\vec{\nu}$ has length about $R_1 = Q^{1/3}$ and every vector $\vec{r}$ has length about $Q^{2/3}/g$; with these sizes the moduli $ge\,q_1$ of the final step never exceed Q, and the lattice counts in the normalisation are as accurate as Lemma 3.2 allows.

Proposition 4.2. *Let $Q \geqslant 1$ and let $\vec{n}\in\mathbb{Z}^3$ with $|\vec{n}| < \frac{1}{2}Q^{4/3}$. In the sums below, $g = \gcd(\vec{\nu})$, and $\vec{r} = e\vec{c}$ with $e = \gcd(\vec{r})$. For any $N > 0$ we have*

$$\delta_{\vec{n}} = \frac{c_2c_3}{Q^{13/3}}\sum_{\vec{\nu}\neq\vec{0}}\omega\Big(\frac{|\vec{\nu}|}{R_1}\Big)|\vec{\nu}|\sum_{\vec{r}\in L_{\vec{\nu}}\setminus\{\vec{0}\}}\omega\Big(\frac{|\vec{r}|}{\varrho_{\vec{\nu}}}\Big)\,g^3e^2\,\mathbf{1}_{g|\vec{n}}\,\mathbf{1}_{e|B_{\vec{\nu}}^T\vec{n}/g}$$

$$\times\sum_{q_1\leqslant Q/(ge)}\sideset{}{^*}\sum_{a \bmod q_1}\mathrm{e}_{q_1}\Big(\frac{a\,\vec{c}\cdot\vec{n}}{ge}\Big)h\Big(\frac{geq_1}{Q},\frac{ge\,\vec{c}\cdot\vec{n}}{Q^2}\Big) \tag{4.1}$$

$$-\frac{2c_2c_3}{Q^{7/3}}\sum_{\vec{\nu}\neq\vec{0}}\omega\Big(\frac{|\vec{\nu}|}{R_1}\Big)|\vec{\nu}|\,g\,\mathbf{1}_{g|\vec{n}}\sum_{e'|B_{\vec{\nu}}^T\vec{n}/g}\omega\Big(\frac{|\vec{n}\times\vec{\nu}_0|}{e'Q^{2/3}}\Big) \tag{4.2}$$

$$-\frac{2c_3}{Q}\sum_{g'|\vec{n}}\omega\Big(\frac{|\vec{n}|}{g'R_1}\Big)+O_N(Q^{-N}), \tag{4.3}$$

where the implicit constant does not depend on $\vec{n}$.

Proof. The proof follows the three steps just described. At each of the first two steps we apply Lemma 4.1, which says that a weighted count of the vectors detecting $\vec{n}$ equals a divisor sum when $\vec{n}\neq\vec{0}$; the difference of the two is therefore a multiple of $\delta_{\vec{n}}$, and we only have to compute the multiple, which is a lattice point count. At the last step we insert (3.3).

The outer step. Since $\mathbf{1}_{\vec{x}\in\mathbb{Z}\vec{v}_0} = \delta_{B_{\vec{\nu}}^T\vec{x}}$, Lemma 4.1(i) with $W(t) = \omega(t/R_1)$ gives, for $\vec{n}\neq\vec{0}$,

$$\sum_{\vec{\nu}}\omega(|\vec{\nu}|/R_1)\,\mathbf{1}_{g|\vec{n}}\,\delta_{B_{\vec{\nu}}^T\vec{n}/g}-2\sum_{g'|\vec{n}}\omega\Big(|\vec{n}|/(g'R_1)\Big)=0. \tag{4.4}$$

For $\vec{n}=\vec{0}$ the second sum is empty, and the first sum equals

$$S_3 = \sum_{\vec{\nu}\in\mathbb{Z}^3}\omega(|\vec{\nu}|/R_1) = c_3^{-1}Q(1+O_N(Q^{-N}))$$

by Lemma 3.2. Thus $\delta_{\vec{n}}$ equals S_3^{-1} times the left side of (4.4). In words: $\vec{n}$ vanishes if and only if every $\vec{\nu}$ "sees" it, and the count of all $\vec{\nu}$ is the normalising factor S_3.

The inner step. Fix $\vec{\nu}$ with $g \mid \vec{n}$, and put $\vec{x} = \vec{n}/g$ and $\vec{m} = B_{\vec{\nu}}^T \vec{x}$. Lemma 4.1(ii) with $W(t) = \omega(t/\varrho_{\vec{\nu}})$ gives, for $\vec{m} \neq \vec{0}$,

$$\sum_{\vec{r}=e\vec{c}\in L_{\vec{\nu}}\setminus\{\vec{0}\}} \omega(|\vec{r}|/\varrho_{\vec{\nu}})\, \mathbf{1}_{e|\vec{m}}\, \delta_{\vec{c}\cdot\vec{x}/e} - 2\sum_{e'|\vec{m}} \omega\Big(\frac{|\vec{x}\times\vec{\nu}_0|}{e'\varrho_{\vec{\nu}}}\Big) = 0. \tag{4.5}$$

Note that $e \mid \vec{m}$ implies $e \mid \vec{c}\cdot\vec{x}$. For $\vec{m} = \vec{0}$ the second sum is empty and the first equals $S_2(\vec{\nu}) = \sum_{\vec{r}\in L_{\vec{\nu}}} \omega(|\vec{r}|/\varrho_{\vec{\nu}})$. By Lemma 3.2 in the plane $\vec{\nu}^\perp$, with $\lambda = 1/|\vec{\nu}_0|$ and $\lambda\varrho_{\vec{\nu}} = Q^{2/3}/|\vec{\nu}| \geqslant Q^{1/3}$ when $|\vec{\nu}| \leqslant R_1$, we have $S_2(\vec{\nu}) = \varrho_{\vec{\nu}}^2 c_2^{-1} |\vec{\nu}_0|^{-1}(1 + O_N(Q^{-N}))$. Thus $\delta_{\vec{m}}$ equals $S_2(\vec{\nu})^{-1}$ times the left side of (4.5). Since $\vec{m} = B_{\vec{\nu}}^T \vec{n}/g$ vanishes exactly when $\vec{n}/g \in \mathbb{Z}\vec{\nu}_0$, this replaces the factor $\delta_{B_{\vec{\nu}}^T \vec{n}/g}$ in (4.4) by a sum over the vectors $\vec{r}$ in the plane $L_{\vec{\nu}}$, each of which only asks for the vanishing of the single integer $\vec{c}\cdot\vec{x}/e$.

The last step. We detect $t = \vec{c}\cdot\vec{x}/e = 0$ by (3.3) with $Q_1 = Q/(ge)$. If $\omega(|\vec{r}|/\varrho_{\vec{\nu}}) \neq 0$ then $e|\vec{c}| < \varrho_{\vec{\nu}}$, so $|t| \leqslant |\vec{c}||\vec{n}|/(ge) < \frac{1}{2}Q_1^2$, and only $q_1 \leqslant Q_1$ occur by Lemma 3.1(1). Moreover $ge \leqslant g|\vec{r}| < Q^{2/3}$, so $Q_1 > Q^{1/3}$ and $c_{Q_1} = 1 + O_N(Q^{-N})$. Collecting the factors

$$\frac{c_3}{Q}\cdot\frac{c_2|\vec{\nu}_0|}{\varrho_{\vec{\nu}}^2}\cdot\frac{1}{Q_1^2} = \frac{c_2c_3|\vec{\nu}|g^3e^2}{Q^{13/3}}, \qquad \frac{c_3}{Q}\cdot\frac{c_2|\vec{\nu}_0|}{\varrho_{\vec{\nu}}^2} = \frac{c_2c_3|\vec{\nu}|g}{Q^{7/3}},$$

we obtain (4.1) and (4.2), where we used $|\vec{x}\times\vec{\nu}_0|/\varrho_{\vec{\nu}} = |\vec{n}\times\vec{\nu}_0|/Q^{2/3}$. There are $O(Q^{7/3})$ pairs $(\vec{\nu},\vec{r})$ with non-zero weight and every term is $O(Q^C)$ for some absolute C, so the relative errors add up to $O_N(Q^{-N})$. □

It is useful to keep in mind the sizes of the three kinds of terms. A point term carries the factor Q^{-1} and there is one for each divisor of $\vec{n}$; a plane term carries $Q^{-7/3}|\vec{\nu}|$, and there are about Q vectors $\vec{\nu}$; a flag term carries $Q^{-13/3}|\vec{\nu}|$ times a one-dimensional symbol, and there are about Q^2 pairs $(\vec{\nu},\vec{r})$ (up to logarithms, for each $\vec{\nu}$ with $\gcd(\vec{\nu}) = 1$ the plane $L_{\vec{\nu}}$ contains about $Q^{4/3}/|\vec{\nu}| \asymp Q$ vectors $\vec{r}$ of length at most $Q^{2/3}$). As a check, at $\vec{n} = \vec{0}$ the one-dimensional symbol contributes about Q_1^2, and $Q^{-13/3}\cdot Q^{1/3}\cdot Q^2\cdot Q^2 = 1$, as it should be. After Fourier inversion, which we carry out next, these three families become the point, plane and flag weights of Theorem 1.2.

Remark 4.3. If $c_3^{-1}Q$, $\varrho_{\vec{\nu}}^2/(c_2|\vec{\nu}_0|)$ and c_{Q_1} are replaced by the exact quantities S_3, $S_2(\vec{\nu})$ and $Q_1(\sum_m \omega(m/Q_1))^{-1}$, the proof gives an identity without error term for every $\vec{n}$ with $|\vec{n}| < \frac{1}{2}Q^{4/3}$. Everything below holds for either version.

4.2. **The weight functions.** Next, we use Fourier inversion to pass from Proposition 4.2 to the form (1.4). The idea is simple. Each term of Proposition 4.2 is a congruence condition on $\vec{n}$ times a smooth function of $\vec{n}$. The congruence condition is written with additive characters $\mathrm{e}(\vec{a}\cdot\vec{n}/q)$, and the smooth function is written as a Fourier integral $\int p(\vec{w})\mathrm{e}(\vec{w}\cdot\vec{n})\,d\vec{w}$. The functions p obtained in this way are the weights. For $q,k\in\mathbb{N}$, a non-zero $\vec{\nu}\in\mathbb{Z}^3$, a flag $(\vec{\nu},\vec{\rho})$ and $\vec{w}\in\mathbb{R}^3$ we define

$$p_{1,q}(\vec{w}) = -\frac{2c_3}{Q}\int_{\mathbb{R}^3} \omega_0\Big(\frac{|\vec{x}|}{Q^{4/3}}\Big)\sum_{j\in\mathbb{N}}\frac{1}{(qj)^3}\,\omega\Big(\frac{|\vec{x}|}{jqQ^{1/3}}\Big)\mathrm{e}(-\vec{w}\cdot\vec{x})\,d\vec{x}, \tag{4.6}$$

$$p_{2,\vec{\nu},k,q}(\vec{w}) = -\frac{2c_2c_3|\vec{\nu}|}{Q^{7/3}}\int_{\mathbb{R}^3} \omega_0\Big(\frac{|\vec{x}|}{Q^{4/3}}\Big)\sum_{j\in\mathbb{N}}\frac{1}{(qkj)^2}\,\omega\Big(\frac{|\vec{x}\times\vec{\nu}|}{qkjQ^{2/3}}\Big)\mathrm{e}(-\vec{w}\cdot\vec{x})\,d\vec{x}, \tag{4.7}$$

$$p_{3,\vec{\nu},\vec{\rho},k,q}(\vec{w}) = \frac{c_2c_3|\vec{\nu}|}{Q^{13/3}}\int_{\mathbb{R}^3} \omega_0\Big(\frac{|\vec{x}|}{Q^{4/3}}\Big)h\Big(\frac{kq}{Q},\frac{\vec{\rho}\cdot\vec{x}}{Q^2}\Big)\mathrm{e}(-\vec{w}\cdot\vec{x})\,d\vec{x}. \tag{4.8}$$

Each integrand is a smooth function of $\vec{x}$ with support in $|\vec{x}| < \frac{1}{2}Q^{4/3}$: the function $\omega(|\vec{x} \times \vec{\nu}|/(\cdots))$ vanishes near the line $\mathbb{R}\vec{\nu}$, and $h(y,z)$ does not depend on z for $|z| < y/2$ by Lemma 3.1(1). So each p-function is a Schwartz function. In the next three subsections we prove the bounds (1.7)–(1.9). All three proofs follow the same pattern: we change variables so that the amplitude becomes a fixed smooth function with bounded derivatives, and then integrate by parts in the directions in which the phase $\mathrm{e}(-\vec{w}\cdot\vec{x})$ oscillates. The scale of the amplitude in each direction determines the size of the region in $\vec{w}$ where the weight is not negligible. For the point weight the amplitude lives at scale $qQ^{1/3}$ in every direction, so $p_{1,q}$ is concentrated on a ball of radius $(qQ^{1/3})^{-1}$; for the plane and flag weights one direction has the large scale $Q^{4/3}$, which produces the thickness $Q^{-4/3}$ in the table of §2. Throughout them $1 \leqslant q \leqslant Q$, and we write $T = Q^{4/3}$.

4.3. **Bounds for the point weights $p_{1,q}$.**

Lemma 4.4. *The bound (1.7) holds.*

Proof. The j-th term of (4.6) is a bump of radius about $jqQ^{1/3}$ in $\vec{x}$, so its Fourier transform is concentrated on $|\vec{w}| \lesssim (jqQ^{1/3})^{-1}$; summing over j produces the factor Q/q and the decay in $|\vec{w}|qQ^{1/3}$. Substituting $\vec{x} = jqQ^{1/3}\vec{y}$ in (4.6) gives

$$p_{1,q}(\vec{w}) = -2c_3 \sum_{j\in\mathbb{N}} \int_{\mathbb{R}^3} \omega_0\Big(\frac{jq|\vec{y}|}{Q}\Big)\omega(|\vec{y}|)\,\mathrm{e}(-jqQ^{1/3}\vec{w}\cdot\vec{y})\,d\vec{y}.$$

Since $|\vec{y}| > \frac{1}{2}$ on the support of ω, only $j < Q/q$ occur. The amplitude has bounded derivatives, uniformly for $j < Q/q$, so integration by parts bounds the j-th integral by $\ll_N (1+jqQ^{1/3}|\vec{w}|)^{-N}$. Summing over $j < Q/q$ gives (1.7). (The factor $(1+|\vec{w}|Q^{4/3})^{-1}$ comes from the sum over j: the j-th integral is negligible unless $j \lesssim (qQ^{1/3}|\vec{w}|)^{-1}$, so only $\ll 1 + \min(Q/q, (qQ^{1/3}|\vec{w}|)^{-1})$ values of j contribute.) □

4.4. **Bounds for the plane weights $p_{2,\vec{\nu},k,q}$.**

Lemma 4.5. *Let $\frac{1}{2}Q^{1/3} < |\vec{\nu}| < Q^{1/3}$. Then $p_{2,\vec{\nu},k,q} = 0$ unless $kq < Q$, and the bound (1.8) holds.*

Proof. Here the amplitude has two different scales: it is spread over the whole ball of radius $T = Q^{4/3}$ in the direction of $\vec{\nu}$, but only over a distance $\lambda_j \approx qkjQ^{1/3}$ in the directions of the plane $\vec{\nu}^{\perp}$. We therefore use coordinates adapted to this splitting. Put $\lambda_j = qkjQ^{2/3}/|\vec{\nu}|$ and write $\vec{x} = Ts\hat{\vec{\nu}} + \lambda_j\vec{y}$ with $s \in \mathbb{R}$ and $\vec{y} \in \vec{\nu}^{\perp}$, so that $|\vec{x}\times\vec{\nu}| = \lambda_j|\vec{\nu}||\vec{y}|$. Then

$$p_{2,\vec{\nu},k,q}(\vec{w}) = -\frac{2c_2c_3Q^{1/3}}{|\vec{\nu}|}\sum_{j\in\mathbb{N}}\int_{\mathbb{R}}\int_{\vec{\nu}^{\perp}} \omega_0\Big(\sqrt{s^2+\theta_j^2|\vec{y}|^2}\Big)\omega(|\vec{y}|)\,\mathrm{e}\Big(-Ts\,\vec{w}\cdot\hat{\vec{\nu}} - \lambda_j\vec{w}\cdot\vec{y}\Big)\,d\vec{y}\,ds,$$

with $\theta_j = \lambda_j/T$. The integrand vanishes unless $\theta_j < 1$, that is, $qkj < Q^{2/3}|\vec{\nu}| < Q$; in particular $p_{2,\vec{\nu},k,q} = 0$ if $kq \geqslant Q$. The amplitude has bounded derivatives, and $\vec{w}\cdot\vec{y}$ only involves the component of $\vec{w}$ in $\vec{\nu}^{\perp}$, which has length $|\vec{w}\times\hat{\vec{\nu}}|$. Hence the j-th integral is $\ll_N (1+T|\vec{w}\cdot\hat{\vec{\nu}}|)^{-N}(1+qkjQ^{1/3}|\vec{w}\times\hat{\vec{\nu}}|)^{-N}$, and $Q^{1/3}/|\vec{\nu}| \ll 1$. Summing over $j < Q/(kq)$ gives (1.8). □

4.5. **Bounds for the flag weights $p_{3,\vec\nu,\vec\rho,k,q}$.**

Lemma 4.6. *Let $\frac12 Q^{1/3} < |\vec\nu| < Q^{1/3}$ and $\frac12 Q^{2/3} < |\vec\rho| < Q^{2/3}$. Then $p_{3,\vec\nu,\vec\rho,k,q} = 0$ unless $kq < Q$, and the bound (1.9) holds.*

Proof. Now the roles of the two scales are reversed: the function h only depends on $\vec\rho\cdot\vec x$, and it varies on the scale $kqQ/|\vec\rho| \approx kqQ^{1/3}$ in the direction $\hat{\vec\rho}$, while in the plane $\vec\rho^{\perp}$ the only cut-off is ω_0 at scale T. The new difficulty is that h is not a fixed smooth function, but Lemma 3.1(4) bounds the L^1 norms of its derivatives, which is all that integration by parts needs. Put $y = kq/Q$. Write $\vec x = z\hat{\vec\rho} + \vec x'$ with $\vec x' \in \vec\rho^{\perp}$ and substitute $z = Q^2Z/|\vec\rho|$:

$$p_{3,\vec\nu,\vec\rho,k,q}(\vec w) = \frac{c_2c_3|\vec\nu|}{Q^{7/3}|\vec\rho|}\int_{\vec\rho^{\perp}}\int_{\mathbb R}\omega_0\Big(\frac{\sqrt{(Q^2Z/|\vec\rho|)^2+|\vec x'|^2}}{T}\Big)h(y,Z) \\ \times \mathrm e\Big(-\frac{Q^2Z\,\vec w\cdot\hat{\vec\rho}}{|\vec\rho|} - \vec w\cdot\vec x'\Big)dZ\,d\vec x'.$$

On the support of ω_0 we have $|Z| \leqslant |\vec\rho|T/(2Q^2) < \frac12$, so $h(y,Z) = 0$ if $y \geqslant 1$, by Lemma 3.1(1). Suppose that $y < 1$. Since $y \geqslant Q^{-1}$, Lemma 3.1(4) gives

$$\int_{|Z|\leqslant 1/2}|\partial_Z^i h(y,Z)|\,dZ \ll Q^{\varepsilon}y^{-i}$$

for $i \geqslant 0$. The derivatives of the amplitude are $\ll 1$ in Z, because $Q^2/(|\vec\rho|T) < 2$, and $\ll T^{-1}$ in $\vec x'$. Integrating by parts i times in Z and i' times in $\vec x'$ we obtain

$$p_{3,\vec\nu,\vec\rho,k,q}(\vec w) \ll \frac{|\vec\nu|T^2}{Q^{7/3}|\vec\rho|}\,Q^{\varepsilon}\Big(\frac{|\vec\rho|}{Q^2y|\vec w\cdot\hat{\vec\rho}|}\Big)^{i}\big(T|\vec w\times\hat{\vec\rho}|\big)^{-i'}.$$

The first factor is $\ll |\vec\nu|Q^{1/3}/|\vec\rho| \ll 1$, and $Q^2y/|\vec\rho| = kqQ/|\vec\rho| > kqQ^{1/3}$. Taking $i, i' \in \{0, N\}$ gives (1.9). □

5. Proof of the identity

In this section we prove Theorem 1.2. The identity of Proposition 4.2 is exact but its terms are written with periodic functions of $\vec n$; to obtain (1.4) we must write each periodic function as a sum over fractions $\vec a/q$ in lowest terms and check that the resulting weight depends only on the lattice $\Lambda(\vec a, q)$. This is done in the first two subsections, with elementary lattice arguments. In §5.4 we prove the main-term property, and at the end of the section the mass bound.

5.1. **Lattices attached to a fraction.** To prove Theorem 1.2 we need some properties of the lattices $\Lambda(\vec a, q)$ and $\Lambda^{\perp}(\vec a, q) = \{\vec\nu \in \mathbb Z^3 :\ q \mid \vec\nu\cdot\vec a\}$.

Geometrically, $\Lambda(\vec a, q)$ consists of the integer vectors $\vec\rho$ which are congruent modulo q to a multiple of $\vec a$; when the multiple is coprime to q, the rational line $\mathbb R\vec\rho$ passes through the point $\vec a/q$ of the torus $\mathbb R^3/\mathbb Z^3$. Similarly $\Lambda^{\perp}(\vec a, q)$ consists of the normal vectors $\vec\nu$ for which $\vec a/q$ lies on one of the parallel planes $\vec\nu\cdot\vec x \in \mathbb Z$. Part (iii) is the reason why the weights are invariant under $\vec a \mapsto \lambda\vec a$, which is the basis of the Kloosterman refinement.

Lemma 5.1. *Let $q \in \mathbb N$ and $\vec a \in \mathbb Z^3$ with $\gcd(\vec a, q) = 1$.*

(i) $\Lambda(\vec a, q) = \{\vec\rho \in \mathbb Z^3 :\ q \mid \vec\rho\times\vec a\}$, and $\Lambda(\vec a, q)$ has covolume q^2.

(ii) $\Lambda^{\perp}(\vec a, q) = q\,\Lambda(\vec a, q)^*$, and $\Lambda^{\perp}(\vec a, q)$ has covolume q.

(iii) $\Lambda(\vec{a},q)=\Lambda(\vec{b},q)$ *if and only if* $\vec{b}\equiv\lambda\vec{a}\pmod q$ *for some* λ *with* $\gcd(\lambda,q)=1$.

Proof. Write $\vec{a}=t\vec{a}_0$ with $\vec{a}_0$ primitive. Everything is reduced to the case where $\vec{a}$ is a basis vector, by a unimodular change of coordinates. Then $\gcd(t,q)=1$, $\Lambda(\vec{a},q)=\Lambda(\vec{a}_0,q)$, and the conditions in (i) and (ii) do not change if $\vec{a}$ is replaced by $\vec{a}_0$. Complete $\vec{a}_0$ to a basis $\vec{a}_0,\vec{a}',\vec{a}''$ of $\mathbb{Z}^3$. Then $\Lambda(\vec{a}_0,q)=\mathbb{Z}\vec{a}_0+q\mathbb{Z}\vec{a}'+q\mathbb{Z}\vec{a}''$, which has covolume q^2. Since $\vec{a}_0\times\vec{a}'$ and $\vec{a}_0\times\vec{a}''$ form a basis of the primitive lattice $\vec{a}_0^{\perp}\cap\mathbb{Z}^3$, the cross product of $x\vec{a}_0+y\vec{a}'+z\vec{a}''$ with $\vec{a}_0$ is divisible by q if and only if $q\mid y$ and $q\mid z$. This proves (i). Part (ii) is the definition of the dual lattice. For (iii), suppose first that $\vec{b}\equiv\lambda\vec{a}\pmod q$ with $\gcd(\lambda,q)=1$. Then $\vec{b}\in\Lambda(\vec{a},q)$, so $\Lambda(\vec{b},q)\subseteq\Lambda(\vec{a},q)$; choosing λ' with $\lambda\lambda'\equiv1\pmod q$ we have $\vec{a}\equiv\lambda'\vec{b}$, and the reverse inclusion follows in the same way. Conversely, if $\Lambda(\vec{b},q)=\Lambda(\vec{a},q)$, then $\vec{b}\in\Lambda(\vec{a},q)$, so $\vec{b}=k\vec{a}+q\vec{y}$ for some $k\in\mathbb{Z}$ and $\vec{y}\in\mathbb{Z}^3$. If a prime p divided both k and q, then $p\mid\vec{b}$, contradicting $\gcd(\vec{b},q)=1$. So $\gcd(k,q)=1$, and $\vec{b}\equiv k\vec{a}\pmod q$. (This is [LRV24, Lemma 4.1] in three dimensions.) □

The periodic parts of the terms in Proposition 4.2 are characters of finite abelian groups. We now write them in lowest terms. The next two lemmas are pure bookkeeping, but they are the heart of the matter: they show that, after expanding the congruence conditions of a flag term and of a plane term into additive characters and reducing each fraction to lowest terms, the flag $(\vec{\nu},\vec{\rho})$ is attached to the fraction $\vec{a}/q$ exactly when $\vec{\nu}\in\Lambda^{\perp}(\vec{a},q)$ and $\vec{\rho}\in\Lambda(\vec{a},q)$. In other words, a flag "belongs" to the fractions lying on the line $\mathbb{R}\vec{\rho}$ and on the plane $\vec{\nu}^{\perp}$, modulo $\mathbb{Z}^3$.

Lemma 5.2. *Let* $\vec{\nu}=g\vec{\nu}_0$ *with* $\vec{\nu}_0$ *primitive, let* $\vec{r}=e\vec{c}\in L_{\vec{\nu}}$ *with* $\vec{c}$ *primitive, and put* $G=ge$ *and* $\vec{\rho}=g\vec{r}=G\vec{c}$. *Then for* $q_1\in\mathbb{N}$ *and* $\vec{n}\in\mathbb{Z}^3$

$$g^3e^2\,\mathbf{1}_{g\mid\vec{n}}\,\mathbf{1}_{e\mid B_{\vec{\nu}}^T\vec{n}/g}\sideset{}{^*}\sum_{a\bmod q_1}\mathrm{e}_{q_1}\Big(\frac{a\,\vec{c}\cdot\vec{n}}{G}\Big)=\sum_{\substack{q\in\mathbb{N}\\ q/\gcd(q,G)=q_1}}\ \sideset{}{^*}\sum_{\substack{\vec{a}\bmod q\\ \vec{\rho}\in\Lambda(\vec{a},q),\ \vec{\nu}\in\Lambda^{\perp}(\vec{a},q)}}\mathrm{e}_q(\vec{a}\cdot\vec{n}).$$

Proof. The left side is a sum of characters over a finite group, namely the indicator function of a lattice of $\vec{n}$'s times a Ramanujan-type sum. We write it as one sum of characters $\mathrm{e}(\vec{A}\cdot\vec{n}/(Gq_1))$, identify the set of frequencies $\vec{A}$ by counting, and then reduce each frequency to lowest terms. The set $\{\vec{n}:\ g\mid\vec{n},\ e\mid B_{\vec{\nu}}^T\vec{n}/g\}=g(\mathbb{Z}\vec{\nu}_0+e\mathbb{Z}^3)$ is a lattice of index g^3e^2, and its dual lattice is $G^{-1}\{\vec{y}\in\mathbb{Z}^3:\ e\mid\vec{\nu}_0\cdot\vec{y}\}$. Hence the left side equals $\sum_{\vec{A}}\mathrm{e}_{Gq_1}(\vec{A}\cdot\vec{n})$, where $\vec{A}=a\vec{c}+q_1\vec{y}$ with $\gcd(a,q_1)=1$ and $\vec{y}$ running modulo G over the solutions of $e\mid\vec{\nu}_0\cdot\vec{y}$. These $\varphi(q_1)g^3e^2$ vectors are distinct modulo Gq_1, and they satisfy

$$\vec{A}\equiv a\vec{c}\pmod{q_1}\text{ for some }a\text{ with }\gcd(a,q_1)=1,\qquad\vec{\nu}_0\cdot\vec{A}\equiv0\pmod{q_1e}.$$

The number of classes modulo Gq_1 with these two properties is also $\varphi(q_1)g^3e^2$, so the $\vec{A}$ are exactly those classes. Write $\vec{A}/(Gq_1)=\vec{a}/q$ in lowest terms, so that $\vec{A}=j\vec{a}$ with $jq=Gq_1$. The first property and the primitivity of $\vec{c}$ give $\gcd(\vec{A},q_1)=1$, hence $\gcd(j,q_1)=1$, $j\mid G$, and $q=dq_1$ with $d=G/j=\gcd(q,G)$. The first property becomes $q_1\mid\vec{c}\times\vec{a}$, which is equivalent to $q\mid\vec{\rho}\times\vec{a}$ because $\vec{\rho}=dj\vec{c}$ and $\gcd(j,q_1)=1$; by Lemma 5.1(i) this says $\vec{\rho}\in\Lambda(\vec{a},q)$. Multiplying the second property by g gives $q_1dj\mid j\,\vec{\nu}\cdot\vec{a}$, that is, $q\mid\vec{\nu}\cdot\vec{a}$. The converse is proved by reversing these steps. □

Lemma 5.3. *Let $\vec{\nu} = g\vec{\nu}_0$ with $\vec{\nu}_0$ primitive. For $e' \in \mathbb{N}$ and $\vec{n} \in \mathbb{Z}^3$*

$$g^3 e'^2\, \mathbf{1}_{g|\vec{n}}\, \mathbf{1}_{e'|B_{\vec{\nu}}^T \vec{n}/g} = \sum_{q | ge'} \sideset{}{^*}\sum_{\substack{\vec{a} \bmod q \\ \vec{\nu} \in \Lambda^\perp(\vec{a},q)}} \mathrm{e}_q(\vec{a}\cdot\vec{n}).$$

For a given q, the e' with $q \mid ge'$ are those with $ge' = qkj$ for some $j \in \mathbb{N}$, where $k = g/\gcd(g,q)$.

Proof. As in the previous proof, the left side equals $\sum_{\vec{y}} \mathrm{e}_{ge'}(\vec{y}\cdot\vec{n})$ over $\vec{y}$ modulo ge' with $e' \mid \vec{\nu}_0 \cdot \vec{y}$. Writing $\vec{y} = j\vec{a}$ in lowest terms with $jq = ge'$, the condition $e' \mid j\vec{\nu}_0 \cdot \vec{a}$ is equivalent to $ge' \mid j\,\vec{\nu}\cdot\vec{a}$, that is, to $q \mid \vec{\nu}\cdot\vec{a}$. □

5.2. **The weight attached to a fraction.** We are now ready to state the formula for $p_{\Lambda(\vec{a},q)}$. It is obtained by collecting, for each reduced fraction $\vec{a}/q$, all the contributions which Lemmas 5.2 and 5.3 attach to it. For $q \in \mathbb{N}$ and $\vec{a} \in \mathbb{Z}^3$ with $\gcd(\vec{a}, q) = 1$ we define

$$\begin{aligned} p_{\Lambda(\vec{a},q)}(\vec{w}) = p_{1,q}(\vec{w}) &+ \sum_{\substack{\vec{\nu} \in \Lambda^\perp(\vec{a},q) \\ k = \gcd(\vec{\nu})/\gcd(\vec{\nu},q)}} \omega\Big(\frac{|\vec{\nu}|}{Q^{1/3}}\Big) p_{2,\vec{\nu},k,q}(\vec{w}) \\ &+ \sum_{\substack{(\vec{\nu},\vec{\rho})\ \text{flag},\ \vec{\nu}\in\Lambda^\perp(\vec{a},q),\ \vec{\rho}\in\Lambda(\vec{a},q) \\ k = \gcd(\vec{\rho})/\gcd(\vec{\rho},q)}} \omega\Big(\frac{|\vec{\nu}|}{Q^{1/3}}\Big)\omega\Big(\frac{|\vec{\rho}|}{Q^{2/3}}\Big) p_{3,\vec{\nu},\vec{\rho},k,q}(\vec{w}), \end{aligned} \tag{5.1}$$

with the p-functions of (4.6)–(4.8). By Lemma 5.1, the conditions $\vec{\nu} \in \Lambda^\perp(\vec{a},q)$ and $\vec{\rho} \in \Lambda(\vec{a},q)$ depend only on the lattice $\Lambda(\vec{a},q)$, and so does $p_{\Lambda(\vec{a},q)}$. The sums are finite. The next lemma shows that the weight attached to $\vec{a}/q$ is concentrated on the ball of radius $r_q = (qQ^{1/3})^{-1}$ around $\vec{a}/q$, up to the harmless factor Q^ε. This is the three-dimensional analogue of the arcs of the classical circle method, and r_q is exactly the radius suggested by Dirichlet's theorem in $\mathbb{R}^3$.

Lemma 5.4. *Let $1 \leqslant q \leqslant Q$ and $\gcd(\vec{a},q) = 1$, and put $r_q = (qQ^{1/3})^{-1}$. Then $p_{\Lambda(\vec{a},q)}(\vec{w}) \ll Q^4$, and for any $\varepsilon, N > 0$*

$$p_{\Lambda(\vec{a},q)}(\vec{w}) \ll_{\varepsilon,N} Q^{-N}(1+|\vec{w}|)^{-N} \qquad \text{if } |\vec{w}| \geqslant Q^\varepsilon r_q.$$

Proof. Each of the three kinds of weight decays rapidly once $\vec{w}$ leaves a box around $\vec{a}/q$ whose longest side is at most r_q. So we only have to check, for each kind, that a vector $\vec{w}$ of length at least $Q^{3\varepsilon} r_q$ is far outside the box in at least one direction. The sums in (5.1) have $O(Q^{7/3})$ terms, and each is $\ll Q^{1+\varepsilon}$ by Lemmas 4.4–4.6. Let $|\vec{w}| \geqslant Q^{3\varepsilon} r_q$. For $p_{1,q}$ the factor $(1 + |\vec{w}| qQ^{1/3})^{-2N}$ is at most $Q^{-3\varepsilon N}(1+|\vec{w}|)^{-N}$, as $qQ^{1/3} \geqslant 1$. For $p_{2,\vec{\nu},k,q}$, either $|\vec{w} \times \hat{\vec{\nu}}| \geqslant |\vec{w}|/2$, and then $|\vec{w} \times \hat{\vec{\nu}}| kqQ^{1/3} \geqslant Q^{3\varepsilon}/2$, or

$$|\vec{w} \cdot \hat{\vec{\nu}}| \geqslant |\vec{w}|/2 \geqslant Q^{3\varepsilon} Q^{-4/3}.$$

The same alternative holds for $p_{3,\vec{\nu},\vec{\rho},k,q}$ with $\hat{\vec{\rho}}$ in place of $\hat{\vec{\nu}}$, and the factor Q^ε in (1.9) is absorbed. Replacing 3ε by ε gives the lemma. □

5.3. **Assembling the identity.** Since $\omega_0(0) = 1$ we have $\delta_{\vec{n}} = \omega_0(|\vec{n}|/Q^{4/3})\delta_{\vec{n}}$, and $\omega_0(|\vec{n}|/Q^{4/3}) = 0$ when $|\vec{n}| \geqslant \frac{1}{2}Q^{4/3}$. We multiply Proposition 4.2 by $\omega_0(|\vec{n}|/Q^{4/3})$; for $|\vec{n}| \geqslant \frac{1}{2}Q^{4/3}$ both sides of the resulting identity vanish. We treat the three terms in turn. In each case the plan is the same: Lemma 5.2 or 5.3 rewrites the congruence conditions as a sum over reduced fractions $\vec{a}/q$, the smooth factor which remains is a function of $\vec{n}$ supported in $|\vec{n}| < \frac{1}{2}Q^{4/3}$, and Fourier inversion turns it into the

corresponding p-function. The only thing to check is that no modulus larger than Q appears.

The term (4.1). Replace $\vec{r}$ by $\vec{\rho} = g\vec{r}$. Then $(\vec{\nu}, \vec{\rho})$ runs over the flags, $\omega(|\vec{r}|/\varrho_{\vec{\nu}}) = \omega(|\vec{\rho}|/Q^{2/3})$, and $\gcd(\vec{\rho}) = ge = G$. By Lemma 5.2, and since $Gq_1 = kq$ with $k = G/\gcd(q, G)$ when $q = \gcd(q, G)q_1$, the term equals

$$\sum_{q\in\mathbb{N}} \sideset{}{^*}\sum_{\vec{a} \bmod q} \mathrm{e}_q(\vec{a}\cdot\vec{n}) \sum_{\substack{(\vec{\nu},\vec{\rho}) \text{ flag} \\ \vec{\nu}\in\Lambda^\perp(\vec{a},q),\ \vec{\rho}\in\Lambda(\vec{a},q)}} \omega\Big(\frac{|\vec{\nu}|}{Q^{1/3}}\Big)\omega\Big(\frac{|\vec{\rho}|}{Q^{2/3}}\Big)\frac{c_2c_3|\vec{\nu}|}{Q^{13/3}}\,\omega_0\Big(\frac{|\vec{n}|}{Q^{4/3}}\Big)h\Big(\frac{kq}{Q}, \frac{\vec{\rho}\cdot\vec{n}}{Q^2}\Big).$$

The function $\vec{x} \mapsto \omega_0(|\vec{x}|/Q^{4/3})h(kq/Q, \vec{\rho}\cdot\vec{x}/Q^2)$ vanishes identically when $kq \geqslant Q$, since $|\vec{\rho}\cdot\vec{x}|/Q^2 < \frac{1}{2}$ on the support of ω_0. So only $q \leqslant Q$ occur, and Fourier inversion in $\vec{n}$ gives the flag part of (5.1).

The term (4.2). By Lemma 5.3, writing $ge' = qkj$, the coefficient of $\mathrm{e}_q(\vec{a}\cdot\vec{n})$ for $\vec{\nu} \in \Lambda^\perp(\vec{a}, q)$ is

$$-\frac{2c_2c_3|\vec{\nu}|g}{Q^{7/3}}\sum_{j\in\mathbb{N}}\frac{1}{g^3e'^2}\,\omega\Big(\frac{|\vec{n}\times\vec{\nu}_0|}{e'Q^{2/3}}\Big) = -\frac{2c_2c_3|\vec{\nu}|}{Q^{7/3}}\sum_{j\in\mathbb{N}}\frac{1}{(qkj)^2}\,\omega\Big(\frac{|\vec{n}\times\vec{\nu}|}{qkjQ^{2/3}}\Big).$$

On the supports of $\omega_0(|\vec{n}|/Q^{4/3})$ and ω we have $qkj < 2|\vec{n}||\vec{\nu}|/Q^{2/3} < Q$. Fourier inversion gives the plane part of (5.1).

The term (4.3). We write

$$\mathbf{1}_{g'|\vec{n}} = g'^{-3}\sum_{q|g'\vec{a}} \sideset{}{^*}\sum_{\bmod q} \mathrm{e}_q(\vec{a}\cdot\vec{n})$$

and $g' = qj$. On the supports $qj < 2|\vec{n}|/Q^{1/3} < Q$, and Fourier inversion gives $p_{1,q}$.

This proves (1.4) and (1.6). The bounds (1.7)–(1.9) are Lemmas 4.4–4.6. It remains to prove (1.5).

5.4. **The main-term property.** For $\gcd(\vec{a}, q) = 1$ we let

$$\Phi_{\vec{a},q}(\vec{y}) = \int_{\mathbb{R}^3} p_{\Lambda(\vec{a},q)}(\vec{w})\,\mathrm{e}(\vec{w}\cdot\vec{y})\,d\vec{w} \qquad (\vec{y}\in\mathbb{R}^3).$$

By the proof above, $\Phi_{\vec{a},q}$ is a smooth function supported in $|\vec{y}| < \frac{1}{2}Q^{4/3}$ whose Fourier transform is $p_{\Lambda(\vec{a},q)}$, and (1.4) can be written as

$$\delta_{\vec{n}} = \sum_{q\leqslant Q} \sideset{}{^*}\sum_{\vec{a} \bmod q} \mathrm{e}_q(\vec{a}\cdot\vec{n})\,\Phi_{\vec{a},q}(\vec{n}) + O_N(Q^{-N}), \tag{5.2}$$

where the error term vanishes when $|\vec{n}| \geqslant \frac{1}{2}Q^{4/3}$.

The main-term property (1.5) says that the weight of a fraction with a small denominator behaves like a Dirac mass when it is tested against functions which vary on the scale $Q^{-4/3}$. Before giving the proof, we explain why this should follow from the identity alone. Apply (5.2) to the function $\vec{n} \mapsto g(\vec{n}/T)\mathrm{e}(-\vec{a}^*\cdot\vec{n}/q^*)$ and sum over $\vec{n}$. On the left we get $g(\vec{0})$. On the right, by Poisson summation, the term of $\vec{a}/q$ is essentially the Fourier transform of $\Phi_{\vec{a},q}g(\cdot/T)$ evaluated at the distance from $\vec{a}/q$ to $\vec{a}^*/q^*$ modulo $\mathbb{Z}^3$. Since distinct fractions with denominators at most Q and q^* are at distance at least $1/(qq^*)$, which is much larger than r_q when q^* is small, all these terms are negligible except the one with $\vec{a}/q = \vec{a}^*/q^*$, and that term is exactly the left side of (1.5). The weights therefore "know" that they must integrate to 1 against such test functions, without any explicit computation.

Proposition 5.5. *The property* (1.5) *holds.*

Proof. Write $T = Q^{4/3}$. Fix $q^* < Q^{1/3-\delta}$ and $\vec{a}^*$ with $\gcd(\vec{a}^*, q^*) = 1$, and let g be a Schwartz function on $\mathbb{R}^3$. We multiply (5.2) by $g(\vec{n}/T)\,\mathrm{e}(-\vec{a}^* \cdot \vec{n}/q^*)$ and sum over $\vec{n} \in \mathbb{Z}^3$. On the left we obtain $g(\vec{0})$, and the error term contributes $O_N(Q^{4-N})$. On the right we obtain, for each pair $(\vec{a}, q)$, the sum

$$\Sigma(\vec{a}, q) = \sum_{\vec{n} \in \mathbb{Z}^3} \mathrm{e}(\vec{\theta} \cdot \vec{n})\, \Psi_{\vec{a},q}(\vec{n}), \qquad \vec{\theta} = \frac{\vec{a}}{q} - \frac{\vec{a}^*}{q^*}, \qquad \Psi_{\vec{a},q}(\vec{y}) = \Phi_{\vec{a},q}(\vec{y})\, g(\vec{y}/T).$$

The function $\Psi_{\vec{a},q}$ is smooth with compact support, and its Fourier transform is $\hat{\Psi}_{\vec{a},q} = p_{\Lambda(\vec{a},q)} * G_T$, where $G_T(\vec{\xi}) = T^3 \hat{g}(T\vec{\xi})$. By Poisson summation, $\Sigma(\vec{a}, q) = \sum_{\vec{k} \in \mathbb{Z}^3} \hat{\Psi}_{\vec{a},q}(\vec{k} - \vec{\theta})$.

We claim that $\hat{\Psi}_{\vec{a},q}(\vec{\xi}) \ll_N Q^{-N}(1 + |\vec{\xi}|)^{-N}$ whenever $|\vec{\xi}| \geqslant \rho_0 = 2Q^{\delta/2} r_q$. To see this, split

$$\hat{\Psi}_{\vec{a},q}(\vec{\xi}) = \int p_{\Lambda(\vec{a},q)}(\vec{\xi} - \vec{\eta}) G_T(\vec{\eta})\, d\vec{\eta}$$

according to whether $|\vec{\eta}| \geqslant |\vec{\xi}|/2$ or $|\vec{\xi} - \vec{\eta}| \geqslant |\vec{\xi}|/2$. In the first range $G_T(\vec{\eta}) \ll_N T^3 (T|\vec{\eta}|)^{-2N}$ with

$$T|\vec{\eta}| \geqslant T\rho_0/2 \geqslant Q^{1+\delta/2}/q \geqslant Q^{\delta/2},$$

while $p_{\Lambda(\vec{a},q)} \ll Q^4$. In the second range Lemma 5.4 gives $p_{\Lambda(\vec{a},q)}(\vec{\xi} - \vec{\eta}) \ll_N Q^{-N}(1 + |\vec{\xi}|)^{-N}$, and $\int |G_T| \ll 1$.

Now consider the pairs $(\vec{a}, q)$ with $q \leqslant Q$ and $\vec{a}$ reduced modulo q. If $\vec{a}/q \not\equiv \vec{a}^*/q^* \pmod{\mathbb{Z}^3}$, then $qq^*\vec{\theta} \in \mathbb{Z}^3 \setminus qq^*\mathbb{Z}^3$, so $|\vec{k} - \vec{\theta}| \geqslant 1/(qq^*)$ for every $\vec{k}$. Since $q^* < Q^{1/3-\delta}$ we have $1/(qq^*) > Q^{\delta} r_q \geqslant \rho_0$, so every term of $\Sigma(\vec{a}, q)$ is covered by the claim, and $\Sigma(\vec{a}, q) \ll_N Q^{-N}$. If $\vec{a}/q \equiv \vec{a}^*/q^*$, then $q = q^*$ and $\vec{a} \equiv \vec{a}^* \pmod{q^*}$, so this pair occurs exactly once; then $\vec{\theta} \in \mathbb{Z}^3$, and the terms with $\vec{k} \neq \vec{\theta}$ satisfy $|\vec{k} - \vec{\theta}| \geqslant 1 \geqslant \rho_0$ and are negligible by the claim. Since there are at most Q^4 pairs $(\vec{a}, q)$, we conclude that

$$g(\vec{0}) = \hat{\Psi}_{\vec{a}^*,q^*}(\vec{0}) + O_N(Q^{-N}).$$

Finally

$$\hat{\Psi}_{\vec{a}^*,q^*}(\vec{0}) = \int \Phi_{\vec{a}^*,q^*}(\vec{y}) g(\vec{y}/T)\, d\vec{y} = \int p_{\Lambda(\vec{a}^*,q^*)}(\vec{w})\, T^3 \hat{g}(-T\vec{w})\, d\vec{w}.$$

□

Remark 5.6. The proof uses only that the identity holds for all $\vec{n}$ with an error independent of $\vec{n}$, that the weight of $\vec{a}/q$ is concentrated within $Q^{\varepsilon}(qQ^{1/R})^{-1}$ of $\vec{a}/q$, where R is the dimension, and that it is polynomially bounded. Two distinct fractions with denominators $q \leqslant Q$ and q^* are at distance at least $1/(qq^*)$, so the argument gives the main-term property for $q^* < Q^{1/R-\delta}$, tested against functions at scale $Q^{-1-1/R}$. For $R = 1$ and the symbol (3.3) this is the statement $\int h(y, z) f(z)\, dz = f(0) + O_N(y^N)$ for $y \leqslant Q^{-\delta}$ of [HB96, Lemma 9], compare [LRV24, Lemma 2.1(2)]. For $R = 2$ and the symbol of [LRV24] it is a weak form of [LRV24, (1.6)].

5.5. **The mass of the weights.** We now prove Proposition 1.3. The mass is the quantity that measures how much the trivial bound loses: every error term in the paper is first bounded by the size of the integrand times the mass of the weights in the range considered. The proof is a matter of counting how many residues $\vec{a}$ and how many flags are attached to each modulus, and multiplying by the integral of a single weight, which we read off from (1.7)–(1.9). We first count the residues and the flags which occur.

Lemma 5.7. *Let $q \in \mathbb{N}$.*

(i) For a non-zero $\vec{\nu} \in \mathbb{Z}^3$ with $g = \gcd(\vec{\nu})$, the number of $\vec{a} \bmod q$ with $q \mid \vec{\nu} \cdot \vec{a}$ is $q^2 \gcd(q, g)$.

(ii) Let $\vec{\rho} = G\vec{c}$ with $\vec{c}$ primitive and $d = \gcd(q, G)$. The residues $\vec{a} \bmod q$ with $\gcd(\vec{a}, q) = 1$ and $\vec{\rho} \in \Lambda(\vec{a}, q)$ are those with $\vec{a} \equiv \lambda\vec{c} \pmod{q/d}$ for some λ coprime to q/d. There are at most qd^2 of them.

Proof. (i) The condition is $q/\gcd(q, g) \mid \vec{\nu}_0 \cdot \vec{a}$, and $\vec{a} \mapsto \vec{\nu}_0 \cdot \vec{a}$ maps $(\mathbb{Z}/q\mathbb{Z})^3$ onto $\mathbb{Z}/q\mathbb{Z}$. (ii) By Lemma 5.1(i), $\vec{\rho} \in \Lambda(\vec{a}, q)$ means $q \mid G\,\vec{c} \times \vec{a}$, that is, $(q/d) \mid \vec{c} \times \vec{a}$, since $\gcd(q/d, G/d) = 1$. As $\gcd(\vec{a}, q/d) = 1$, this says $\vec{c} \in \Lambda(\vec{a}, q/d)$, hence $\Lambda(\vec{c}, q/d) = \Lambda(\vec{a}, q/d)$, and Lemma 5.1(iii) applies. There are at most q/d choices of λ and d^3 lifts modulo q. □

Lemma 5.8. *For $G \in \mathbb{N}$ and $U, V \geqslant 1$, the number of flags $(\vec{\nu}, \vec{\rho})$ with $|\vec{\nu}| \leqslant U$, $|\vec{\rho}| \leqslant V$ and $\gcd(\vec{\rho}) = G$ is*

$$\ll \frac{U^2V^2}{G^2} + \frac{UV^{5/2}}{G^{5/2}}.$$

In particular it is $\ll Q^2/G^2$ for $U = Q^{1/3}$ and $V = Q^{2/3}$.

Proof. For a fixed direction $\vec{c}$ we count the vectors $\vec{\nu}$ in the plane $\vec{c}^{\perp}$; such a count has a "volume" term and a "boundary" term, the latter coming from planes which contain an unusually short vector. The volume terms are easily summed over $\vec{c}$, and the boundary terms are summed by grouping the $\vec{c}$ according to that short vector. Write $\vec{\rho} = G\vec{c}$ with $\vec{c}$ primitive and $|\vec{c}| \leqslant V' = V/G$. The lattice $\vec{c}^{\perp} \cap \mathbb{Z}^3$ has rank 2 and covolume $|\vec{c}|$, so the number of its non-zero vectors of length at most U is $\ll U/\lambda_1(\vec{c}) + U^2/|\vec{c}|$, where $\lambda_1(\vec{c}) \ll |\vec{c}|^{1/2}$ is the length of its shortest non-zero vector. The sum of $U^2/|\vec{c}|$ over $|\vec{c}| \leqslant V'$ is $\ll U^2V'^2$. For the other term we choose for each $\vec{c}$ a primitive $\vec{y} \perp \vec{c}$ with $|\vec{y}| = \lambda_1(\vec{c})$, and obtain

$$\sum_{|\vec{c}| \leqslant V'} \frac{1}{\lambda_1(\vec{c})} \leqslant \sum_{\substack{\vec{y} \text{ primitive} \\ |\vec{y}| \ll V'^{1/2}}} \frac{\#\{\vec{c}\colon\ |\vec{c}| \leqslant V',\ \vec{c} \perp \vec{y}\}}{|\vec{y}|} \ll \sum_{|\vec{y}| \ll V'^{1/2}} \Big(\frac{V'^2}{|\vec{y}|^2} + \frac{V'}{|\vec{y}|}\Big) \ll V'^{5/2}. \quad \square$$

Proof of Proposition 1.3. By Lemma 5.4 we may assume that $W \leqslant Q^{\varepsilon} r_q$. We treat the point, plane and flag weights in turn; in each case we integrate the bound of Theorem 1.2 over $|\vec{w}| \leqslant W$ and multiply by the number of pairs $(\vec{a}, \text{term})$ that occur. Write $T = Q^{4/3}$. We use the elementary estimates

(5.3)
$$\int_{\vec{w} \in \mathbb{R}^3,\, |\vec{w}| \leqslant U} \frac{d\vec{w}}{1 + |\vec{w}|T} \asymp U^2 \min(U, T^{-1}), \qquad \int_{\vec{y} \in \mathbb{R}^2,\, |\vec{y}| \leqslant U} \frac{d\vec{y}}{1 + |\vec{y}|T} \asymp U \min(U, T^{-1}).$$

By (1.7) and (5.3), $\int_{|\vec{w}| \leqslant W} |p_{1,q}| \ll (Q/q) \min(W, r_q)^2 \min(W, T^{-1})$, and there are at most q^3 residues $\vec{a}$. This gives the first term of the proposition.

By (1.8) and (5.3), integrating first along $\hat{\vec{\nu}}$ and then over $\vec{\nu}^{\perp}$,

$$\int_{|\vec{w}| \leqslant W} |p_{2,\vec{\nu},k,q}| \ll (Q/(kq)) \min(W, r_q) \min(W, T^{-1})^2.$$

By Lemma 5.7(i) the number of $\vec{a}$ with $\vec{\nu} \in \Lambda^{\perp}(\vec{a}, q)$ is at most $q^2 \gcd(q, g)$, and $k = g/\gcd(g, q)$. So the plane terms contribute

$$\ll qQ \min(W, r_q) \min(W, T^{-1})^2 \sum_{|\vec{\nu}| \leqslant Q^{1/3}} \frac{\gcd(q, g)^2}{g},$$

and the last sum is

$$\ll \sum_g \gcd(q,g)^2 g^{-1}(Q^{1/3}/g)^3 \leqslant Q \sum_{d|q} d^2 \sum_{d|g} g^{-4} \ll Q.$$

By (1.9),

$$\int_{|\vec{w}|\leqslant W} |p_{3,\vec{\nu},\vec{\rho},k,q}| \ll Q^{2\varepsilon} \min(W, r_q) \min(W, T^{-1})^2.$$

By Lemma 5.7(ii) a flag with $\gcd(\vec{\rho}) = G$ occurs for at most $q \gcd(q,G)^2$ residues $\vec{a}$, and by Lemma 5.8

$$\sum_{\text{flags}} \gcd(q,G)^2 \ll \sum_G \gcd(q,G)^2 \frac{Q^2}{G^2} \leqslant Q^2 \sum_{d|q} d^2 \sum_{d|G} G^{-2} \ll Q^{2+\varepsilon}.$$

So the flag terms contribute $\ll qQ^{2+3\varepsilon} \min(W, r_q) \min(W, T^{-1})^2$. This proves the bound of the proposition. Taking $W = Q^{\varepsilon} r_q$, the bound is $\ll Q^{-1+4\varepsilon}$ for each q, and summing over $q \leqslant Q$ gives the last assertion. □

We note that for $q \asymp Q$ and $|\vec{w}| \leqslant Q^{-4/3}$ all three kinds of terms contribute a mass $\asymp Q^{-1}$ for each q, so the total mass in this range is $\asymp 1$. This is the range which dominates the error terms in Theorem 6.5.

6. Setting up the circle method

All the notation introduced in this section will be used throughout the rest of the paper. We first recall the geometry of nets of quadrics that we need, and use it to split the weights $p_{\Lambda(\vec{a},q)}$ into a generic and an exceptional part. We then reduce Theorem 6.5 to estimates for three sums N_0, N_1, N_2.

Why do we need geometry at this point? After Poisson summation, the exponential sums that appear are Gauss sums for the single quadratic form $F_{\vec{a}} = \vec{a} \cdot \vec{F}$ modulo q, and their size depends on how singular $F_{\vec{a}}$ is modulo q. So we need to know how the singular members of the net $\{F_{\vec{\beta}}\}$ are distributed. The answer, given by Lemma 6.1, is that the singular members form a plane curve $\Delta = 0$ of degree s, almost all of whose points have corank one, and that there are only finitely many members of corank two. The fractions $\vec{a}/q$ which lie close to this curve are the ones which need special care, and they are collected into the exceptional part of the weights at the end of §6.1.

6.1. Geometry of nets of quadrics.

Let $\vec{F} = (F_1, F_2, F_3)$ be as in Theorem 6.5, and let M_1, M_2, M_3 be the symmetric integer matrices with $F_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T M_i \mathbf{x}$. For $\vec{\beta}$ in $\mathbb{Q}^3$, or in any field containing $\mathbb{Q}$, we put

$$M_{\vec{\beta}} = \beta_1 M_1 + \beta_2 M_2 + \beta_3 M_3, \qquad F_{\vec{\beta}} = \vec{\beta} \cdot \vec{F}, \qquad \Delta(\vec{\beta}) = \det M_{\vec{\beta}}.$$

Then Δ is a ternary form of degree s with integer coefficients, which we call the discriminant of the net. Let D_F be a positive integer, divisible by every prime $p \leqslant s$ and by the content of Δ, such that the reduction modulo p of the variety $F_1 = F_2 = F_3 = 0$ is a non-singular complete intersection of codimension 3 for every prime $p \nmid D_F$. Such an integer exists because the variety is non-singular over $\mathbb{Q}$. As in [LRV24, §5.3] we call the primes dividing D_F *bad* and the others *good*.

Lemma 6.1. *Let k be an algebraically closed field of characteristic different from 2, such that $F_1 = F_2 = F_3 = 0$ defines a non-singular complete intersection of codimension 3 over k; this holds for $k = \overline{\mathbb{Q}}$, and for $k = \overline{\mathbb{F}}_p$ when p is good.*

(i) Every non-zero $\vec{\beta} \in k^3$ satisfies $\operatorname{corank} M_{\vec{\beta}} \leqslant 2$.

(ii) If $\operatorname{corank} M_{\vec{\beta}} = 1$ *and* $\mathbf{v}$ *spans* $\ker M_{\vec{\beta}}$*, then* $\nabla\Delta(\vec{\beta}) = c\,(F_1(\mathbf{v}), F_2(\mathbf{v}), F_3(\mathbf{v}))$ *for some* $c \neq 0$*, and this vector is non-zero. In particular the points of corank* 1 *are non-singular points of the curve* $\Delta = 0$ *in* $\mathbb{P}^2$*.*

(iii) The set $\mathcal{Z}_2$ *of* $[\vec{\beta}] \in \mathbb{P}^2(k)$ *with* $\operatorname{corank} M_{\vec{\beta}} = 2$ *is finite.*

(iv) Δ *is non-zero in* $k[\vec{\beta}]$ *and has no repeated factors.*

(v) Let $[\vec{\beta}_0] \in \mathcal{Z}_2$ *and* $W = \ker M_{\vec{\beta}_0}$*. The linear map* $\vec{b} \mapsto M_{\vec{b}}|_W$ *from* $k^3/k\vec{\beta}_0$ *to the space of quadratic forms on* W *is injective, and the determinant does not vanish identically on its image.*

Proof. All five parts are consequences of the non-singularity of the variety $F_1 = F_2 = F_3 = 0$, through one simple principle: a common zero of the three forms which also lies in the kernel of some member $M_{\vec{\beta}}$ is a singular point. In each part we assume the conclusion fails and construct such a point. We use repeatedly the following observation: *if* $\mathbf{x} \neq \mathbf{0}$*,* $M_{\vec{\beta}}\mathbf{x} = \mathbf{0}$ *for some non-zero* $\vec{\beta}$*, and* $\vec{F}(\mathbf{x}) = \vec{0}$*, then* $[\mathbf{x}]$ *is a singular point of the variety*, since the rows $M_i\mathbf{x}$ of its Jacobian matrix at $\mathbf{x}$ satisfy $\sum_i \beta_i M_i\mathbf{x} = \mathbf{0}$.

(i) If $\dim\ker M_{\vec{\beta}} \geqslant 3$, choose $\vec{\beta}', \vec{\beta}''$ such that $\vec{\beta}, \vec{\beta}', \vec{\beta}''$ is a basis of k^3. The quadratic forms $F_{\vec{\beta}'}$ and $F_{\vec{\beta}''}$ restricted to $\ker M_{\vec{\beta}}$ have a common non-trivial zero $\mathbf{x}$, and $F_{\vec{\beta}}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T M_{\vec{\beta}}\mathbf{x} = 0$. So $\vec{F}(\mathbf{x}) = \vec{0}$, contradicting the observation. (Two quadratic forms in at least three variables always have a common non-trivial zero over an algebraically closed field. The third form $F_{\vec{\beta}}$ needs no condition, since it vanishes identically on $\ker M_{\vec{\beta}}$.)

(ii) By Jacobi's formula, $\partial_i\Delta(\vec{\beta}) = \operatorname{tr}(\operatorname{adj}(M_{\vec{\beta}})M_i)$. Since $M_{\vec{\beta}}$ is symmetric of corank 1, $\operatorname{adj}(M_{\vec{\beta}}) = c'\mathbf{v}\mathbf{v}^T$ with $c' \neq 0$, and so $\partial_i\Delta(\vec{\beta}) = 2c'F_i(\mathbf{v})$. If this vanished for all i we would have $\vec{F}(\mathbf{v}) = \vec{0}$ and $M_{\vec{\beta}}\mathbf{v} = \mathbf{0}$. The second statement follows, since a point of the curve $\Delta = 0$ at which $\nabla\Delta \neq \vec{0}$ is a non-singular point.

(iii) Suppose that $\mathcal{Z}_2$ contains an irreducible curve C. The idea is that, moving along C, the two-dimensional kernel moves with it, and differentiating the relation $M_{\vec{\gamma}(t)}\mathbf{v}(t) = \mathbf{0}$ shows that a second member of the net, besides $F_{\vec{\gamma}_0}$, also vanishes on the kernel W. This leaves only one quadratic condition on the plane W, which then has a non-trivial zero. Take a general point $[\vec{\gamma}_0]$ of C and a formal arc $\vec{\gamma}(t) \in k[[t]]^3$ on the cone over C with $\vec{\gamma}(0) = \vec{\gamma}_0$ and $\vec{\gamma}'(0) \notin k\vec{\gamma}_0$. The kernel of $M_{\vec{\gamma}(t)}$ over $k((t))$ has dimension 2, and its intersection with $k[[t]]^s$ is a free module whose reduction at $t = 0$ is $W = \ker M_{\vec{\gamma}_0}$. So every $\mathbf{v} \in W$ lifts to $\mathbf{v}(t)$ with $M_{\vec{\gamma}(t)}\mathbf{v}(t) = \mathbf{0}$. Differentiating at $t = 0$ and pairing with $\mathbf{v}' \in W$ gives $\mathbf{v}'^T M_{\vec{\gamma}'(0)}\mathbf{v} = 0$. Thus $F_{\vec{\gamma}_0}$ and $F_{\vec{\gamma}'(0)}$ both vanish on W, so the restrictions of F_1, F_2, F_3 to the plane W span a space of binary quadratic forms of dimension at most 1. These forms have a common zero $\mathbf{x} \in W \setminus \{\mathbf{0}\}$, and $M_{\vec{\gamma}_0}\mathbf{x} = \mathbf{0}$, contradicting the observation.

(iv) If Δ vanished identically or had a repeated factor Δ_1, then $\nabla\Delta$ would vanish on the curve $\Delta_1 = 0$ (or on all of $\mathbb{P}^2$). By (ii) all points of this curve would have corank at least 2, contradicting (iii). (Indeed, by (ii) a point of corank exactly one has $\nabla\Delta \neq \vec{0}$.)

(v) If $\vec{b} \notin k\vec{\beta}_0$ and $M_{\vec{b}}|_W = 0$, the argument of (iii) applies with $\vec{b}$ in place of $\vec{\gamma}'(0)$. So the image S of the map has dimension 2. If the determinant vanished on S, every element of S would be a symmetric 2×2 matrix of rank at most 1; such matrices form the cone over a non-singular conic, which contains no line. □

We call a primitive vector $\vec{c} \in \mathbb{Z}^3$ *singular* if $\Delta(\vec{c}) = 0$, that is, if the quadratic form $F_{\vec{c}}$ is singular, and *non-singular* otherwise. We write $\mathcal{C}_F$ for the set of singular primitive vectors, and $r(\vec{c}) = \operatorname{corank} M_{\vec{c}} \in \{1, 2\}$ for $\vec{c} \in \mathcal{C}_F$. By Lemma 6.1(iii) only finitely many $\vec{c} \in \mathcal{C}_F$ have $r(\vec{c}) = 2$, and since $\Delta \neq 0$

$$\#\{\vec{c} \in \mathcal{C}_F : \ |\vec{c}| \leqslant B\} \ll B^2. \tag{6.1}$$

For a pair of quadrics there are at most s singular directions [LRV24, §5.3]; for a net they form the rational points of a plane curve, and if the curve contains a rational line there are $\asymp B^2$ of them. This does happen: if F_1 and F_2 are both singular at a point $\mathbf{x}_0$ with $F_3(\mathbf{x}_0) \neq 0$, every member of the pencil spanned by F_1 and F_2 is singular, and the variety may still be non-singular.

For a primitive $\vec{\nu}_0 \in \mathbb{Z}^3$ let $\ell_{\vec{\nu}_0}$ be the line $\vec{\nu}_0 \cdot \vec{\beta} = 0$ in $\mathbb{P}^2$. Lines of this form will appear in the plane terms of the delta symbol: a plane term with normal vector $\vec{\nu}$ collects the fractions $\vec{a}/q$ with $q \mid \vec{\nu} \cdot \vec{a}$, that is, the fractions lying near the line $\ell_{\vec{\nu}_0}$. So we need to know how $\ell_{\vec{\nu}_0}$ meets the discriminant curve. Choose a basis $\vec{w}_1, \vec{w}_2$ of the lattice $\vec{\nu}_0^{\perp} \cap \mathbb{Z}^3$ with $|\vec{w}_1|, |\vec{w}_2| \ll |\vec{\nu}_0|$; this is possible since the lattice has rank 2 and determinant $|\vec{\nu}_0|$. Put

$$G_{\vec{\nu}_0}(x, y) = \Delta(x\vec{w}_1 + y\vec{w}_2),$$

a binary form of degree s with integer coefficients of size $\ll |\vec{\nu}_0|^s$, which describes the restriction of Δ to $\ell_{\vec{\nu}_0}$.

Definition 6.2. Let $\mathcal{N}_F$ be the set of primitive $\vec{\nu}_0 \in \mathbb{Z}^3$ such that either $G_{\vec{\nu}_0}$ vanishes identically (that is, $\ell_{\vec{\nu}_0}$ is contained in the curve $\Delta = 0$), or $G_{\vec{\nu}_0}$ has a root of multiplicity at least 3 in $\mathbb{P}^1(\overline{\mathbb{Q}})$.

A general line meets the curve $\Delta = 0$ transversally, in s distinct points. The lines in $\mathcal{N}_F$ are the special ones: lines contained in the curve, and lines which are inflectional tangents or pass through a corank-two point in a tangent direction. The following lemma shows that there are only finitely many of them, and that for all other lines the transversality survives reduction modulo almost every prime, with an exceptional set of primes controlled by a single integer $R_{\vec{\nu}_0}$ of polynomial size.

Lemma 6.3. *(i) The set $\mathcal{N}_F$ is finite.*

(ii) For every primitive $\vec{\nu}_0 \notin \mathcal{N}_F$ there is a non-zero integer $R_{\vec{\nu}_0}$ with $|R_{\vec{\nu}_0}| \ll |\vec{\nu}_0|^C$, where C depends only on s, such that for every prime $p \nmid D_F R_{\vec{\nu}_0}$ the reduction of $G_{\vec{\nu}_0}$ modulo p is non-zero and all its roots in $\mathbb{P}^1(\overline{\mathbb{F}}_p)$ have multiplicity at most 2.

Proof. Part (i) is classical plane curve geometry. For part (ii) we encode the condition "no root of multiplicity three" as the non-vanishing of a resultant, which is an integer; the primes for which the condition fails modulo p must divide it.

(i) The curve $\Delta = 0$ is reduced (Lemma 6.1(iv)), so it contains only finitely many lines. A line not contained in the curve meets it with multiplicity at least 3 at some point P only in two situations. If P is a non-singular point of the curve, the line is the tangent at P and P is a flex of a non-linear component. A reduced plane curve over a field of characteristic 0 has only finitely many flexes on its non-linear components, since they are cut out by the Hessian. If P is a singular point, then P has corank 2 by Lemma 6.1(ii), so it is one of the finitely many points of $\mathcal{Z}_2$. It has multiplicity exactly 2 on the curve, because the quadratic term of Δ at P is a non-zero multiple of $\det(M_{\vec{b}}|_W)$, as in the proof of Lemma 8.6. A line through a double point meets

the curve there with multiplicity at least 3 only if it lies in the tangent cone, which consists of at most two lines. Hence only finitely many lines occur.

(ii) Let $G = G_{\vec{\nu}_0}$. For $p > s$ (and over $\overline{\mathbb{Q}}$), a root of G has multiplicity at least 3 if and only if G and its three second partial derivatives G_{xx}, G_{xy}, G_{yy} all vanish there. Since $\vec{\nu}_0 \notin \mathcal{N}_F$, at each of the at most s roots of G in $\mathbb{P}^1(\overline{\mathbb{Q}})$ the vector (G_{xx}, G_{xy}, G_{yy}) is non-zero. The integer pairs (t_1, t_2) for which $G_{xx} + t_1 G_{xy} + t_2 G_{yy}$ vanishes at one of these roots lie on at most s lines in $\mathbb{R}^2$. Each line contains at most $s+1$ points of $\{0, \dots, s\}^2$, so we may choose $t_1, t_2 \in \{0, \dots, s\}$ off all of them. Put $A = G_{xx} + t_1 G_{xy} + t_2 G_{yy}$ and let $R_{\vec{\nu}_0} = \operatorname{Res}(A, G)$ be the resultant of the binary forms A and G. It is non-zero because A and G have no common root in $\mathbb{P}^1(\overline{\mathbb{Q}})$. It is a polynomial of bounded degree in the coefficients, so $|R_{\vec{\nu}_0}| \ll |\vec{\nu}_0|^C$. If $p \nmid R_{\vec{\nu}_0}$, then the reductions of A and G have no common root in $\mathbb{P}^1(\overline{\mathbb{F}}_p)$; in particular G is non-zero modulo p, since otherwise the resultant would vanish modulo p. If moreover $p > s$, a root of G modulo p of multiplicity at least 3 would be a common root of G and A modulo p. This proves (ii), since D_F is divisible by all primes $p \leqslant s$. □

The discriminant curve may contain lines, including lines defined over $\mathbb{Q}$. The following lemma describes exactly when this happens; we use it in §14.11.

Lemma 6.4. *Let k be an algebraically closed field of characteristic 0, and let $\ell \subset \mathbb{P}^2$ be a line with $\Delta|_\ell \equiv 0$, spanned by $[\vec{\beta}_1], [\vec{\beta}_2]$. Then the members of the pencil ℓ have a common kernel vector: there is $\mathbf{x}_0 \in k^s \setminus \{\mathbf{0}\}$, unique up to scaling, with $M_{\vec{\beta}_1}\mathbf{x}_0 = M_{\vec{\beta}_2}\mathbf{x}_0 = \mathbf{0}$; moreover $F_{\vec{\gamma}}(\mathbf{x}_0) \neq 0$ for every $\vec{\gamma} \notin \ell$. If ℓ is defined over $\mathbb{Q}$, then $\mathbf{x}_0$ may be taken primitive in $\mathbb{Z}^s$, and after a unimodular change of variables with $\mathbf{x}_0 = \mathbf{e}_1$ the system takes the form*

$$F_{\vec{\beta}_1} = G_1(\mathbf{x}'), \qquad F_{\vec{\beta}_2} = G_2(\mathbf{x}'), \qquad F_{\vec{\gamma}} = cx_1^2 + 2x_1 L(\mathbf{x}') + G_3(\mathbf{x}'),$$

with $\mathbf{x}' = (x_2, \dots, x_s)$, quadratic forms G_i, a linear form L and $c = F_{\vec{\gamma}}(\mathbf{x}_0) \neq 0$. Conversely, every system of this shape has ℓ in its discriminant curve.

Proof. A pencil of symmetric matrices whose determinant vanishes identically has a kernel vector depending polynomially on the parameter. The proof shows, using the non-singularity of the variety, that this kernel vector must in fact be constant. This is a special case of the Kronecker canonical form for pencils [Tho91], and we give a direct argument. Write $A = M_{\vec{\beta}_1}$ and $B = M_{\vec{\beta}_2}$. Since $\det(sA + tB) \equiv 0$, the matrix $sA + tB$ has a non-zero kernel over $k(s,t)$; clearing denominators, there is a non-zero vector $\mathbf{x}(s,t) = \sum_{i=0}^{\varepsilon} s^{\varepsilon - i} t^i \mathbf{x}_i$ of forms of degree $\varepsilon \geqslant 0$ with $(sA + tB)\mathbf{x}(s,t) \equiv \mathbf{0}$, and we take ε minimal. Then $\mathbf{x}(s_0, t_0) \neq \mathbf{0}$ for every $(s_0, t_0) \neq (0,0)$, for otherwise $\mathbf{x}(s,t)$ would be divisible by $t_0 s - s_0 t$, contradicting minimality. Comparing coefficients gives

$$A\mathbf{x}_0 = \mathbf{0}, \qquad A\mathbf{x}_i + B\mathbf{x}_{i-1} = \mathbf{0} \quad (1 \leqslant i \leqslant \varepsilon), \qquad B\mathbf{x}_\varepsilon = \mathbf{0}.$$

Suppose that $\varepsilon \geqslant 1$. For $0 \leqslant i \leqslant \varepsilon - 1$ and $1 \leqslant j \leqslant \varepsilon$ these relations give

$$\mathbf{x}_i^T A \mathbf{x}_j = -\mathbf{x}_i^T B \mathbf{x}_{j-1} = -(B\mathbf{x}_i)^T \mathbf{x}_{j-1} = \mathbf{x}_{i+1}^T A \mathbf{x}_{j-1}.$$

Iterating, every $\mathbf{x}_i^T A \mathbf{x}_j$ equals either some $\mathbf{x}_m^T A \mathbf{x}_0 = 0$ or some $\mathbf{x}_\varepsilon^T A \mathbf{x}_m$, and the latter is 0 for $m = 0$ and equals $-(B\mathbf{x}_\varepsilon)^T \mathbf{x}_{m-1} = 0$ for $m \geqslant 1$. So the span W of $\mathbf{x}_0, \dots, \mathbf{x}_\varepsilon$ is totally isotropic for A, and in the same way for B. Let $\vec{\gamma}$ complete $\vec{\beta}_1, \vec{\beta}_2$ to a basis. The binary form $F_{\vec{\gamma}}(\mathbf{x}(s,t))$ has degree $2\varepsilon \geqslant 2$, so it has a zero (s_0, t_0) in $\mathbb{P}^1(k)$ (or vanishes identically). The vector $\mathbf{w} = \mathbf{x}(s_0, t_0) \neq \mathbf{0}$ satisfies $F_{\vec{\beta}_1}(\mathbf{w}) = F_{\vec{\beta}_2}(\mathbf{w}) = F_{\vec{\gamma}}(\mathbf{w}) = 0$, so $[\mathbf{w}] \in V$, and $(s_0 A + t_0 B)\mathbf{w} = \mathbf{0}$, so the rows

$A\mathbf{w}, B\mathbf{w}$ of the Jacobian matrix at $\mathbf{w}$ are linearly dependent. This contradicts the non-singularity of V. Hence $\varepsilon = 0$, and $\mathbf{x}_0 = \mathbf{x}(s,t)$ is a common kernel vector.

If $F_{\vec\gamma}(\mathbf{x}_0) = 0$, then $[\mathbf{x}_0] \in V$ and the Jacobian matrix at $\mathbf{x}_0$ has rank at most 1, a contradiction; since $F_{\vec\beta_1}(\mathbf{x}_0) = F_{\vec\beta_2}(\mathbf{x}_0) = 0$, this gives $F_{\vec\gamma}(\mathbf{x}_0) \neq 0$ for all $\vec\gamma \notin \ell$. If the common kernel had dimension at least 2, the form $F_{\vec\gamma}$ would have a zero $\mathbf{w} \neq \mathbf{0}$ on it, and the same argument would give a singular point of V. If ℓ is defined over $\mathbb{Q}$ we may take $\vec\beta_1, \vec\beta_2 \in \mathbb{Z}^3$; the common kernel of A and B is then a line defined over $\mathbb{Q}$, spanned by a primitive $\mathbf{x}_0 \in \mathbb{Z}^s$, and completing $\mathbf{x}_0$ to a basis of $\mathbb{Z}^s$ gives a unimodular change of variables with $\mathbf{x}_0 = \mathbf{e}_1$. In the new variables the first row and column of A and B vanish, which is the stated shape, with $c = F_{\vec\gamma}(\mathbf{e}_1) = F_{\vec\gamma}(\mathbf{x}_0)$. The converse is clear, since every member of the pencil then has $\mathbf{e}_1$ in its kernel. □

For example, if $F_i = a_i x_1^2 + G_i(\mathbf{x}')$ with $\vec a = (a_1, a_2, a_3) \neq \vec 0$, then $\Delta(\vec\beta) = (\vec a \cdot \vec\beta) \det M'_{\vec\beta}$, where $M'_{\vec\beta}$ is the matrix of $\vec\beta \cdot \mathbf{G}$, so the line $\vec a \cdot \vec\beta = 0$ lies in the discriminant curve; and V is non-singular for a generic choice of the G_i. (A singular point with $x_1 \neq 0$ would require $\mathbf{x}'$ to lie in the kernel of a member $G_{\vec\lambda}$ with $\vec\lambda \perp \vec a$, together with the three equations $a_i x_1^2 + G_i(\mathbf{x}') = 0$; for generic G_i the pencil $\{G_{\vec\lambda} : \vec\lambda \perp \vec a\}$ has finitely many singular members, each with a one-dimensional kernel, and these equations have no common solution. A singular point with $x_1 = 0$ would be a singular point of $G_1 = G_2 = G_3 = 0$.) So rational lines in the discriminant curve do occur.

We can now say which parts of the delta symbol need special attention. A flag term with direction $\vec c$ collects fractions $\vec a/q$ close to the line through $\vec c$, so the exponential sums it produces behave like Gauss sums for the form $F_{\vec c}$; if $F_{\vec c}$ is singular, these sums are larger than usual. Similarly a plane term with normal $\vec\nu_0$ collects fractions near the line $\ell_{\vec\nu_0}$ of $\mathbb{P}^2$, and it is harmless unless this line meets the discriminant curve in a degenerate way. We therefore separate these two kinds of terms.

We split

$$p_{\Lambda(\vec a,q)} = p^{\mathrm{gen}}_{\Lambda(\vec a,q)} + p^{\mathrm{exc}}_{\Lambda(\vec a,q)},$$

where $p^{\mathrm{exc}}_{\Lambda(\vec a,q)}$ consists of the terms of (5.1) of the following two kinds: the terms $p_{2,\vec\nu,k,q}$ with $\vec\nu/\gcd(\vec\nu) \in \mathcal{N}_F$ (Definition 6.2), and the terms $p_{3,\vec\nu,\vec\rho,k,q}$ with $\vec\rho/\gcd(\vec\rho) \in \mathcal{C}_F$.

6.2. **The unrefined method and the sums N_i.** The first step towards Theorem 1.1 is the following result, which uses the delta symbol of Theorem 1.2 with absolute values throughout, and no average over the numerators. Its proof occupies Sections 6–12, and it provides the framework, the notation and most of the estimates for the Kloosterman refinement in Sections 14 and 15.

Theorem 6.5 (the unrefined method)**.** *Under the hypotheses of Theorem 1.1, the asymptotic formula* (1.1) *holds for $s \geqslant 25$, with any $\Delta < (s-24)/4$.*

For the rest of the paper we adopt the following conventions. We let $P \geqslant 1$ be a large parameter. Let $\mathbf{z}_0$ be a non-zero real zero of $\vec F$. Since the variety is non-singular, the Jacobian matrix of $\vec F$ at $\mathbf{z}_0$ has rank 3. We fix a closed ball U around $\mathbf{z}_0$, not containing $\mathbf{0}$, on which the Jacobian matrix has rank 3, and we take $w \in C_c^\infty(\mathbb{R}^s)$ supported in U. We work with the counting function

$$N(P) = \sum_{\substack{\mathbf{x}\in\mathbb{Z}^s \\ \vec F(\mathbf{x})=\vec 0}} w(\mathbf{x}/P) = \sum_{\mathbf{x}\in\mathbb{Z}^s} w(\mathbf{x}/P)\, \delta_{\vec F(\mathbf{x})}.$$

Let $M_U = \max(1, \max_{\mathbf{x}\in U} |\vec{F}(\mathbf{x})|)$. We choose

$$Q = (4M_U P^2)^{3/4}, \qquad \text{so that} \qquad Q^{4/3} = 4M_U P^2,$$

and $|\vec{F}(\mathbf{x})| \leqslant \frac{1}{4}Q^{4/3}$ whenever $w(\mathbf{x}/P) \neq 0$. We fix $\delta > 0$ small and put $q_0 = Q^{1/3-\delta}$. The exponential sums and integrals are

$$S_q(\vec{a}, \mathbf{u}) = \sum_{\mathbf{b} \bmod q} \mathrm{e}_q\Big(\vec{a}\cdot\vec{F}(\mathbf{b}) + \mathbf{b}\cdot\mathbf{u}\Big), \tag{6.2}$$

$$I_q(\vec{w}, \mathbf{u}) = \int_{\mathbb{R}^s} w\Big(\frac{\mathbf{x}}{P}\Big)\mathrm{e}\Big(\vec{w}\cdot\vec{F}(\mathbf{x})\Big)\mathrm{e}_q(-\mathbf{x}\cdot\mathbf{u})\,d\mathbf{x}. \tag{6.3}$$

Lemma 6.6. *For any $N > 0$ we have*

$$N(P) = N_0(P) + N_1(P) + N_2(P) + O_N(P^{-N}),$$

where

$$N_0(P) = \sum_{1\leqslant q<q_0} q^{-s} \sideset{}{^*}\sum_{\vec{a} \bmod q} S_q(\vec{a}, \mathbf{0}) \int_{\mathbb{R}^3} p_{\Lambda(\vec{a},q)}(\vec{w})\, I_q(\vec{w}, \mathbf{0})\, d\vec{w},$$

$$N_1(P) = \sum_{1\leqslant q\leqslant Q} q^{-s} \sideset{}{^*}\sum_{\vec{a} \bmod q} \sum_{\mathbf{u}\in\mathbb{Z}^s}{}' S_q(\vec{a}, \mathbf{u}) \int_{\mathbb{R}^3} p^{\mathrm{gen}}_{\Lambda(\vec{a},q)}(\vec{w})\, I_q(\vec{w}, \mathbf{u})\, d\vec{w},$$

$N_2(P)$ is defined as $N_1(P)$ with $p^{\mathrm{exc}}_{\Lambda(\vec{a},q)}$ in place of $p^{\mathrm{gen}}_{\Lambda(\vec{a},q)}$, and $\sum'$ runs over all $\mathbf{u} \neq \mathbf{0}$, together with $\mathbf{u} = \mathbf{0}$ when $q \geqslant q_0$.

Proof. We apply Theorem 1.2 with $\vec{n} = \vec{F}(\mathbf{x})$ and sum over $\mathbf{x}$; the error term contributes $O_N(P^s Q^{-N})$. For fixed q, $\vec{a}$ and $\vec{w}$ we break the sum over $\mathbf{x}$ into residue classes modulo q and apply Poisson summation:

$$\sum_{\mathbf{x}\in\mathbb{Z}^s} w\Big(\frac{\mathbf{x}}{P}\Big)\mathrm{e}\Big(\Big(\frac{\vec{a}}{q} + \vec{w}\Big)\cdot\vec{F}(\mathbf{x})\Big) = q^{-s}\sum_{\mathbf{u}\in\mathbb{Z}^s} S_q(\vec{a}, \mathbf{u})\, I_q(\vec{w}, \mathbf{u}).$$

Since $p_{\Lambda(\vec{a},q)}$ is a Schwartz function and $I_q(\vec{w}, \mathbf{u})$ decays rapidly in $\mathbf{u}$ (Lemma 7.1), we may interchange the sum over $\mathbf{u}$ with the integral over $\vec{w}$. The terms are then sorted into N_0, N_1 and N_2. □

Our principal task will be to prove the following lemma. It immediately implies Theorem 6.5. Its three parts correspond to the three sources of contributions: the main term N_0, coming from small moduli and the zero dual vector; the generic error terms N_1; and the exceptional error terms N_2, coming from the singular directions of the net. The bound required for N_1 and N_2 is $P^{3s/4+\varepsilon}$, which is smaller than the main term P^{s-6} exactly when $s > 24$.

Lemma 6.7. *Let $\varepsilon > 0$. With the notation above, for all sufficiently small δ we have*

(a) $N_0(P) = \mathfrak{S}\mathfrak{J}P^{s-6} + O\Big(P^{s-6-(s-8)/4+\varepsilon}\Big)$ *for $s \geqslant 11$;*
(b) $N_1(P) \ll P^{s-6-(s-24)/4+\varepsilon}$;
(c) $N_2(P) \ll P^{s-6-(s-24)/4+\varepsilon}$.

Part (a) is proved in Section 9, part (b) in Section 11, and part (c) in Section 12. Note that $s - 6 - (s-24)/4 = 3s/4$: the bounds in (b) and (c) are the quantity $Q^{s/2}$ of Section 1.3.

Proof of Theorem 6.5. Combining Lemma 6.6 with Lemma 6.7, we obtain (1.1) with any $\Delta < (s-24)/4$, since $(s-8)/4 > (s-24)/4$. □

7. Exponential integrals

We need two facts about the integrals: that the integral is negligible unless the dual vector $\mathbf{u}$ is small (by repeated integration by parts), and that it decays like $X^{-(s-2)/2}$ when $\vec{w}$ is large (by stationary phase, using that every member of the net has rank at least $s-2$). Both are standard.

Recall the exponential integral $I_q(\vec{w},\mathbf{u})$ defined in (6.3). After the change of variables $\mathbf{x}\mapsto P\mathbf{x}$ we have

$$I_q(\vec{w},\mathbf{u}) = P^s\int_{\mathbb{R}^s} w(\mathbf{x})\,\mathrm{e}\Big(P^2\vec{w}\cdot\vec{F}(\mathbf{x}) - \frac{P}{q}\,\mathbf{x}\cdot\mathbf{u}\Big)d\mathbf{x}. \tag{7.1}$$

Throughout this section and later we write $X_{\vec{w}} = 1+P^2|\vec{w}|$. The quantity $X_{\vec{w}}$ measures how much the phase $P^2\vec{w}\cdot\vec{F}(\mathbf{x})$ oscillates on the support of w: for $|\vec{w}|\leqslant P^{-2}$ it hardly oscillates and the integral is essentially P^s, while for larger $\vec{w}$ the phase makes about $X_{\vec{w}}$ turns. Lemma 7.1 says that two things then happen. The dual vectors $\mathbf{u}$ with $|\mathbf{u}|$ much larger than $qX_{\vec{w}}/P$ contribute nothing, because the linear phase $\mathbf{x}\cdot\mathbf{u}/q$ oscillates faster than the quadratic phase can compensate; and the integral itself shrinks by the factor $X_{\vec{w}}^{-1/2}$ for each direction in which the quadratic form $\vec{w}\cdot\vec{F}$ is non-degenerate.

Lemma 7.1. *There is a constant C_1, depending on $\vec{F}$ and U, such that the following holds.*

(i) If $|\mathbf{u}|\geqslant C_1qX_{\vec{w}}/P$, then $I_q(\vec{w},\mathbf{u})\ll_N P^s(P|\mathbf{u}|/q)^{-N}$.

(ii) For all $\vec{w}$ and $\mathbf{u}$ we have $I_q(\vec{w},\mathbf{u})\ll P^sX_{\vec{w}}^{-(s-2)/2}$.

Proof. (i) This is the principle of non-stationary phase: if the gradient of the phase never vanishes on the support of the amplitude, and is large compared with the second derivatives, then repeated integration by parts makes the integral as small as we like. Write $\mathbf{v}=\mathbf{u}/q$. The phase $\phi(\mathbf{x}) = P^2\vec{w}\cdot\vec{F}(\mathbf{x}) - P\mathbf{v}\cdot\mathbf{x}$ has gradient $P^2M_{\vec{w}}\mathbf{x}-P\mathbf{v}$, and $|P^2M_{\vec{w}}\mathbf{x}|\leqslant\frac12C_1P^2|\vec{w}|$ on U for a suitable C_1. If $P|\mathbf{v}|\geqslant C_1X_{\vec{w}}$, then $|\nabla\phi|\geqslant\frac12P|\mathbf{v}|$ on U, while all second derivatives of ϕ are $\ll P^2|\vec{w}|\leqslant P|\mathbf{v}|$ and the higher derivatives vanish. Integrating by parts N times with the operator $(2\pi i)^{-1}|\nabla\phi|^{-2}\nabla\phi\cdot\nabla$, each step gains a factor $\ll(P|\mathbf{v}|)^{-1}$.

(ii) After diagonalising the quadratic form $\vec{w}\cdot\vec{F}$ by an orthogonal change of variables, the integral becomes a product-like integral of one-dimensional Fresnel integrals $\int\mathrm{e}(\beta y^2+by)\,\psi(y)\,dy$, each of which is $\ll|\beta|^{-1/2}$. Since every member of the net has rank at least $s-2$, at least $s-2$ of the coefficients β_i have size $\asymp P^2|\vec{w}|$, and this gives the factor $X_{\vec{w}}^{-(s-2)/2}$. The only technical point is that the amplitude is not a product, which we handle with Parseval's identity. Let $\vec{w}\neq\vec{0}$ and let $\beta_1,\dots,\beta_s$ be the eigenvalues of the real symmetric matrix $\frac12P^2M_{\vec{w}}$. By Lemma 6.1(i), $\operatorname{rank}M_{\hat{\vec{w}}}\geqslant s-2$ for every unit vector $\hat{\vec{w}}$. By continuity and compactness, at least $s-2$ eigenvalues of $M_{\hat{\vec{w}}}$ have absolute value at least $c_F>0$ (the $(s-2)$-th largest absolute value of an eigenvalue of $M_{\hat{\vec{w}}}$ is a continuous function of $\hat{\vec{w}}$, positive on the unit sphere, hence bounded below there), and so at least $s-2$ of the $|\beta_i|$ are $\geqslant\frac12c_FP^2|\vec{w}|$. It therefore suffices to show that

$$\Big|\int_{\mathbb{R}^s}\psi(\mathbf{y})\,\mathrm{e}\Big(\sum_i\beta_iy_i^2+\mathbf{b}\cdot\mathbf{y}\Big)d\mathbf{y}\Big|\ll_\psi\prod_{|\beta_i|\geqslant1}|\beta_i|^{-1/2}$$

for every smooth compactly supported ψ, uniformly in $\mathbf{b}\in\mathbb{R}^s$; we apply this with $\psi(\mathbf{y})=w(O\mathbf{y})$ for an orthogonal matrix O which diagonalises $M_{\vec{w}}$, and the bounds

for the derivatives of ψ are uniform in O. Split $\mathbf{y} = (\mathbf{y}', \mathbf{y}'')$, where $\mathbf{y}''$ consists of the coordinates with $|\beta_i| \geqslant 1$. For fixed $\mathbf{y}'$, Parseval's identity and the formula for the Fourier transform of $\mathrm{e}(\beta t^2)$ give

$$\Big| \int \psi(\mathbf{y}', \mathbf{y}'')\, \mathrm{e}\Big({\sum}'' \beta_i y_i^2 + \mathbf{b}'' \cdot \mathbf{y}'' \Big) d\mathbf{y}'' \Big| \leqslant \big\| \widehat{\psi(\mathbf{y}', \cdot)} \big\|_{L^1} \prod{}'' |2\beta_i|^{-1/2},$$

and $\|\widehat{\psi(\mathbf{y}', \cdot)}\|_{L^1}$ is bounded uniformly in $\mathbf{y}'$. Integrating over $\mathbf{y}'$ in a bounded set gives the claim. □

The bound (ii) loses the factor $X_{\vec{w}}$ for the two eigenvalues which may be small. For a pair of quadrics the corresponding loss concerns one eigenvalue, and in [LRV24, (5.10) and §6] it is controlled through the distance of $\vec{w}$ from the directions of the singular members of the pencil; we do not need such a refinement here.

7.1. **The zero frequency.** For the main term we need the integral $I_q(\vec{w}, \mathbf{0})$ integrated against the weight $p_{\Lambda(\vec{a},q)}$. The point of the next lemma is that this integral can be evaluated exactly, in terms of the density σ of the values of $\vec{F}$, and that the main-term property of the delta symbol turns it into the singular integral.

Let

$$\sigma(\vec{t}) = \int_{\vec{F}(\mathbf{x})=\vec{t}} w(\mathbf{x}) \frac{d\mathcal{H}^{s-3}(\mathbf{x})}{\sqrt{\det\big(J(\mathbf{x})J(\mathbf{x})^T\big)}} \qquad (\vec{t} \in \mathbb{R}^3)$$

be the density of the push-forward of the measure $w(\mathbf{x})\, d\mathbf{x}$ under $\vec{F}$, where J is the Jacobian matrix of $\vec{F}$ and $\mathcal{H}^{s-3}$ is the Hausdorff measure. Since $\vec{F}$ is a submersion on U, the function σ is smooth with compact support, and $\int w(\mathbf{x}) f(\vec{F}(\mathbf{x}))\, d\mathbf{x} = \int_{\mathbb{R}^3} f(\vec{t})\sigma(\vec{t})\, d\vec{t}$ for continuous f. The singular integral is

$$\mathfrak{J} = \sigma(\vec{0}) = \lim_{\eta \to 0} (2\eta)^{-3} \int_{\max_i |F_i(\mathbf{x})| \leqslant \eta} w(\mathbf{x})\, d\mathbf{x}. \tag{7.2}$$

Lemma 7.2. *For* $\gcd(\vec{a}, q) = 1$ *we have*

$$\int_{\mathbb{R}^3} p_{\Lambda(\vec{a},q)}(\vec{w})\, I_q(\vec{w}, \mathbf{0})\, d\vec{w} = P^{s-6} \int_{\mathbb{R}^3} \Phi_{\vec{a},q}(\vec{y})\, \sigma(\vec{y}/P^2)\, d\vec{y},$$

with $\Phi_{\vec{a},q}$ *as in Section 5.4. If* $q < q_0 = Q^{1/3-\delta}$, *this equals* $P^{s-6}\big(\mathfrak{J} + O_N(P^{-N})\big)$.

Proof. Integrating $I_q(\vec{w}, \mathbf{0})$ against $p_{\Lambda(\vec{a},q)}$ simply replaces $\mathrm{e}(\vec{w} \cdot \vec{F}(\mathbf{x}))$ by its Fourier transform $\Phi_{\vec{a},q}(\vec{F}(\mathbf{x}))$; the rest is a change of variables. By (7.1), Fubini's theorem and Fourier inversion,

$$\int p_{\Lambda(\vec{a},q)}(\vec{w})\, I_q(\vec{w}, \mathbf{0})\, d\vec{w} = P^s \int w(\mathbf{x})\, \Phi_{\vec{a},q}(P^2\vec{F}(\mathbf{x}))\, d\mathbf{x} = P^s \int_{\mathbb{R}^3} \Phi_{\vec{a},q}(P^2\vec{t})\sigma(\vec{t})\, d\vec{t},$$

and the substitution $\vec{y} = P^2\vec{t}$ gives the first assertion. Now $\sigma(\vec{y}/P^2) = g(\vec{y}/Q^{4/3})$ with $g(\vec{x}) = \sigma(4M_U\vec{x})$, a fixed smooth function with compact support. By (1.5) in the form of Proposition 5.5, $\int \Phi_{\vec{a},q}(\vec{y}) g(\vec{y}/Q^{4/3})\, d\vec{y} = g(\vec{0}) + O_N(Q^{-N})$, and $g(\vec{0}) = \mathfrak{J}$. □

8. Exponential sums

Recall the exponential sum $S_q(\vec{a}, \mathbf{u})$ defined in (6.2). For fixed $\vec{a}$ it is a Gauss sum attached to the single quadratic form $F_{\vec{a}} = \vec{a} \cdot \vec{F}$, whose matrix is $M_{\vec{a}}$. For $\vec{a} \in \mathbb{Z}^3$ and $q \in \mathbb{N}$ we put

$$\mathcal{K}(\vec{a}, q) = \{\mathbf{h} \in (\mathbb{Z}/q\mathbb{Z})^s :\ M_{\vec{a}}\mathbf{h} \equiv \mathbf{0} \pmod q\}, \qquad K(\vec{a}, q) = \#\mathcal{K}(\vec{a}, q).$$

The whole section revolves around one simple fact (Lemma 8.1): the absolute value of $S_q(\vec{a},\mathbf{u})$ is either 0 or exactly $q^{s/2}K(\vec{a},q)^{1/2}$. If $F_{\vec{a}}$ is non-singular modulo every prime dividing q, then $K(\vec{a},q)=1$ and we have perfect square-root cancellation. The loss $K(\vec{a},q)^{1/2}$ measures how singular $F_{\vec{a}}$ is modulo q, and since the singular members of the net form the curve $\Delta=0$, it is large only when $\Delta(\vec{a})$ is divisible by a large power of the primes dividing q. The main result of the section, Proposition 8.7, shows that this is rare: on average over the residues $\vec{a}$, the loss is only q^{ε} (apart from the bad primes), both over all residues and over the residues on a plane.

8.1. **The modulus of a Gauss sum.**

Lemma 8.1. *For $\vec{a}\in\mathbb{Z}^3$, $q\in\mathbb{N}$ and $\mathbf{u}\in\mathbb{Z}^s$ we have*

$$|S_q(\vec{a},\mathbf{u})|^2 = q^s \sum_{\mathbf{h}\in\mathcal{K}(\vec{a},q)} \mathrm{e}_q\Big(F_{\vec{a}}(\mathbf{h})+\mathbf{u}\cdot\mathbf{h}\Big),$$

and $\mathbf{h}\mapsto \mathrm{e}_q(F_{\vec{a}}(\mathbf{h})+\mathbf{u}\cdot\mathbf{h})$ is a character of the group $\mathcal{K}(\vec{a},q)$. Consequently

$$|S_q(\vec{a},\mathbf{u})| = q^{s/2}K(\vec{a},q)^{1/2}\,\mathbf{1}_{\mathbf{u}\in\mathcal{M}(\vec{a},q)},$$

where

$$\mathcal{M}(\vec{a},q)=\{\mathbf{u} \bmod q:\ F_{\vec{a}}(\mathbf{h})+\mathbf{u}\cdot\mathbf{h}\equiv 0 \pmod q \text{ for all } \mathbf{h}\in\mathcal{K}(\vec{a},q)\}$$

is either empty or a coset of a subgroup of index $K(\vec{a},q)$ in $(\mathbb{Z}/q\mathbb{Z})^s$.

Proof. We use the standard "squaring and differencing" argument for quadratic exponential sums: in $|S_q|^2$ the quadratic terms cancel up to a linear term in one of the variables, and summing over that variable forces the difference $\mathbf{h}$ into the kernel of $M_{\vec{a}}$ modulo q. Writing $\mathbf{b}=\mathbf{b}'+\mathbf{h}$ in

$$|S_q|^2=\sum_{\mathbf{b},\mathbf{b}'}\mathrm{e}_q(F_{\vec{a}}(\mathbf{b})-F_{\vec{a}}(\mathbf{b}')+\mathbf{u}\cdot(\mathbf{b}-\mathbf{b}'))$$

and using $F_{\vec{a}}(\mathbf{b}'+\mathbf{h})-F_{\vec{a}}(\mathbf{b}')=(M_{\vec{a}}\mathbf{h})\cdot\mathbf{b}'+F_{\vec{a}}(\mathbf{h})$, the sum over $\mathbf{b}'$ equals q^s if $\mathbf{h}\in\mathcal{K}(\vec{a},q)$ and 0 otherwise. For $\mathbf{h}_1\in\mathcal{K}(\vec{a},q)$ we have $F_{\vec{a}}(\mathbf{h}_1+\mathbf{h}_2)\equiv F_{\vec{a}}(\mathbf{h}_1)+F_{\vec{a}}(\mathbf{h}_2)$ (mod q), so the summand is a character, and its sum over $\mathcal{K}(\vec{a},q)$ is $K(\vec{a},q)$ or 0. The set $\mathcal{M}(\vec{a},q)$ is the set of $\mathbf{u}$ for which a fixed character of $\mathcal{K}(\vec{a},q)$ coincides with $\mathbf{h}\mapsto\mathrm{e}_q(-\mathbf{u}\cdot\mathbf{h})$, so it is empty or a coset of the annihilator of $\mathcal{K}(\vec{a},q)$; the annihilator has index $K(\vec{a},q)$, as the pairing $\mathbf{u}\cdot\mathbf{h}$ on $(\mathbb{Z}/q\mathbb{Z})^s$ is perfect. □

The next lemma collects some elementary properties of $K(\vec{a},q)$.

Lemma 8.2. *Let $\vec{a}\in\mathbb{Z}^3$ and $q\in\mathbb{N}$.*

- *(i) If $\rho_1\mid\rho_2\mid\cdots\mid\rho_s$ are the invariant factors in the Smith normal form of $M_{\vec{a}}$, then $K(\vec{a},q)=\prod_i\gcd(\rho_i,q)$, where $\gcd(0,q)=q$. In particular $K(\vec{a},q)$ is multiplicative in q.*
- *(ii) If $q'\mid q$, then $K(\vec{a},q)\leqslant K(\vec{a},q')K(\vec{a},q/q')$.*
- *(iii) $K(\lambda\vec{a},q)=K(\vec{a},q)$ if $\gcd(\lambda,q)=1$.*
- *(iv) If $\Delta(\vec{a})\neq 0$, and $D_{s-1}(\vec{a})$ denotes the greatest common divisor of the $(s-1)\times(s-1)$ minors of $M_{\vec{a}}$, then $K(\vec{a},q)\leqslant D_{s-1}(\vec{a})\gcd(\Delta(\vec{a}),q)$.*

Proof. Part (i) follows from the Smith normal form, and (iii) is clear. For (ii), reduction modulo q' maps $\mathcal{K}(\vec{a},q)$ to $\mathcal{K}(\vec{a},q')$, and the kernel consists of the $q'\mathbf{y}$ with $M_{\vec{a}}\mathbf{y}\equiv\mathbf{0}$ (mod q/q'). For (iv), $\rho_1\cdots\rho_{s-1}=D_{s-1}(\vec{a})$ and $\rho_s=\Delta(\vec{a})/D_{s-1}(\vec{a})$, so $\prod_i\gcd(\rho_i,q)\leqslant\rho_1\cdots\rho_{s-1}\gcd(\rho_s,q)$. □

When $F_{\vec{c}}$ is singular, $K(\vec{a},q)$ is large for the $\vec{a}$ close to $\vec{c}$, but then the exponential sum vanishes for most $\mathbf{u}$. This compensation is important: a large K comes together with a restriction of $\mathbf{u}$ to a sublattice of index about K, so that, summed over $\mathbf{u}$, the loss largely disappears. The next lemma makes this precise for the kernel vectors of $M_{\vec{c}}$.

Lemma 8.3. *Let $\vec{c}\in\mathcal{C}_F$, and let $\mathbf{h}_1,\dots,\mathbf{h}_r$ be a basis of the lattice $\ker M_{\vec{c}}\cap\mathbb{Z}^s$, where $r=r(\vec{c})$. Let $q'\mid q$ and $\vec{a}\equiv\lambda\vec{c}\pmod{q'}$. If $S_q(\vec{a},\mathbf{u})\neq 0$, then $\mathbf{u}\cdot\mathbf{h}_j\equiv 0\pmod{q'}$ for $1\leqslant j\leqslant r$.*

Proof. The vectors $(q/q')\mathbf{h}_j$ are elements of the group $\mathcal{K}(\vec{a},q)$ on which the character of Lemma 8.1 is trivial, so the condition defining $\mathcal{M}(\vec{a},q)$ restricts $\mathbf{u}$. Put $\vec{b}=(\vec{a}-\lambda\vec{c})/q'\in\mathbb{Z}^3$ and $\mathbf{x}_j=(q/q')\mathbf{h}_j$. Then

$$M_{\vec{a}}\mathbf{x}_j=(q/q')(\lambda M_{\vec{c}}\mathbf{h}_j+q'M_{\vec{b}}\mathbf{h}_j)=qM_{\vec{b}}\mathbf{h}_j,$$

so $\mathbf{x}_j\in\mathcal{K}(\vec{a},q)$. Moreover

$$F_{\vec{a}}(\mathbf{h}_j)=\lambda F_{\vec{c}}(\mathbf{h}_j)+q'F_{\vec{b}}(\mathbf{h}_j)=q'F_{\vec{b}}(\mathbf{h}_j),$$

so $F_{\vec{a}}(\mathbf{x}_j)=q(q/q')F_{\vec{b}}(\mathbf{h}_j)\equiv 0\pmod{q}$. By Lemma 8.1, $S_q(\vec{a},\mathbf{u})\neq 0$ implies $\mathbf{u}\cdot\mathbf{x}_j\equiv 0\pmod{q}$, that is, $\mathbf{u}\cdot\mathbf{h}_j\equiv 0\pmod{q'}$. □

8.2. **Local averages.** We now show that $K(\vec{a},q)^{1/2}$ is small on average, first over all residues $\vec{a}$ modulo q and then over the residues on a plane. For a pair of quadrics the corresponding analysis in [LRV24, §7] uses the factorisation of the binary form $\det(xM_1+yM_2)$. Here the discriminant Δ is a ternary form, and we use instead the description of the curve $\Delta=0$ in Lemma 6.1. The argument is elementary but has several cases, so we first isolate three local lemmas. Throughout this subsection p is a prime, μ denotes the Haar measure on $\mathbb{Z}_p^3$ normalised by $\mu(\mathbb{Z}_p^3)=1$, and $v=v_p$ is the p-adic valuation, with $v(0)=\infty$. A vector $\vec{x}\in\mathbb{Z}_p^3$ is called *primitive* if $p\nmid\vec{x}$. The bad primes, and the integer D_F, were defined in §6.1. We write

$$b(q)=\prod_{p\mid D_F}p^{v_p(q)}$$

for the part of q composed of bad primes.

Here is the heuristic behind the local averages. Fix a good prime p. For most residues $\vec{a}$ modulo p^n the form $F_{\vec{a}}$ is non-singular modulo p, and $K(\vec{a},p^n)=1$. If $\Delta(\vec{a})\equiv 0\pmod{p}$, which happens for a proportion about $1/p$ of the residues, then $F_{\vec{a}}$ has corank 1 (or, rarely, 2) modulo p, and $K(\vec{a},p^n)$ can be as large as $p^{v(\Delta(\vec{a}))}$ (Lemma 8.4). So we need to know how often $\Delta(\vec{a})$ is divisible by p^j for large j. At a point of corank one the curve $\Delta=0$ is smooth, so Δ behaves like a linear function and is divisible by p^j with probability about p^{-j}; this exactly compensates the loss $p^{j/2}$ in $K^{1/2}$ with room to spare. At a point of corank two the curve has a double point, Δ behaves like a quadratic function, and p^j divides it with probability about $p^{-j/2}$, which is just enough. Lemma 8.6 makes these statements precise, and Lemma 8.5 is the tool for the quadratic case.

Lemma 8.4. *(i) Let p be good, $n\geqslant 1$, and $\vec{a}\in\mathbb{Z}^3$ with $p\nmid\vec{a}$. Then $K(\vec{a},p^n)\leqslant p^{\min(v(\Delta(\vec{a})),2n)}$.*

(ii) There is a constant $c_F\geqslant 1$ such that $K(\vec{a},q)\leqslant c_F\,b(q)\,q^2$ for all $q\in\mathbb{N}$ and all $\vec{a}$ with $\gcd(\vec{a},q)=1$. More precisely, for each bad prime p there is $\mu_p\geqslant 0$ with $K(\vec{a},p^n)\leqslant p^{\mu_p}p^{2n}$ whenever $p\nmid\vec{a}$.

Proof. In both parts the point is that at most two invariant factors of $M_{\vec{a}}$ can be divisible by high powers of p, because the corank of every member of the net is at most two. At a good prime the other invariant factors are units; at a bad prime they are bounded by a constant power of p, by compactness. Let $\rho_1 \mid \cdots \mid \rho_s$ be the invariant factors of $M_{\vec{a}}$ over $\mathbb{Z}_p$, so that $K(\vec{a}, p^n) = \prod_i \gcd(\rho_i, p^n)$ by Lemma 8.2(i).

(i) Since p is good and $p \nmid \vec{a}$, Lemma 6.1(i) over $\overline{\mathbb{F}}_p$ shows that $M_{\vec{a}}$ has corank at most 2 modulo p. Hence $\rho_1, \dots, \rho_{s-2}$ are units, and $K(\vec{a}, p^n) = \gcd(\rho_{s-1}, p^n)\gcd(\rho_s, p^n)$. This is at most p^{2n}, and, if $\Delta(\vec{a}) \neq 0$, at most $p^{v(\rho_{s-1}\rho_s)} = p^{v(\Delta(\vec{a}))}$, because $\rho_1 \cdots \rho_s$ and $\Delta(\vec{a})$ differ by a unit.

(ii) Let p be bad. By Lemma 6.1(i) over $\overline{\mathbb{Q}}_p$, for every primitive $\vec{x} \in \mathbb{Z}_p^3$ the matrix $M_{\vec{x}}$ has rank at least $s-2$, so some $(s-2) \times (s-2)$ minor of $M_{\vec{x}}$ is non-zero. The function $\vec{x} \mapsto \max |m(\vec{x})|_p$, the maximum taken over these minors, is continuous and positive on the compact set of primitive vectors, so it is bounded below by $p^{-\mu_p}$ for some μ_p. Then $v(\rho_1 \cdots \rho_{s-2}) \leqslant \mu_p$, and

$$K(\vec{a}, p^n) \leqslant p^{\mu_p} \gcd(\rho_{s-1}, p^n)\gcd(\rho_s, p^n) \leqslant p^{\mu_p + 2n}.$$

Combining this with (i) for the good primes and using multiplicativity gives (ii), with $c_F = \prod_{p \mid D_F} p^{\mu_p}$. □

The next lemma is a p-adic version of the fact that a quadratic polynomial in one variable is small only on a short interval.

Lemma 8.5. *Let $h(t) = \alpha + \beta t + t^2 + \sum_{i \geqslant 3} \delta_i t^i$ be a polynomial with $\alpha, \beta \in \mathbb{Q}_p$ and $\delta_i \in p\mathbb{Z}_p$ for $i \geqslant 3$. Then for every integer $m \geqslant 1$,*

$$\mu_1\{t \in \mathbb{Z}_p : \ v(h(t)) \geqslant m\} \leqslant 2p^{-m/2},$$

where μ_1 is the Haar probability measure on $\mathbb{Z}_p$.

Proof. Over $\mathbb{R}$, the set where $|t^2 + \beta t + \alpha| \leqslant \eta$ is contained in at most two intervals of length $O(\eta^{1/2})$, around the roots. The p-adic argument is the same: after factoring off a unit, h is a monic quadratic, and $v(h(t)) \geqslant m$ forces t to be within $p^{-m/2}$ of one of its two roots. The other cases, in which the constant or linear term dominates, are easier.

Suppose first that $\alpha, \beta \in \mathbb{Z}_p$. Then $h \in \mathbb{Z}_p[t]$, and its reduction modulo p is the monic quadratic $t^2 + \bar{\beta}t + \bar{\alpha}$. By the Weierstrass preparation theorem for $\mathbb{Z}_p[t]$ we may write $h(t) = (t^2 + \beta' t + \alpha')u(t)$ with $\alpha', \beta' \in \mathbb{Z}_p$ and $u \in \mathbb{Z}_p[t]$, $u \equiv 1 \pmod p$, so that $u(t)$ is a unit for $t \in \mathbb{Z}_p$. (Equivalently, by Hensel's lemma for factorisations: the reduction $\bar{h} = (t^2 + \bar{\beta}t + \bar{\alpha}) \cdot 1$ is a product of two coprime polynomials, the first monic of degree $\deg \bar{h} = 2$, so it lifts to a factorisation $h = gu$ over $\mathbb{Z}_p$ with g monic of degree 2 and $u \equiv 1 \pmod p$. For $t \in \mathbb{Z}_p$ we have $u(t) \equiv 1 \pmod p$, so $v(h(t)) = v(g(t))$.) Let r_1, r_2 be the roots of $t^2 + \beta' t + \alpha'$ in an extension of $\mathbb{Q}_p$ of degree at most 2, with the valuation extended. If $v(h(t)) \geqslant m$, then $v(t - r_1) + v(t - r_2) \geqslant m$, so $v(t - r_i) \geqslant m/2$ for some i. For each i the set $T_i = \{t \in \mathbb{Z}_p : \ v(t - r_i) \geqslant m/2\}$ is empty or contained in a ball: if $t_0 \in T_i$, then every $t \in T_i$ has $v(t - t_0) \geqslant m/2$, hence $v(t - t_0) \geqslant \lceil m/2 \rceil$. So $\mu_1(T_i) \leqslant p^{-\lceil m/2 \rceil}$, and the bound follows.

Suppose next that $v(\beta) < 0$ and $v(\beta) \leqslant v(\alpha)$. For $t, t' \in \mathbb{Z}_p$ we have $h(t) - h(t') = (t - t')(\beta + \gamma)$ with $v(\gamma) \geqslant 0$, so $v(h(t) - h(t')) = v(t - t') + v(\beta)$. If t and t' both satisfy $v(h) \geqslant m$, then $v(t - t') \geqslant m - v(\beta) > m$, so the set in question lies in a ball of measure at most p^{-m}.

In the remaining case $v(\alpha) < \min(0, v(\beta))$, and then $v(h(t)) = v(\alpha) < 0$ for all $t \in \mathbb{Z}_p$, so the set is empty. □

The following lemma describes how often Δ is highly divisible by p near a given point. It is the only place where the geometry of the net enters the local analysis.

Lemma 8.6. *Let p be good, $k \geqslant 1$, let $\vec{x}_0 \in \mathbb{Z}_p^3$ be primitive, and let $B = \vec{x}_0 + p^k\mathbb{Z}_p^3$. Let r be the corank of $M_{\vec{x}_0}$ modulo p. Then for every integer $j \geqslant 0$,*

$$\frac{\mu(B \cap \{v(\Delta) \geqslant j\})}{\mu(B)} \leqslant \begin{cases} \mathbf{1}_{j=0} & \text{if } r = 0, \\ \min(1, p^{k-j}) & \text{if } r = 1, \\ \min(1, 2p^{k-j/2}) & \text{if } r = 2. \end{cases}$$

Proof. We expand Δ around $\vec{x}_0$ in a Taylor series. If $r = 1$ the linear term has a unit coefficient (the point is a smooth point of the curve modulo p), and Hensel's lemma shows that Δ is equidistributed p-adically. If $r = 2$ the linear term vanishes modulo p but the quadratic term does not, by Lemma 6.1(v), and we restrict to lines on which the quadratic term is non-zero and apply Lemma 8.5. Write $\vec{x} = \vec{x}_0 + p^k\vec{y}$ with $\vec{y} \in \mathbb{Z}_p^3$, so that $\mu(B \cap E)/\mu(B)$ is the $\vec{y}$-measure of the corresponding set. If $r = 0$, then $\Delta(\vec{x}) \equiv \Delta(\vec{x}_0) \not\equiv 0 \pmod{p}$ on B.

Let $r = 1$. By Lemma 6.1(ii) over $\overline{\mathbb{F}}_p$, the gradient $\nabla\Delta(\vec{x}_0)$ is not divisible by p. Since Δ has integer coefficients,

$$\Delta(\vec{x}_0 + p^k\vec{y}) = \Delta(\vec{x}_0) + p^k\nabla\Delta(\vec{x}_0) \cdot \vec{y} + p^{2k}E(\vec{y})$$

with $E \in \mathbb{Z}_p[\vec{y}]$. If $v(\Delta(\vec{x}_0)) < k$, then $v(\Delta) = v(\Delta(\vec{x}_0))$ on B, and the bound holds because $j \leqslant v(\Delta(\vec{x}_0)) < k$ for the relevant j. Otherwise $\Delta(\vec{x}_0 + p^k\vec{y}) = p^kH(\vec{y})$, where $H(\vec{y}) = \Delta(\vec{x}_0)/p^k + \nabla\Delta(\vec{x}_0) \cdot \vec{y} + p^kE(\vec{y}) \in \mathbb{Z}_p[\vec{y}]$ has a reduction modulo p which is a non-constant affine-linear function. Choose a coordinate y_1 in which its coefficient is a unit. For fixed values of the other coordinates, $y_1 \mapsto H$ has exactly one zero modulo p, which is simple, so by Hensel's lemma it has exactly one zero modulo p^m for each $m \geqslant 1$. Hence the y_1-measure of $\{v(H) \geqslant m\}$ is p^{-m}, and integrating over the other coordinates gives $\mu\{v(\Delta) \geqslant j\}/\mu(B) = p^{-(j-k)}$ for $j \geqslant k$.

Let $r = 2$. Then every $(s-1) \times (s-1)$ minor of $M_{\vec{x}_0}$ vanishes modulo p, so $\operatorname{adj}(M_{\vec{x}_0}) \equiv 0$, and Jacobi's formula $\partial_i\Delta = \operatorname{tr}(\operatorname{adj}(M_{\vec{x}})M_i)$ shows that $\nabla\Delta(\vec{x}_0) \equiv \vec{0} \pmod{p}$. Let $\bar{Q}(\vec{b})$ be the coefficient of t^2 in $\bar{\Delta}(\vec{\bar{x}}_0 + t\vec{b})$, a quadratic form in $\vec{b} \in \mathbb{F}_p^3$. In a basis adapted to $\mathbb{F}_p^s = W \oplus W'$, with $W = \ker M_{\vec{\bar{x}}_0}$, we have $M_{\vec{\bar{x}}_0} = \operatorname{diag}(0, N)$ with N invertible, and expanding the determinant gives $\bar{Q}(\vec{b}) = \det N \cdot \det(M_{\vec{b}}|_W)$. By Lemma 6.1(v) over $\overline{\mathbb{F}}_p$ this is not the zero polynomial. A non-zero quadratic form over $\mathbb{F}_p$ with p odd does not vanish on all of $\mathbb{F}_p^3$, since it has at most $2p^2$ zeros, so we may choose $\vec{v} \in \mathbb{Z}_p^3$ with $\bar{Q}(\vec{\bar{v}}) \neq 0$. Complete $\vec{v}$ to a basis of $\mathbb{Z}_p^3$, and write $\vec{y} = \vec{y}_0 + t\vec{v}$, with $\vec{y}_0$ in the span of the other two basis vectors. For fixed $\vec{y}_0$ put $\vec{z} = \vec{x}_0 + p^k\vec{y}_0$ and

$$g(t) = \Delta(\vec{z} + p^kt\vec{v}) = \sum_{i=0}^{s} A_it^i.$$

The coefficients of the polynomial $t \mapsto \Delta(\vec{z} + t\vec{v})$ are integral, so $A_i \in p^{ik}\mathbb{Z}_p$. The coefficient of t^2 in $\Delta(\vec{z} + t\vec{v})$ is congruent modulo p to $\bar{Q}(\vec{\bar{v}})$, since $\vec{z} \equiv \vec{x}_0 \pmod{p}$; hence $v(A_2) = 2k$. Lemma 8.5, applied to $h = g/A_2$, whose coefficients $\delta_i = A_i/A_2$ for $i \geqslant 3$ have valuation at least $(i-2)k \geqslant 1$, shows that the t-measure of $\{v(g) \geqslant j\} = \{v(h) \geqslant j - 2k\}$ is at most $2p^{-(j-2k)/2}$ when $j > 2k$. Integrating over $\vec{y}_0$ gives the claim. □

We can now average $K(\vec{a},q)^{1/2}$. Part (i) of the next proposition is the average over all residues, which is what the point terms need, and part (ii) is the average over the residues on a plane $q \mid \vec{\nu}\cdot\vec{a}$, which is what the plane terms need. In part (ii) the discriminant restricted to the plane is a binary form, and the role of the condition $\vec{\nu}_0 \notin \mathcal{N}_F$ is to guarantee that its roots modulo p have multiplicity at most 2, so that the heuristic above still applies. Part (iii) collects the elementary facts we use to sum the factors $b(q)$ and $m_{\vec{\nu}_0}(q)$ over q. Recall the set $\mathcal{N}_F$ of exceptional directions and the integers $R_{\vec{\nu}_0}$ from §6.1. For a primitive $\vec{\nu}_0 \notin \mathcal{N}_F$ we put

$$m_{\vec{\nu}_0}(q) = \prod_{p \mid D_F R_{\vec{\nu}_0}} p^{v_p(q)}.$$

Proposition 8.7. *(i) For every $q \in \mathbb{N}$ we have*

$$\sideset{}{^*}\sum_{\vec{a} \bmod q} K(\vec{a},q)^{1/2} \ll_\varepsilon q^{3+\varepsilon} b(q).$$

(ii) Let $\vec{\nu} = g\vec{\nu}_0$ with $g \in \mathbb{N}$ and $\vec{\nu}_0 \notin \mathcal{N}_F$ primitive. Then

$$\sideset{}{^*}\sum_{\substack{\vec{a} \bmod q \\ q \mid \vec{\nu}\cdot\vec{a}}} K(\vec{a},q)^{1/2} \ll_\varepsilon q^{2+\varepsilon} \gcd(q,g)\, m_{\vec{\nu}_0}(q).$$

(iii) For $Y \geqslant 1$, every non-zero integer N and every integer $R \neq 0$ we have

$$\sum_{q \leqslant Y} \gcd(N,q) \leqslant Y\tau(N), \qquad \#\{m \leqslant Y : \ m \mid R^\infty\} \ll_\varepsilon (Y|R|)^\varepsilon,$$

and consequently $\sum_{q\leqslant Y} b(q) \ll_\varepsilon Y^{1+\varepsilon}$ and

$$\sum_{q \leqslant Y} m_{\vec{\nu}_0}(q) \ll_\varepsilon Y^{1+\varepsilon} |\vec{\nu}_0|^\varepsilon.$$

Proof. (iii) The first bound follows from

$$\sum_{q\leqslant Y} \gcd(N,q) \leqslant \sum_{d \mid N} d \cdot \#\{q \leqslant Y : \ d \mid q\} \leqslant Y\tau(N).$$

For the second, Rankin's trick gives, for $0 < \eta \leqslant 1/2$,

$$\#\{m \leqslant Y : \ m \mid R^\infty\} \leqslant Y^\eta \sum_{m \mid R^\infty} m^{-\eta} = Y^\eta \prod_{p \mid R} (1-p^{-\eta})^{-1} \leqslant Y^\eta C_\eta^{\omega(R)},$$

and $C_\eta^{\omega(R)} \ll_{\eta,\varepsilon} |R|^\varepsilon$ because $\omega(R) \ll \log|R| / \log\log|R|$. Finally, if $m(q)$ denotes the part of q composed of the primes dividing R, then

$$\sum_{q\leqslant Y} m(q) = \sum_{\substack{m \leqslant Y \\ m \mid R^\infty}} m \cdot \#\{q \leqslant Y : \ m(q) = m\} \leqslant Y \cdot \#\{m \leqslant Y : \ m \mid R^\infty\},$$

which gives the last two bounds, since $|R_{\vec{\nu}_0}| \ll |\vec{\nu}_0|^C$.

(i) By Lemma 8.2(i) and the Chinese remainder theorem, both sides are multiplicative in q, so it suffices to consider $q = p^n$. The plan is to cover the primitive vectors by the p^3-1 balls of radius p^{-1}, sort the balls according to the corank $r \in \{0,1,2\}$ of their centre modulo p, and apply Lemma 8.6 on each ball. The balls with $r = 0$ contribute the main term 1; there are $O(p^2)$ balls with $r = 1$, each contributing $O(p^{-5/2})$; and $O(p)$ balls with $r = 2$, each contributing $O(np^{-2})$. If p is bad, Lemma 8.4(ii) gives the bound $p^{\mu_p/2}p^{4n}$. Let p be good. Since $K(\vec{a},p^n)$ depends only on $\vec{a}$ modulo p^n,

$$\sideset{}{^*}\sum_{\vec{a} \bmod p^n} K(\vec{a},p^n)^{1/2} = p^{3n} \int_{\vec{x} \text{ primitive}} K(\vec{x},p^n)^{1/2}\, d\mu(\vec{x}).$$

We split the set of primitive $\vec{x}$ into the p^3-1 balls $B=\vec{x}_0+p\mathbb{Z}_p^3$, each of measure p^{-3}. On each ball we use Lemma 8.4(i) in the form

$$\int_B K^{1/2}\,d\mu \leqslant \int_B p^{\min(v(\Delta),2n)/2}\,d\mu \leqslant \sum_{j=0}^{2n} p^{j/2}\mu\Big(B\cap\{v(\Delta)\geqslant j\}\Big),$$

and Lemma 8.6 with $k=1$. The balls with $r=0$ contribute at most 1 in total. On a ball with $r=1$ the right side is at most $p^{-3}(1+p^{1/2}+\sum_{j\geqslant 2}p^{1-j/2})\leqslant 6p^{-5/2}$. The number of such balls is at most the number of non-zero zeros of $\bar{\Delta}$ in $\mathbb{F}_p^3$, which is at most sp^2, since $\bar{\Delta}$ is a non-zero form of degree s (recall that $p\nmid D_F$, so p does not divide the content of Δ). On a ball with $r=2$ the right side is at most

$$p^{-3}\Big(1+p^{1/2}+p+\sum_{j=3}^{2n}2p\Big)\leqslant 7(n+1)p^{-2}.$$

The vectors $\bar{\vec{x}}_0$ of corank 2 lie on the lines through the points of $\mathcal{Z}_2$ over $\mathbb{F}_p$. These points are the common zeros in $\mathbb{P}^2(\overline{\mathbb{F}}_p)$ of the $(s-1)\times(s-1)$ minors of $M_{\vec{\beta}}$, which are forms of degree $s-1$. The set is finite by Lemma 6.1(iii), so two generic linear combinations of the minors have no common component, and Bézout's theorem bounds the number of points by $(s-1)^2$. So there are at most $(s-1)^2(p-1)$ such balls. Altogether

$$\sum_{\vec{a}\bmod p^n}^{*} K(\vec{a},p^n)^{1/2}\leqslant p^{3n}\Big(1+6sp^{-1/2}+7s^2(n+1)p^{-1}\Big)\leqslant p^{3n}(n+1)(1+13s^2p^{-1/2}).$$

The product of $(n_p+1)(1+13s^2p^{-1/2})$ over $p^{n_p}\parallel q$ is at most

$$\tau(q)\prod_{p\mid q}(1+13s^2p^{-1/2})\ll_\varepsilon q^{\varepsilon},$$

which proves (i).

(ii) Again both sides are multiplicative, and we take $q=p^n$. The structure is the same as in (i), but now the residues lie on the plane $\vec{\nu}_0\cdot\vec{x}\equiv 0 \pmod{p^{n-f}}$, so we must understand Δ on a thin neighbourhood of a plane rather than on balls. After reducing modulo p, everything is governed by the binary form $\bar{\Delta}|_{\bar{\vec{\nu}}_0^{\perp}}$, and the cases below correspond to the multiplicity m of its root at the centre $\bar{\vec{x}}_0$ and the corank r of $M_{\vec{x}_0}$. Let $p^f=\gcd(p^n,g)$, so that $p^n\mid\vec{\nu}\cdot\vec{a}$ if and only if $p^{n-f}\mid\vec{\nu}_0\cdot\vec{a}$. The number of residues in question is at most $p^{3n}\cdot p^{-(n-f)}=p^{2n+f}$. If $p\mid D_F R_{\vec{\nu}_0}$, we use $K^{1/2}\leqslant c_F^{1/2}p^n$ from Lemma 8.4. If $f=n$ the condition is empty and (i) applies. So let $p\nmid D_F R_{\vec{\nu}_0}$ and $n-f\geqslant 1$, and put $\tau_0=n-f-1\geqslant 0$. As in (i) the sum equals $p^{3n}\int_A K^{1/2}d\mu$, where $A=\{\vec{x}\text{ primitive}:\ v(\vec{\nu}_0\cdot\vec{x})\geqslant n-f\}$ has measure at most $p^{-(n-f)}$.

Every $\vec{x}\in A$ reduces to a non-zero vector of the plane $\bar{\vec{\nu}}_0^{\perp}\subset\mathbb{F}_p^3$. Fix such a vector, and choose a lift $\vec{x}_0$ of it in the lattice $\vec{\nu}_0^{\perp}\cap\mathbb{Z}_p^3$; this is possible because $\vec{\nu}_0$ is primitive, so that this lattice reduces onto $\bar{\vec{\nu}}_0^{\perp}$. Since $\vec{x}_0$ is primitive, we can choose $\vec{w}$ such that $\vec{x}_0,\vec{w}$ is a basis of $\vec{\nu}_0^{\perp}\cap\mathbb{Z}_p^3$, and $\vec{w}_3$ with $\vec{\nu}_0\cdot\vec{w}_3=1$; then $\vec{x}_0,\vec{w},\vec{w}_3$ is a basis of $\mathbb{Z}_p^3$. The part of A in the ball $\vec{x}_0+p\mathbb{Z}_p^3$ is

$$A_{\vec{x}_0}=\{\vec{x}_0+p(y_2\vec{x}_0+y_1\vec{w}+t\vec{w}_3):\ y_1,y_2\in\mathbb{Z}_p,\ v(t)\geqslant\tau_0\},$$

of measure $p^{-3}p^{-\tau_0}=p^{-2}p^{-(n-f)}$. Since Δ is homogeneous of degree s,

$$\Delta\Big(\vec{x}_0+p(y_2\vec{x}_0+y_1\vec{w}+t\vec{w}_3)\Big)=(1+py_2)^s\,\Delta\Big(\vec{x}_0+p(y_1'\vec{w}+t'\vec{w}_3)\Big)$$

with $y_1' = y_1/(1+py_2)$ and $t' = t/(1+py_2)$. For fixed y_2, the map $(y_1,t)\mapsto(y_1',t')$ preserves measure and the condition $v(t)\geqslant\tau_0$, so it suffices to study

$$G(y_1,t) = \Delta(\vec{x}_0 + py_1\vec{w} + pt\vec{w}_3) = \sum_{i,l\geqslant 0} p^{i+l}c_{il}\,y_1^i t^l, \qquad c_{il}\in\mathbb{Z}_p,$$

on $\mathbb{Z}_p\times p^{\tau_0}\mathbb{Z}_p$. The binary form $\Delta|_{\ell_{\vec{\nu}_0}}$ in the basis $\vec{x}_0,\vec{w}$ is congruent modulo p to the restriction of $\bar{\Delta}$ to $\bar{\vec{\nu}}_0^{\perp}$. By the choice of $R_{\vec{\nu}_0}$ (see §6.1) the latter is non-zero and all its roots have multiplicity at most 2. Let $m\in\{0,1,2\}$ be the multiplicity at $\bar{\vec{x}}_0$. Then $c_{i0}\equiv 0 \pmod p$ for $i<m$ and $c_{m0}\not\equiv 0 \pmod p$. We distinguish four cases.

Case $m=0$. Then Δ is a unit on $A_{\vec{x}_0}$, and $K=1$ there.

Case $m=1$. Then c_{10} is a unit, and for fixed t the function $y_1\mapsto G/p$ has a reduction modulo p which is affine-linear with unit slope. By Hensel's lemma the y_1-measure of $\{v(G)\geqslant j\}$ is at most $p^{-(j-1)}$.

Case $m=2$ and $r=1$, where r is the corank of $M_{\vec{x}_0}$ modulo p. Then $\nabla\bar{\Delta}(\bar{\vec{x}}_0)\neq\vec{0}$ by Lemma 6.1(ii). It is orthogonal to $\bar{\vec{x}}_0$ by Euler's identity and to $\bar{\vec{w}}$ because $m=2$, so $c_{01}=\nabla\Delta(\vec{x}_0)\cdot\vec{w}_3$ is a unit. Write $t=p^{\tau_0}\tau$. Then

$$G(y_1,p^{\tau_0}\tau) = G_0(y_1) + p^{1+\tau_0}\tau\,U(y_1,\tau), \qquad G_0(y_1)=G(y_1,0),$$

where $U\in\mathbb{Z}_p[y_1,\tau]$ and $U\equiv c_{01} \pmod p$. Here $G_0(y_1)=\sum_i p^i c_{i0}y_1^i$ with $c_{00}\equiv c_{10}\equiv 0$ and c_{20} a unit. Applying Lemma 8.5 to $G_0/(p^2c_{20})$ gives $\mu_1\{v(G_0)\geqslant i\}\leqslant \min(1,2p^{1-i/2})$ for all i. For $j\leqslant 1+\tau_0$ the condition $v(G)\geqslant j$ is equivalent to $v(G_0)\geqslant j$. For $j>1+\tau_0$ it requires $v(G_0)\geqslant 1+\tau_0$, and then, for fixed y_1, the condition on τ is that $G_0/p^{1+\tau_0}+\tau U$ has valuation at least $j-1-\tau_0$; by Hensel's lemma this has τ-measure $p^{-(j-1-\tau_0)}$. Hence the relative measure of $\{v(\Delta)\geqslant j\}$ in $A_{\vec{x}_0}$ is at most $\min(1,2p^{1-j/2})$ for $j\leqslant 1+\tau_0$, and at most $2p^{1-(1+\tau_0)/2}p^{-(j-1-\tau_0)}$ for $j>1+\tau_0$.

Case $m=2$ and $r=2$. Now $\nabla\Delta(\vec{x}_0)\equiv\vec{0}$, so $c_{01}\equiv c_{10}\equiv 0 \pmod p$, while c_{20} is a unit. For fixed $t\in p^{\tau_0}\mathbb{Z}_p$, the coefficient of y_1^2 in G is $p^2(c_{20}+O(p))$, of valuation exactly 2, and the coefficient of y_1^i for $i\geqslant 3$ is divisible by p^i. Lemma 8.5 shows that the y_1-measure of $\{v(G)\geqslant j\}$ is at most $\min(1,2p^{1-j/2})$, uniformly in t.

In every case

$$\frac{1}{\mu(A_{\vec{x}_0})}\int_{A_{\vec{x}_0}} K^{1/2}\,d\mu \leqslant \sum_{j=0}^{2n} p^{j/2}\,\frac{\mu(A_{\vec{x}_0}\cap\{v(\Delta)\geqslant j\})}{\mu(A_{\vec{x}_0})} \leqslant C(n+1)p,$$

and it is 1 when $m=0$. In the third case this uses

$$\sum_{j>1+\tau_0} p^{j/2}\cdot 2p^{1-(1+\tau_0)/2}p^{-(j-1-\tau_0)} = 2p\sum_{i\geqslant 1}p^{-i/2}\leqslant 6p.$$

There are at most p^2 vectors $\bar{\vec{x}}_0$, and at most $s(p-1)$ of them have $m\geqslant 1$, because $\bar{\Delta}|_{\bar{\vec{\nu}}_0^{\perp}}$ is a non-zero binary form of degree s. Since $\mu(A_{\vec{x}_0})=p^{-2}\mu(A)$ up to a bounded factor, we obtain

$$\int_A K^{1/2}\,d\mu \leqslant \mu(A)\big(1+Cs(n+1)\big),$$

and hence the sum is at most $p^{2n+f}(1+Cs(n+1))$. Taking the product over the primes dividing q proves (ii). □

The proof shows where the exceptional directions come from. If the plane $\bar{\vec{\nu}}_0^{\perp}$ meets the discriminant curve with multiplicity $m\geqslant 3$ at some point, then the residues near that point give a contribution of size about $p^{n(1-2/m)}$ relative to the expected one.

This is why such planes, which are finite in number over $\overline{\mathbb{Q}}$, are treated separately in Section 12, and why the finitely many primes dividing $R_{\vec{\nu}_0}$ are handled crudely.

9. The main term

In this section we prove Lemma 6.7(a). This is the easy part of the paper. Once the main-term property of the delta symbol is available, the contribution of the small moduli and the zero dual vector factorises as a sum over q, which is the singular series, times an integral, which is the singular integral. The work in this section is to show that the series converges fast enough, and to interpret both factors as local densities. We define the singular series by

$$\mathfrak{S} = \sum_{q=1}^{\infty} q^{-s} \sideset{}{^*}\sum_{\vec{a} \bmod q} S_q(\vec{a}, \mathbf{0}). \tag{9.1}$$

By Lemma 8.1 and Proposition 8.7(i),

$$q^{-s} \sideset{}{^*}\sum_{\vec{a} \bmod q} |S_q(\vec{a}, \mathbf{0})| \leqslant q^{-s/2} \sideset{}{^*}\sum_{\vec{a} \bmod q} K(\vec{a}, q)^{1/2} \ll q^{3-s/2+\varepsilon} b(q). \tag{9.2}$$

Since $b(q) \leqslant q$, the series converges absolutely for $s \geqslant 11$. Moreover, by Proposition 8.7(iii) and partial summation over dyadic ranges,

$$\sum_{q \geqslant Y} q^{3-s/2+\varepsilon} b(q) \ll_{\varepsilon} Y^{4-s/2+2\varepsilon} \qquad (s \geqslant 11). \tag{9.3}$$

Lemma 9.1. *Let $s \geqslant 11$.*

(i) We have $\mathfrak{S} = \prod_p \sigma_p$, where

$$\sigma_p = \lim_{e \to \infty} p^{-e(s-3)} \#\{\mathbf{x} \bmod p^e : \ \vec{F}(\mathbf{x}) \equiv \vec{0} \pmod{p^e}\}.$$

The product converges absolutely, and $\sigma_p = 1 + O(p^{3-s/2+\varepsilon})$.

(ii) If the variety $F_1 = F_2 = F_3 = 0$ has a point over $\mathbb{Q}_p$, then $\sigma_p > 0$. Hence $\mathfrak{S} > 0$ if it has points over every $\mathbb{Q}_p$.

(iii) If $w \geqslant 0$ and $w(\mathbf{z}_0) > 0$, then $\mathfrak{J} > 0$.

Proof. Part (i) is the usual argument: the terms of the series are multiplicative in q, and the sum over the powers of a fixed prime counts solutions modulo prime powers. Part (ii) lifts a non-singular p-adic point to many solutions modulo p^e by Hensel's lemma, and part (iii) is the analogous real statement.

(i) Let $A(q) = q^{-s} \sideset{}{^*}\sum_{\vec{a} \bmod q} S_q(\vec{a}, \mathbf{0})$. By the Chinese remainder theorem A is multiplicative, and by orthogonality of characters

$$\sum_{e \leqslant E} A(p^e) = p^{-E(s-3)} \#\{\mathbf{x} \bmod p^E : \ \vec{F}(\mathbf{x}) \equiv \vec{0}\}.$$

For a good prime p we have $b(p^e) = 1$, so (9.2) gives

$$\sum_{e \geqslant 1} |A(p^e)| \ll \sum_{e \geqslant 1} p^{e(3-s/2+\varepsilon)} \ll p^{3-s/2+\varepsilon}.$$

For each of the finitely many bad primes, $\sum_e |A(p^e)| \ll \sum_e p^{e(4-s/2+\varepsilon)}$ converges. This gives the claims.

(ii) Since the variety is non-singular, a $\mathbb{Q}_p$-point can be scaled to a primitive $\mathbf{x}_0 \in \mathbb{Z}_p^s$ with $\vec{F}(\mathbf{x}_0) = \vec{0}$ at which the Jacobian matrix $J(\mathbf{x}_0)$ has rank 3. Let μ be the smallest valuation of a 3×3 minor of $J(\mathbf{x}_0)$, put $k = 2\mu + 1$, and renumber the variables so

that the minor formed by the first three columns has valuation μ. Write $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ with $\mathbf{y} \in \mathbb{Z}_p^3$, $\mathbf{z} \in \mathbb{Z}_p^{s-3}$, and $\mathbf{x}_0 = (\mathbf{y}_0, \mathbf{z}_0)$. For $\mathbf{z} \equiv \mathbf{z}_0 \pmod{p^k}$ the map $\mathbf{y} \mapsto \vec{F}(\mathbf{y}, \mathbf{z})$ takes at $\mathbf{y}_0$ a value of valuation at least $k > 2\mu$, and its Jacobian determinant at $\mathbf{y}_0$ is congruent modulo p^k to the chosen minor, so it has valuation exactly μ. By the multivariable Hensel lemma there is $\mathbf{y} \in \mathbb{Z}_p^3$ with $\vec{F}(\mathbf{y}, \mathbf{z}) = \vec{0}$. Hence for $e \geqslant k$ each of the $p^{(s-3)(e-k)}$ classes $\mathbf{z} \bmod p^e$ with $\mathbf{z} \equiv \mathbf{z}_0 \pmod{p^k}$ is the $\mathbf{z}$-part of a solution of $\vec{F}(\mathbf{x}) \equiv \vec{0} \pmod{p^e}$, so

$$p^{-e(s-3)} \#\{\mathbf{x} \bmod p^e : \ \vec{F}(\mathbf{x}) \equiv \vec{0}\} \geqslant p^{-(s-3)k}$$

and $\sigma_p \geqslant p^{-(s-3)(2\mu+1)} > 0$. The product $\prod_p \sigma_p$ is then positive, since by (i) it converges absolutely and none of its factors vanishes.

(iii) The singular integral $\mathfrak{J} = \sigma(\vec{0})$ is the integral of $w/\sqrt{\det(JJ^T)}$ over the manifold $\{\vec{F} = \vec{0}\} \cap U$, which passes through $\mathbf{z}_0$. This manifold is smooth of dimension $s-3$, because J has rank 3 on U; the integrand is continuous and non-negative, and positive at $\mathbf{z}_0$, hence positive on a neighbourhood of $\mathbf{z}_0$ in the manifold, which has positive $\mathcal{H}^{s-3}$-measure. So $\mathfrak{J} > 0$. □

Proof of Lemma 6.7(a). We use the main-term property to evaluate each integral in N_0, which leaves a truncated singular series; the truncation error is controlled by the tail bound (9.3). By Lemma 7.2, for $q < q_0$ each integral in $N_0(P)$ equals $P^{s-6}(\mathfrak{J} + O_N(P^{-N}))$. Hence

$$N_0(P) = \mathfrak{J} P^{s-6} \sum_{q<q_0} A(q) + O_N(P^{-N}).$$

By (9.3) the tail $\sum_{q \geqslant q_0} |A(q)|$ is $\ll q_0^{4-s/2+2\varepsilon}$. Since $q_0 = Q^{1/3-\delta}$ and $Q \asymp P^{3/2}$, this is

$$\ll P^{2-s/4+C\delta+\varepsilon} = P^{-(s-8)/4+C\delta+\varepsilon}.$$

Taking δ small enough gives the result. □

Note that the main term comes only from the moduli $q < Q^{1/3-\delta}$, where (1.5) applies. The remaining terms with $\mathbf{u} = \mathbf{0}$ belong to $N_1(P)$ and $N_2(P)$, and they are small because the singular series converges quickly; see the end of the proof of Proposition 11.2.

10. Averages over the arcs

In this section we prepare the estimates for $N_1(P)$ and $N_2(P)$. We need a simple count of lattice points in cosets, and an average of $K(\vec{a}, q)^{1/2}$ against the weights $p^{\mathrm{gen}}_{\Lambda(\vec{a},q)}$.

The logic of the error terms is as follows. For each modulus q, numerator $\vec{a}$ and point $\vec{w}$, the contribution of the dual vectors $\mathbf{u}$ is bounded by summing $|S_q(\vec{a}, \mathbf{u})|\,|I_q(\vec{w}, \mathbf{u})|$ over $|\mathbf{u}| \ll qX/P$. The integral is bounded by Lemma 7.1, and the sum over $\mathbf{u}$ of the Gauss sums is bounded by Lemma 10.2 below, in terms of $K(\vec{a}, q)$. What remains is a sum of $K(\vec{a}, q)^{1/2}$ weighted by the mass of the weights $p_{\Lambda(\vec{a},q)}$, and Proposition 10.7, the main result of this section, shows that this weighted average is no larger than the mass itself, up to Q^{ε}. The lemmas before it are the local and arithmetic ingredients.

Lemma 10.1. *Let $\Lambda' \subseteq \mathbb{Z}^s$ be a lattice of index I, and let $\mathbf{v} + \Lambda'$ be a coset. For $L \geqslant 1$ we have*

$$\#\{\mathbf{u} \in \mathbf{v} + \Lambda' : \ |\mathbf{u}|_\infty \leqslant L\} \ll \frac{L^s}{I} + L^{s-1}.$$

Proof. The bound is the volume term L^s/I plus a boundary term, which can only be large if the lattice is very lopsided. Let $\mu_1 \leqslant \cdots \leqslant \mu_s$ be the successive minima of Λ' with respect to the sup norm. Then $\mu_1 \geqslant 1$ and $\mu_1 \cdots \mu_s \asymp I$, and the number of points of a coset in a cube of side $2L$ is $\ll \prod_i (1 + L/\mu_i)$. If $\mu_s \leqslant L$ this is $\ll L^s/I$, and otherwise it is $\ll \prod_{i<s}(1 + L/\mu_i) \ll L^{s-1}$. □

The next lemma combines Lemma 10.1 with the exact formula of Lemma 8.1. It shows that a large value of $K(\vec{a}, q)$ does no harm as long as $K(\vec{a}, q)^{1/2}$ does not exceed the length L of the range of dual vectors. Indeed, a large K makes each non-zero Gauss sum larger by $K^{1/2}$, but it also makes the non-zero ones rarer by the factor K; the two effects combine to a gain $K^{-1/2}$, until the sublattice becomes so sparse that the box of side L contains at most one of its points on each line.

Lemma 10.2. *Let $q \in \mathbb{N}$, $\vec{a} \in \mathbb{Z}^3$ with $\gcd(\vec{a}, q) = 1$, and $L \geqslant 1$. Then*

$$\sum_{\substack{\mathbf{u} \in \mathbb{Z}^s \\ |\mathbf{u}|_\infty \leqslant L}} |S_q(\vec{a}, \mathbf{u})| \ll q^{s/2} L^s \Big(K(\vec{a}, q)^{-1/2} + \frac{K(\vec{a}, q)^{1/2}}{L} \Big).$$

Proof. By Lemma 8.1, $|S_q(\vec{a}, \mathbf{u})| = q^{s/2} K(\vec{a}, q)^{1/2}$ if $\mathbf{u}$ lies in the preimage of $\mathcal{M}(\vec{a}, q)$ in $\mathbb{Z}^s$, and $S_q(\vec{a}, \mathbf{u}) = 0$ otherwise. This preimage is empty or a coset of a lattice of index $K(\vec{a}, q)$, so by Lemma 10.1 it contains $\ll L^s/K(\vec{a}, q) + L^{s-1}$ points with $|\mathbf{u}|_\infty \leqslant L$. □

We also record a local average over cosets, which is what the flag terms require. For a flag term the numerators $\vec{a}$ are not spread over all residues, but lie close to a fixed direction $\lambda\vec{c}$ modulo a divisor of q; so we need the average of $K^{1/2}$ over a p-adic ball centred on the line through $\vec{c}$. Lemma 8.6 provides it.

Lemma 10.3. *Let p be good, $\delta \geqslant 1$, $1 \leqslant k \leqslant \delta$, and let $\vec{c} \in \mathbb{Z}^3$ be primitive (we allow $\Delta(\vec{c}) = 0$, with the convention $v(0) = \infty$). For every unit λ, the average of $K(\vec{x}, p^\delta)^{1/2}$ over the residues $\vec{x} \bmod p^\delta$ with $\vec{x} \equiv \lambda\vec{c} \pmod{p^k}$ is at most $7(\delta + 1)p^{\min(k, v(\Delta(\vec{c})))}$.*

Proof. If $\Delta(\vec{c})$ is divisible by p only to a small power, then Δ is constant in valuation on the whole ball and there is nothing to average; otherwise Lemma 8.6 controls how often Δ is highly divisible on the ball. The average equals $\mu(B)^{-1} \int_B K(\vec{x}, p^\delta)^{1/2} d\mu$ with $B = \lambda\vec{c} + p^k\mathbb{Z}_p^3$. If $v(\Delta(\vec{c})) < k$, then $\Delta(\vec{x}) \equiv \lambda^s \Delta(\vec{c}) \pmod{p^k}$ shows $v(\Delta(\vec{x})) = v(\Delta(\vec{c}))$ on B, and Lemma 8.4(i) gives $K^{1/2} \leqslant p^{v(\Delta(\vec{c}))/2}$. Otherwise Lemmas 8.4(i) and 8.6 give

$$\frac{1}{\mu(B)} \int_B K^{1/2} d\mu \leqslant \sum_{j=0}^{2\delta} p^{j/2} \min(1, 2p^{k-j/2}) \leqslant \sum_{j \leqslant 2k} p^{j/2} + \sum_{2k < j \leqslant 2\delta} 2p^k \leqslant 7(\delta + 1)p^k. \quad \square$$

We next collect some elementary sums. For $x \in \mathbb{N}$ we write $x_g = x/b(x)$ for the part of x composed of good primes. They express the familiar fact that a gcd with a fixed integer is 1 on average, up to a divisor function.

Lemma 10.4. *Let A, B be non-zero integers, $W, Z \geqslant 1$ and $0 \leqslant \beta \leqslant 1$. Then*

(i)

$$\sum_{x \leqslant W} \gcd(x, A) \gcd(x, B) \leqslant W\tau(A)\tau(B) \gcd(A, B);$$

(ii)

$$\sum_{x \leqslant W} b(x) \gcd(x_g, A) \gcd(x_g, B) \ll_\varepsilon W^{1+\varepsilon} \tau(A)\tau(B) \gcd(A, B);$$

(iii)

$$\sum_{G\leqslant Z}\sum_{d\mid G} d^{\beta} b(d)\gcd(d_g, A) \ll_{\varepsilon} Z^{1+\beta+\varepsilon}\tau(A).$$

Proof. All three parts follow by writing a gcd as a sum of $\varphi(m)$ over common divisors and interchanging the order of summation. (i) We have $\gcd(x, A) = \sum_{m\mid \gcd(x,A)} \varphi(m)$, and similarly for B. Hence the sum is

$$\sum_{m\mid A}\sum_{n\mid B}\varphi(m)\varphi(n)\,\#\{x\leqslant W:\ \operatorname{lcm}(m,n)\mid x\} \leqslant W\sum_{m\mid A}\sum_{n\mid B}\frac{mn}{\operatorname{lcm}(m,n)},$$

and $mn/\operatorname{lcm}(m,n) = \gcd(m,n) \leqslant \gcd(A,B)$.

(ii) Write $x = x_b x_g$, where $x_b = b(x)$. By (i) the sum is at most

$$\sum_{\substack{x_b\leqslant W\\ x_b\mid D_F^{\infty}}} x_b\cdot\frac{W}{x_b}\tau(A)\tau(B)\gcd(A,B) = W\tau(A)\tau(B)\gcd(A,B)\,\#\{x_b\leqslant W:\ x_b\mid D_F^{\infty}\},$$

and the last factor is $\ll_{\varepsilon} W^{\varepsilon}$ by Proposition 8.7(iii).

(iii) The sum equals

$$\sum_{d\leqslant Z} d^{\beta} b(d)\gcd(d_g, A)\lfloor Z/d\rfloor \leqslant Z\sum_{d\leqslant Z} d^{\beta-1} b(d)\gcd(d_g, A).$$

By (ii) with $B = 1$ the partial sums of $b(d)\gcd(d_g, A)$ up to t are $\ll t^{1+\varepsilon}\tau(A)$, and partial summation gives

$$\sum_{d\leqslant Z} d^{\beta-1} b(d)\gcd(d_g, A) \ll Z^{\beta+\varepsilon}\tau(A).$$

□

We also need a count of the vectors $\vec{\nu}$ attached to a given direction $\vec{c}$. For a flag $(\vec{\nu}, \vec{\rho})$ with $\vec{\rho} = G\vec{c}$, the vector $\vec{\nu}$ must lie in the plane $\vec{c}^{\perp}$, so the number of flags with a given direction depends on the shape of the lattice $\vec{c}^{\perp}\cap\mathbb{Z}^3$. Most planes are "round", but some contain an unusually short vector, and part (ii) shows that these are rare enough. For a primitive $\vec{c}\in\mathbb{Z}^3$ let $\lambda_1(\vec{c})$ be the length of the shortest non-zero vector of the lattice $\vec{c}^{\perp}\cap\mathbb{Z}^3$, which has rank 2 and determinant $|\vec{c}|$.

Lemma 10.5. *(i) For a primitive $\vec{c}$ and $U\geqslant 1$, the number $\nu(\vec{c}; U)$ of non-zero $\vec{\nu}\in\mathbb{Z}^3$ with $\vec{\nu}\cdot\vec{c} = 0$ and $|\vec{\nu}|\leqslant U$ satisfies $\nu(\vec{c}; U)\ll U/\lambda_1(\vec{c}) + U^2/|\vec{c}|$.*

(ii) For $R\geqslant 1$ we have

$$\sum_{\vec{c}\ \text{primitive},\ |\vec{c}|\leqslant R}\lambda_1(\vec{c})^{-1} \ll R^{5/2}.$$

(iii) For $V\geqslant 1$ we have

$$\sum_{\vec{c}\ \text{primitive},\ |\vec{c}|\leqslant V}\nu(\vec{c}; Q^{1/3})|\vec{c}|^{-2} \ll Q^{1/3}V^{1/2} + Q^{2/3}\log(2V).$$

Proof. Part (ii) is the only one which needs an idea: if the plane $\vec{c}^{\perp}$ contains a short vector $\vec{y}$, then $\vec{c}$ lies in the plane $\vec{y}^{\perp}$; so we group the directions $\vec{c}$ according to $\vec{y}$, and count the $\vec{c}$ in each plane $\vec{y}^{\perp}$ by part (i). (i) This is the standard bound for the number of non-zero points of a lattice of rank 2 in a disc, together with $\lambda_1(\vec{c})\lambda_2(\vec{c})\asymp|\vec{c}|$.

(ii) By Minkowski's theorem $\lambda_1(\vec{c}) \ll |\vec{c}|^{1/2}$. For each $\vec{c}$ choose a primitive $\vec{y} \perp \vec{c}$ with $|\vec{y}| = \lambda_1(\vec{c})$. Grouping the $\vec{c}$ according to $\vec{y}$, and using (i) in the lattice $\vec{y}^{\perp} \cap \mathbb{Z}^3$ with $\lambda_1 \geqslant 1$, we get

$$\sum_{|\vec{c}|\leqslant R} \lambda_1(\vec{c})^{-1} \leqslant \sum_{0<|\vec{y}|\ll R^{1/2}} \frac{1}{|\vec{y}|}\,\nu(\vec{y};R) \ll \sum_{0<|\vec{y}|\ll R^{1/2}} \Big(\frac{R}{|\vec{y}|} + \frac{R^2}{|\vec{y}|^2}\Big) \ll R^2 + R^{5/2}.$$

(iii) By (i), $\nu(\vec{c};Q^{1/3}) \ll Q^{1/3}/\lambda_1(\vec{c}) + Q^{2/3}/|\vec{c}|$. On a dyadic range $|\vec{c}| \asymp R$ the first term contributes $\ll Q^{1/3}R^{-2}R^{5/2}$ by (ii), and the second $\ll Q^{2/3}R^{-3}R^3$. Summing over the dyadic ranges $R \leqslant V$ gives the bound. □

The next lemma is the local input for the flag terms. It bounds the sum of $K(\vec{a},q)^{1/2}$ over the numerators attached to one flag. The answer depends on the direction $\vec{c}$ only through the quantity $E_{\vec{c}}(q_1)$, which is $\gcd(\Delta(\vec{c}),q)$ in essence. For non-singular $\vec{c}$ this is 1 on average over q, while for singular $\vec{c}$ it is as large as a power of q; this is the basic reason why the singular directions have to be treated separately. Let $(\vec{\nu},\vec{\rho})$ be a flag with $\vec{\rho} = G\vec{c}$, $\vec{c}$ primitive, let $q \in \mathbb{N}$, and put

$$d = \gcd(q,G), \qquad q_1 = q/d, \qquad \mathcal{A} = \{\vec{a} \bmod q :\ \gcd(\vec{a},q) = 1,\ \vec{\rho} \in \Lambda(\vec{a},q)\}.$$

If $\Delta(\vec{c}) \neq 0$ we put $E_{\vec{c}}(q_1) = \gcd(\Delta(\vec{c}),q_{1,g})\gcd(\Delta(\vec{c}),d_g)$. If $\vec{c}$ is singular of corank r over $\mathbb{Q}$, let $D_{\vec{c}}$ be the $(s-1)$-st invariant factor of $M_{\vec{c}}$ over $\mathbb{Z}$ when $r = 1$ (a non-zero integer dividing $D_{s-1}(\vec{c})$), and $D_{\vec{c}} = 1$ when $r = 2$; we then put $E_{\vec{c}}(q_1) = q_{1,g}^{r/2}\gcd(D_{\vec{c}},q_{1,g})^{1/2}\gcd(q_{1,g},d_g)$.

Lemma 10.6. *In this notation, $\#\mathcal{A} \leqslant qd^2$ and*

$$\sum_{\vec{a}\in\mathcal{A}} K(\vec{a},q)^{1/2} \ll_{\varepsilon} q^{1+\varepsilon}d^2\,b(q)\,E_{\vec{c}}(q_1).$$

Proof. We split $q = dq_1$. At the primes dividing q_1 but not d, the numerator $\vec{a}$ is congruent to a multiple of $\vec{c}$, so $K(\vec{a},\cdot)$ equals $K(\vec{c},\cdot)$ and there is nothing to average. At the primes dividing d, the numerators run over a p-adic ball around $\lambda\vec{c}$, and Lemma 10.3 applies. The bad primes are handled crudely. By Lemma 5.7(ii), every $\vec{a} \in \mathcal{A}$ is of the form $\lambda\vec{c} + q_1\vec{b}$ with λ modulo q_1 coprime to q_1 and $\vec{b}$ modulo d, so it suffices to bound the sum over all such pairs $(\lambda,\vec{b})$ for which $\lambda\vec{c} + q_1\vec{b}$ is coprime to q; there are at most $q_1d^3 = qd^2$ of them. Fix λ. We bound $K(\vec{a},q) = \prod_p K(\vec{a},p^{n_p})$ prime by prime.

At a bad prime we use $K(\vec{a},p^n) \leqslant p^{\mu_p}p^{2n}$ (Lemma 8.4(ii)); these primes contribute at most $c_F^{1/2}b(q)$ to $K(\vec{a},q)^{1/2}$. At a good prime $p \nmid d$ we have $\vec{a} \equiv \lambda\vec{c} \pmod{p^{n_p}}$, so $K(\vec{a},p^{n_p}) = K(\vec{c},p^{n_p})$ by Lemma 8.2(iii). At a good prime $p \mid d$, with $p^{\delta} \parallel d$ and $p^{n_1} \parallel q_1$, Lemma 8.2(ii) gives $K(\vec{a},p^{n_1+\delta}) \leqslant K(\vec{c},p^{n_1})K(\vec{a},p^{\delta})$. Hence

$$K(\vec{a},q)^{1/2} \leqslant c_F^{1/2}\,b(q)\,K(\vec{c},q_{1,g})^{1/2}\,K(\vec{a},d_g)^{1/2}.$$

As $\vec{b}$ runs modulo d, the residue $\vec{a} \bmod p^{\delta}$ runs uniformly over $\lambda\vec{c} + p^{k_p}\mathbb{Z}$ modulo p^{δ}, where $k_p = \min(v_p(q_1),\delta)$, independently for the different primes $p \mid d$. If $k_p \geqslant 1$, Lemma 10.3 bounds the average of $K(\vec{a},p^{\delta})^{1/2}$ by $7(\delta+1)p^{\min(k_p,v_p(\Delta(\vec{c})))}$. If $k_p = 0$, the residues run over all residues modulo p^{δ}, and the proof of Proposition 8.7(i) bounds the sum over the primitive ones by $p^{3\delta}(\delta+1)(1+13s^2p^{-1/2})$. So the average of $K(\vec{a},d_g)^{1/2}$ over $\vec{b}$ is $\ll_{\varepsilon} q^{\varepsilon}\gcd(\Delta(\vec{c}),e_g)$, where $e = \gcd(q_1,d)$, and $e_g \leqslant \gcd(q_{1,g},d_g)$.

Finally we bound $K(\vec{c},q_{1,g})$. If $\Delta(\vec{c}) \neq 0$, Lemma 8.4(i) and the inequality $\min(v,2n)/2 \leqslant \min(v,n)$ give $K(\vec{c},q_{1,g})^{1/2} \leqslant \gcd(\Delta(\vec{c}),q_{1,g})$, and $\gcd(\Delta(\vec{c}),e_g) \leqslant$

$\gcd(\Delta(\vec c), d_g)$. If $\vec c$ is singular of corank r, then at a good prime the first $s-2$ invariant factors of $M_{\vec c}$ are units and the last r vanish, so $K(\vec c, p^n) = p^{rn}\gcd(D_{\vec c}, p^n)$ when $r = 1$ and $K(\vec c, p^n) = p^{2n}$ when $r = 2$; moreover $\gcd(0, e_g) = e_g$. In both cases we obtain the claimed bound. □

For $1 \leqslant Y \leqslant Q$ and $W > 0$ we put $r_Y = (YQ^{1/3})^{-1}$ and

(10.1)
$$\mathfrak{M}(Y,W) = Y^2 Q \min(W, r_Y)^2 \min(W, Q^{-4/3}) + YQ^2 \min(W, r_Y)\min(W, Q^{-4/3})^2.$$

By Proposition 1.3, the mass of the weights with $Y < q \leqslant 2Y$ in the ball $|\vec w| \leqslant W$ is $\ll Q^\varepsilon Y \mathfrak{M}(Y,W)$. The next proposition says that the factor $K(\vec a, q)^{1/2}$ costs only Q^ε on average, once the exceptional terms are removed. It is the main technical result of the unrefined method: combined with Lemma 10.2 and the integral bounds, it gives the estimate for $N_1(P)$ in the next section almost immediately.

Proposition 10.7. *Let $1 \leqslant Y \leqslant Q$ and $W > 0$, and put*

$$\mathcal{S}(Y,W) = \sum_{Y < q \leqslant 2Y} \sideset{}{^*}\sum_{\vec a \bmod q} K(\vec a, q)^{1/2} \int_{|\vec w| \leqslant W} \left| p^{\mathrm{gen}}_{\Lambda(\vec a, q)}(\vec w)\right| d\vec w.$$

Then, for every $\varepsilon > 0$,

$$\mathcal{S}(Y,W) \ll_\varepsilon Q^\varepsilon\, Y\, \mathfrak{M}(Y,W).$$

Proof. We follow the proof of Proposition 1.3, keeping the factor $K(\vec a, q)^{1/2}$, and we sum over $q \in (Y, 2Y]$ before summing over the vectors $\vec\nu$ and $\vec\rho$. Throughout we write $T = Q^{4/3}$ and

$$\varpi_1 = \min(W, r_Y)^2 \min(W, T^{-1}), \qquad \varpi_2 = \min(W, r_Y)\min(W, T^{-1})^2,$$

so that $Y\mathfrak{M}(Y,W) = Y^3 Q\,\varpi_1 + Y^2Q^2\,\varpi_2$. Since $r_Y/2 \leqslant r_q \leqslant r_Y$ for $q \in (Y, 2Y]$, we have $\min(W, r_q) \asymp \min(W, r_Y)$, and we use this without further comment. Note also that $Y \leqslant Q$, so that $Y^\varepsilon \leqslant Q^\varepsilon$.

Step 1: splitting the weight. By (5.1) and the definition of $p^{\mathrm{gen}}_{\Lambda(\vec a, q)}$, and since ω is bounded,

$$\left|p^{\mathrm{gen}}_{\Lambda(\vec a,q)}\right| \ll |p_{1,q}| + \sum_{\substack{\vec\nu \in \Lambda^\perp(\vec a, q),\ |\vec\nu| \leqslant Q^{1/3} \\ \vec\nu/\gcd(\vec\nu) \notin \mathcal{N}_F}} |p_{2,\vec\nu,k,q}| + \sum_{\substack{(\vec\nu,\vec\rho)\ \text{flag},\ \vec\nu \in \Lambda^\perp(\vec a,q),\ \vec\rho \in \Lambda(\vec a,q) \\ |\vec\nu| \leqslant Q^{1/3},\ |\vec\rho| \leqslant Q^{2/3},\ \vec\rho/\gcd(\vec\rho) \notin \mathcal{C}_F}} |p_{3,\vec\nu,\vec\rho,k,q}|,$$

with k as in (5.1). Accordingly $\mathcal{S}(Y,W) \ll \mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3$, where $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ are the contributions of the point, plane and flag weights. In $\mathcal{S}_2$ and $\mathcal{S}_3$ we interchange the order of summation, so that the sums over $\vec\nu$, or over the flags $(\vec\nu, \vec\rho)$, are on the outside. We recall from the proof of Proposition 1.3 the bounds

(10.2)
$$\int_{|\vec w| \leqslant W} |p_{1,q}| \ll \frac{Q}{q}\,\varpi_1, \qquad \int_{|\vec w| \leqslant W} |p_{2,\vec\nu,k,q}| \ll \frac{Q}{kq}\,\varpi_2, \qquad \int_{|\vec w| \leqslant W} |p_{3,\vec\nu,\vec\rho,k,q}| \ll Q^{2\varepsilon}\varpi_2,$$

none of which depends on $\vec a$. We also use repeatedly that the functions b and $m_{\vec\nu_0}$ of Proposition 8.7 are completely multiplicative, in the sense that $b(xy) = b(x)b(y)$ and $m_{\vec\nu_0}(xy) = m_{\vec\nu_0}(x)m_{\vec\nu_0}(y)$ for all $x, y \in \mathbb{N}$, since $v_p(xy) = v_p(x) + v_p(y)$; moreover $b(x) \leqslant x$ and $m_{\vec\nu_0}(x) \leqslant x$.

Step 2: the point terms. By (10.2) and Proposition 8.7(i),

$$\mathcal{S}_1 \ll \sum_{Y < q \leqslant 2Y} \frac{Q}{q}\,\varpi_1\, q^{3+\varepsilon} b(q) \ll Q\,\varpi_1\,(2Y)^{2+\varepsilon} \sum_{q \leqslant 2Y} b(q).$$

By Proposition 8.7(iii) the last sum is $\ll Y^{1+\varepsilon}$, so $\mathcal{S}_1 \ll Q^{2\varepsilon}Y^3Q\,\varpi_1$. This is $Q^{2\varepsilon}$ times the first term of $Y\mathfrak{M}(Y,W)$.

Step 3: the plane terms. Fix $\vec{\nu} = g\vec{\nu}_0$ with $\vec{\nu}_0 \notin \mathcal{N}_F$ primitive and $|\vec{\nu}| \leqslant Q^{1/3}$. For each q we have $k = g/\gcd(g,q)$, so $1/k = \gcd(q,g)/g$. The condition $\vec{\nu} \in \Lambda^{\perp}(\vec{a},q)$ means $q \mid \vec{\nu}\cdot\vec{a}$, so Proposition 8.7(ii) bounds the sum of $K(\vec{a},q)^{1/2}$ over the relevant $\vec{a}$ by $q^{2+\varepsilon}\gcd(q,g)m_{\vec{\nu}_0}(q)$. With (10.2), the contribution of $\vec{\nu}$ to $\mathcal{S}_2$ is therefore

$$\ll \sum_{Y<q\leqslant 2Y} \frac{Q\gcd(q,g)}{gq}\,\varpi_2\,q^{2+\varepsilon}\gcd(q,g)\,m_{\vec{\nu}_0}(q) \ll \frac{Q\,\varpi_2\,(2Y)^{1+\varepsilon}}{g}\sum_{q\leqslant 2Y}\gcd(q,g)^2 m_{\vec{\nu}_0}(q).$$

To estimate the last sum we use $\gcd(q,g)^2 \leqslant \sum_{d\mid\gcd(q,g)} d^2$, write $q = dq'$, and use $m_{\vec{\nu}_0}(dq') = m_{\vec{\nu}_0}(d)m_{\vec{\nu}_0}(q') \leqslant d\,m_{\vec{\nu}_0}(q')$ and Proposition 8.7(iii):

$$\begin{aligned}\sum_{q\leqslant 2Y}\gcd(q,g)^2 m_{\vec{\nu}_0}(q) &\leqslant \sum_{d\mid g} d^3 \sum_{q'\leqslant 2Y/d} m_{\vec{\nu}_0}(q') \ll \sum_{d\mid g} d^3\Big(\frac{2Y}{d}\Big)^{1+\varepsilon}|\vec{\nu}_0|^{\varepsilon}\\ &\ll Y^{1+\varepsilon}|\vec{\nu}_0|^{\varepsilon}\sum_{d\mid g} d^2 \leqslant 2Y^{1+\varepsilon}|\vec{\nu}_0|^{\varepsilon}g^2,\end{aligned}$$

where in the last step $\sum_{d\mid g} d^2 = g^2\sum_{d\mid g} d^{-2} \leqslant \zeta(2)g^2$. Since $|\vec{\nu}_0| \leqslant Q^{1/3}$, the contribution of $\vec{\nu}$ is $\ll Q^{3\varepsilon}Y^2Q\,\varpi_2\,g$. Finally, for each $g \leqslant Q^{1/3}$ there are $\ll (Q^{1/3}/g)^3$ primitive vectors $\vec{\nu}_0$ with $|g\vec{\nu}_0| \leqslant Q^{1/3}$, and

$$\sum_{g\leqslant Q^{1/3}} g\Big(\frac{Q^{1/3}}{g}\Big)^3 = Q\sum_{g\leqslant Q^{1/3}}\frac{1}{g^2} \ll Q.$$

Hence $\mathcal{S}_2 \ll Q^{3\varepsilon}Y^2Q^2\,\varpi_2$, which is $Q^{3\varepsilon}$ times the second term of $Y\mathfrak{M}(Y,W)$.

Step 4: the flag terms. Consider a flag $(\vec{\nu},\vec{\rho})$ with $\vec{\rho} = G\vec{c}$, where $\vec{c}$ is primitive and non-singular, so that $\Delta(\vec{c}) \neq 0$. For each q, a residue $\vec{a}$ for which this flag occurs in $p^{\mathrm{gen}}_{\Lambda(\vec{a},q)}$ satisfies $\gcd(\vec{a},q) = 1$ and $\vec{\rho} \in \Lambda(\vec{a},q)$, so it lies in the set $\mathcal{A}$ of Lemma 10.6 (we drop the condition $\vec{\nu} \in \Lambda^{\perp}(\vec{a},q)$, which can only enlarge the sum). By (10.2) and Lemma 10.6, with $d = \gcd(q,G)$ and $q_1 = q/d$,

$$\mathcal{S}_3 \ll Q^{2\varepsilon}\varpi_2 \sum_{(\vec{\nu},\vec{\rho})}\sum_{Y<q\leqslant 2Y}\sum_{\vec{a}\in\mathcal{A}} K(\vec{a},q)^{1/2} \ll Q^{2\varepsilon}\varpi_2\sum_{(\vec{\nu},\vec{\rho})}\sum_{Y<q\leqslant 2Y} q^{1+\varepsilon}d^2\,b(q)\,E_{\vec{c}}(q_1),$$

where $E_{\vec{c}}(q_1) = \gcd(\Delta(\vec{c}),q_{1,g})\gcd(\Delta(\vec{c}),d_g)$, and the outer sums run over the flags with $|\vec{\nu}| \leqslant Q^{1/3}$, $|\vec{\rho}| \leqslant Q^{2/3}$ and $\vec{c}$ non-singular. It therefore suffices to prove that

$$\sum_{(\vec{\nu},\vec{\rho})}\sum_{Y<q\leqslant 2Y} q^{1+\varepsilon}d^2\,b(q)\gcd(\Delta(\vec{c}),q_{1,g})\gcd(\Delta(\vec{c}),d_g) \ll Q^{C\varepsilon}Y^2Q^2 \tag{10.3}$$

for some constant C depending only on s; then $\mathcal{S}_3 \ll Q^{(C+2)\varepsilon}\,Y^2Q^2\,\varpi_2$, which is again of the size of the second term of $Y\mathfrak{M}(Y,W)$. We first sum over q, then over G and $\vec{\nu}$, and finally over $\vec{c}$.

The sum over q. Fix the flag. Every $q \in (Y,2Y]$ can be written uniquely as $q = dq_1$ with $d = \gcd(q,G)$, so $d \mid G$ and $q_1 \leqslant 2Y/d$. Using $q^{1+\varepsilon} \leqslant (2Y)^{1+\varepsilon}$ and $b(q) = b(d)b(q_1)$, the inner sum in (10.3) is at most

$$(2Y)^{1+\varepsilon}\sum_{d\mid G} d^2 b(d)\gcd(\Delta(\vec{c}),d_g)\sum_{q_1\leqslant 2Y/d} b(q_1)\gcd(q_{1,g},\Delta(\vec{c})).$$

By Lemma 10.4(ii) with $A = \Delta(\vec{c})$ and $B = 1$, the sum over q_1 is $\ll (Y/d)^{1+\varepsilon}\tau(\Delta(\vec{c}))$. Since $d^2(Y/d)^{1+\varepsilon} \leqslant Y^{1+\varepsilon}d$, the inner sum in (10.3) is

$$\ll Y^{2+2\varepsilon}\tau(\Delta(\vec{c}))\sum_{d\mid G} d\,b(d)\gcd(\Delta(\vec{c}),d_g).$$

The sums over G and $\vec{\nu}$. Now fix the primitive vector $\vec{c}$. Since $|\vec{\rho}| = G|\vec{c}| \leqslant Q^{2/3}$, the integer G runs over $G \leqslant Z_{\vec{c}} := Q^{2/3}/|\vec{c}|$; in particular $|\vec{c}| \leqslant Q^{2/3}$. The condition $\vec{\nu} \perp \vec{\rho}$ is equivalent to $\vec{\nu} \perp \vec{c}$, so the number of admissible $\vec{\nu}$ is at most $\nu(\vec{c}; Q^{1/3})$, independently of G. By Lemma 10.4(iii) with $\beta = 1$ and $A = \Delta(\vec{c})$,

$$\sum_{G\leqslant Z_{\vec{c}}}\sum_{d\mid G} d\,b(d)\gcd(\Delta(\vec{c}),d_g) \ll Z_{\vec{c}}^{2+\varepsilon}\tau(\Delta(\vec{c})) \leqslant Q^{4/3+\varepsilon}|\vec{c}|^{-2}\tau(\Delta(\vec{c})).$$

The sum over $\vec{c}$. Since $0 < |\Delta(\vec{c})| \ll |\vec{c}|^s \leqslant Q^{2s/3}$, the divisor bound gives $\tau(\Delta(\vec{c})) \ll Q^{\varepsilon}$. Collecting the three steps, the left side of (10.3) is

$$\ll Q^{5\varepsilon}Y^2Q^{4/3}\sum_{\substack{\vec{c}\text{ primitive}\\ |\vec{c}|\leqslant Q^{2/3}}}\frac{\nu(\vec{c};Q^{1/3})}{|\vec{c}|^2}.$$

By Lemma 10.5(iii) with $V = Q^{2/3}$, the last sum is $\ll Q^{1/3}Q^{1/3} + Q^{2/3}\log(2Q) \ll Q^{2/3+\varepsilon}$. Hence the left side of (10.3) is $\ll Q^{6\varepsilon}Y^2Q^2$, which proves (10.3) with $C = 6$.

Conclusion. Combining Steps 2–4,

$$\mathcal{S}(Y,W) \ll Q^{8\varepsilon}\Big(Y^3Q\,\varpi_1 + Y^2Q^2\,\varpi_2\Big) = Q^{8\varepsilon}\,Y\,\mathfrak{M}(Y,W).$$

Since $\varepsilon > 0$ is arbitrary, this proves the proposition. □

For the flag terms the numerators $\vec{a}$ lie on the line through $\vec{c}$ modulo q/d, so the average over $\vec{a}$ does not remove large values of $K(\vec{a},q)$. They are removed instead by the average over the moduli q. This is a simple form of the double Kloosterman refinement of [HB96], and it is the reason why it matters that $\Delta(\vec{c}) \neq 0$: for a singular $\vec{c}$ the factor $\gcd(\Delta(\vec{c}),q)$ becomes q, and these flags are treated separately in Section 12.

11. The generic terms

For each dyadic range of q and of $|\vec{w}|$ the bound is a product of four quantities: the number of dual vectors, the size of the exponential sum, the size of the integral, and the mass of the weights. The outcome, Lemma 11.1, is that all ranges on the curve $q(1+P^2|\vec{w}|) \asymp Q$ contribute the same amount $Q^{s/2}$, which is the computation sketched in §1.3.

In this section we prove Lemma 6.7(b). We decompose the ranges of q and $\vec{w}$ dyadically: $Y < q \leqslant 2Y$ with $Y = 2^{-i}Q$, and $\vec{w}$ either in the ball $|\vec{w}| \leqslant Q^{-4/3}$ or in a shell $W < |\vec{w}| \leqslant 2W$ with $W = 2^jQ^{-4/3}$. By Lemma 5.4 we may assume that $W \leqslant Q^{\varepsilon}r_Y$. In each range $X_{\vec{w}} \asymp X := 1 + P^2W$, and there are $O((\log P)^2)$ ranges.

It may help to see the shape of the final computation before going into it. In a range with $q \asymp Y$ and $X_{\vec{w}} \asymp X$, the trivial bound for the terms of $N_1(P)$ is the product of four factors:

- the number of relevant dual vectors, $\ll L^s$ with $L \asymp YX/P$ (Lemma 7.1(i));
- the size $Y^{-s}\cdot Y^{s/2}$ of the normalised exponential sum, up to the factor $K(\vec{a},q)^{1/2}$ (Lemma 8.1);
- the size $P^sX^{-(s-2)/2}$ of the integral (Lemma 7.1(ii));

- the mass $Q^{\varepsilon}Y\mathfrak{M}(Y,W)$ of the weights in the range, which also absorbs $K(\vec{a},q)^{1/2}$ (Proposition 10.7).

Multiplying out, the powers of P cancel and one is left with $Y^{s/2+1}X^{s/2+1}\mathfrak{M}(Y,W)$, which Lemma 11.1 evaluates.

The following lemma is the computation behind Section 1.3. Its content is that the product YX is the only thing that matters: the bound is a power of YX/Q, and YX can be at most about Q, because the weights are negligible outside the ball of radius $Q^{\varepsilon}r_Y$. So the worst ranges are those on the "critical curve" $YX \asymp Q$, and all of them give the same bound $Q^{s/2}$.

Lemma 11.1. *Let $1 \leqslant Y \leqslant Q$, $0 < W \leqslant Q^{\varepsilon}r_Y$ and $X = 1+P^2W$. Then $YX \ll Q^{1+\varepsilon}$, and*

$$Y^{s/2+1}X^{s/2+1}\mathfrak{M}(Y,W) \ll Q^{s/2+C\varepsilon}\Big(\frac{YX}{Q}\Big)^{s/2+2}.$$

In particular the left side is $\ll Q^{s/2+C\varepsilon} \asymp P^{3s/4+C\varepsilon}$.

Proof. We consider separately the central ball $W \leqslant Q^{-4/3}$, where $X \asymp 1$, and the shells beyond it, where $X \asymp Q^{4/3}W$; in each case we insert the definition (10.1) of $\mathfrak{M}$. Recall that $Q^{4/3} \asymp P^2$ and $r_Y = (YQ^{1/3})^{-1}$, and note that $Q^{-4/3} \leqslant r_Y$ since $Y \leqslant Q$. We have $YX \leqslant Y + P^2Q^{\varepsilon}Q^{-1/3} \ll Q^{1+\varepsilon}$.

If $W \leqslant Q^{-4/3}$, then $X \asymp 1$ and $\mathfrak{M}(Y,W) \leqslant (Y^2Q + YQ^2)Q^{-4}$, so the left side is $\ll Y^{s/2+3}Q^{-3} + Y^{s/2+2}Q^{-2} \ll Q^{s/2}(Y/Q)^{s/2+2}$.

If $Q^{-4/3} < W \leqslant Q^{\varepsilon}r_Y$, then $X \asymp Q^{4/3}W$ and

$$\mathfrak{M}(Y,W) \ll Q^{2\varepsilon}\big(Y^2QW^2Q^{-4/3} + YQ^2WQ^{-8/3}\big) \asymp Q^{2\varepsilon}\big(Y^2X^2Q^{-3} + YXQ^{-2}\big),$$

so the left side is

$$\ll Q^{2\varepsilon}\big((YX)^{s/2+3}Q^{-3} + (YX)^{s/2+2}Q^{-2}\big),$$

which is at most $Q^{s/2+C\varepsilon}(YX/Q)^{s/2+2}$ since $YX \ll Q^{1+\varepsilon}$. □

Thus every range on the curve $q(1+P^2|\vec{w}|) \asymp Q$ contributes the same amount, and the ranges below the curve contribute less.

Proposition 11.2. *We have $N_1(P) \ll P^{3s/4+\varepsilon}$. In particular Lemma 6.7(b) holds.*

Proof. We bound the terms with $\mathbf{u} \neq \mathbf{0}$ by the four-factor computation described above. The terms with $\mathbf{u} = \mathbf{0}$ and $q \geqslant q_0$ need a separate, easier argument, since for them there is no saving from the number of dual vectors; instead they are small because the moduli are at least q_0 and the singular series converges fast.

The terms with $\mathbf{u} \neq \mathbf{0}$. Fix a dyadic range (Y,W). By Lemma 7.1(i), the terms with $|\mathbf{u}| \geqslant L := 2C_1P^{\varepsilon}YX/P$ contribute $O_N(P^{-N})$. If $L < 1$ there are no other terms with $\mathbf{u} \neq \mathbf{0}$. If $L \geqslant 1$ there are $\ll L^s$ remaining vectors $\mathbf{u}$, and for each of them $|S_q(\vec{a},\mathbf{u})| \leqslant q^{s/2}K(\vec{a},q)^{1/2}$ by Lemma 8.1 and $|I_q(\vec{w},\mathbf{u})| \ll P^sX^{-(s-2)/2}$ by Lemma 7.1(ii). By Proposition 10.7 the range contributes

$$\ll Y^{-s}\cdot Y^{s/2}\cdot L^s\cdot P^sX^{-(s-2)/2}\cdot Q^{\varepsilon}Y\mathfrak{M}(Y,2W) \ll P^{C\varepsilon}\,Y^{s/2+1}X^{s/2+1}\mathfrak{M}(Y,2W),$$

which is $\ll P^{3s/4+C\varepsilon}$ by Lemma 11.1.

The terms with $\mathbf{u} = \mathbf{0}$ and $q \geqslant q_0$. Here there is a single $\mathbf{u}$, and $|I_q(\vec{w},\mathbf{0})| \ll P^sX^{-(s-2)/2}$. The range contributes $\ll P^{s+\varepsilon}Y^{1-s/2}X^{-(s-2)/2}\mathfrak{M}(Y,2W)$. For $W \leqslant Q^{-4/3}$ this is $\ll P^{s+\varepsilon}(Y^{3-s/2}Q^{-3} + Y^{2-s/2}Q^{-2})$, which for $s \geqslant 7$ is largest at $Y = q_0 = Q^{1/3-\delta}$, where it is $\ll P^{s+C\delta}Q^{-4/3-s/6} \asymp P^{3s/4-2+C\delta}$. For $Q^{-4/3} < W \leqslant Q^{\varepsilon}r_Y$ we

have $\mathfrak{M}(Y,2W) \ll Q^{2\varepsilon}(Y^2X^2Q^{-3} + YXQ^{-2})$, as in the proof of Lemma 11.1, so the range contributes $\ll P^{s+C\varepsilon}(Y^{3-s/2}X^{3-s/2}Q^{-3} + Y^{2-s/2}X^{2-s/2}Q^{-2})$. For $s \geqslant 7$ both exponents of X are negative, so this is at most the bound for $X = 1$, which we have just estimated.

Summing over the $O((\log P)^2)$ ranges and taking δ small gives the proposition. □

12. The exceptional terms

In this section we prove Lemma 6.7(c). Recall that $p^{\mathrm{exc}}_{\Lambda(\vec{a},q)}$ consists of the plane terms with $\vec{\nu}/\gcd(\vec{\nu}) \in \mathcal{N}_F$ and the flag terms with $\vec{\rho}/\gcd(\vec{\rho}) \in \mathcal{C}_F$. In the first case the numerators $\vec{a}$ lie on a plane all of whose points give singular forms, or on a plane which is very tangent to the discriminant curve; in the second case they lie on a line through a singular form $F_{\vec{c}}$. In both cases $K(\vec{a},q)$ is large, and Proposition 10.7 does not apply. We compensate for this in two ways: the exceptional terms carry a small part of the mass of the weights, and the exponential sums vanish unless the dual vector satisfies additional congruences (Lemma 8.3). The situation is similar to that of the bad vectors $\vec{c}$ in [LRV24, §§12–13], but simpler, because we do not need a Kloosterman refinement.

To see why these two compensations suffice, consider the flags with singular direction, which are the more delicate of the two kinds. If $\vec{c}$ has corank one, then $K(\vec{a},q)$ can be as large as q_1 for the numerators on the line through $\vec{c}$, a loss of $q_1^{1/2}$ against the generic terms. On the other hand, Lemma 8.3 shows that the exponential sum vanishes unless $\mathbf{u}\cdot\mathbf{h}_1 \equiv 0 \pmod{q_1}$, which cuts the number of dual vectors by the factor q_1 as long as the range of $\mathbf{u}$ is long compared with q_1; when it is not, the loss is controlled by the length of the range instead, and the bookkeeping below shows that this is still affordable. Moreover there are only $O(B^2)$ singular directions of height at most B, against B^3 directions in total, so the singular flags carry a small part of the mass. For the exceptional plane terms the compensation is even simpler: there are only finitely many exceptional directions $\vec{\nu}_0$, so these terms carry a proportion $Q^{-2/3}$ of the mass, which pays for the crude bound $K(\vec{a},q) \ll q^2 b(q)$.

Proposition 12.1. *We have $N_2(P) \ll P^{3s/4+\varepsilon}$. In particular Lemma 6.7(c) holds.*

We decompose into dyadic ranges (Y,W) as in Section 11, and compare the exceptional terms in each range with the generic bound

$$B(Y,W) = P^{C\varepsilon}Y^{s/2+1}X^{s/2+1}\mathfrak{M}(Y,2W) \ll P^{3s/4+C\varepsilon}\Big(\frac{YX}{Q}\Big)^{s/2+2}$$

obtained in the proof of Proposition 11.2. It suffices to show that in each range the exceptional terms are $\ll P^{\varepsilon}B(Y,W)$. We treat the plane terms and the flag terms separately.

Lemma 12.2. *The plane terms with $\vec{\nu}/\gcd(\vec{\nu}) \in \mathcal{N}_F$ contribute $\ll P^{C\varepsilon}B(Y,W)$ to $N_2(P)$ in each dyadic range, and $O(P^{3s/4-3/2+C\varepsilon})$ through the vector $\mathbf{u} = \mathbf{0}$.*

Proof. We first show that these terms carry a proportion $Q^{-2/3+\varepsilon}$ of the mass of the plane terms. Then, for $\mathbf{u} \neq \mathbf{0}$, we use the worst-case bound for $K(\vec{a},q)$ inside Lemma 10.2, which loses at most a factor P, and this is more than paid for by the small mass. The vector $\mathbf{u} = \mathbf{0}$ is treated similarly.

The mass of these terms. Since $\mathcal{N}_F$ is finite (Lemma 6.3), these terms have $\vec{\nu} = g\vec{\nu}_0$ with $\vec{\nu}_0$ in a fixed finite set and $g \ll Q^{1/3}$. In the proof of Proposition 1.3, the plane

terms with a given q contributed $qQ\min(W,r_q)\min(W,T^{-1})^2$ times $\sum_{\vec\nu}\gcd(q,g)^2/g$, and this sum over all $\vec\nu$ was $\ll Q$. Restricted to our $\vec\nu$ it is

$$\ll \sum_{g\ll Q^{1/3}} \frac{\gcd(q,g)^2}{g} \leqslant \sum_{\substack{d\mid q\\ d\ll Q^{1/3}}} d^2 \sum_{\substack{g\ll Q^{1/3}\\ d\mid g}} \frac{1}{g} \ll \tau(q)Q^{1/3}\log Q.$$

So these terms carry at most a proportion $\ll Q^{-2/3+\varepsilon}$ of the mass of the plane terms, for every q.

The dual vectors $\mathbf{u}\neq\mathbf{0}$. Fix a range, let $L=2C_1P^\varepsilon YX/P$ as in the proof of Proposition 11.2, and assume $L\geqslant 1$; otherwise there are no such terms, up to $O(P^{-N})$. For each $(q,\vec a,\vec w)$ we use $|I_q(\vec w,\mathbf{u})|\ll P^sX^{-(s-2)/2}$ and Lemma 10.2 in place of the pointwise bound for S_q:

$$\sum_{0<|\mathbf{u}|\leqslant L}|S_q(\vec a,\mathbf{u})| \ll q^{s/2}L^s\Big(1+\frac{K(\vec a,q)^{1/2}}{L}\Big) \ll q^{s/2}L^s\Big(1+\frac{c_F^{1/2}b(q)^{1/2}q}{L}\Big),$$

by Lemma 8.4(ii). Since $q/L\ll P^{1-\varepsilon}/X\leqslant P$, the factor in brackets is $\ll P\,b(q)^{1/2}$. Write $\mathfrak{M}_q^{\mathrm{exc}}$ for the integral over $|\vec w|\leqslant 2W$ of $\sideset{}{^*}\sum_{\vec a}|p_{2,\vec\nu,k,q}|$, summed over the exceptional $\vec\nu$. By the mass computation above,

$$\mathfrak{M}_q^{\mathrm{exc}} \ll Q^{-2/3+\varepsilon}\cdot qQ^2\min(W,r_q)\min(W,Q^{-4/3})^2 \ll Q^{-2/3+\varepsilon}\mathfrak{M}(Y,2W).$$

Hence the range contributes

$$\begin{aligned}&\ll \sum_{Y<q\leqslant 2Y} q^{-s}\cdot q^{s/2}L^sP\,b(q)^{1/2}\cdot P^sX^{-(s-2)/2}\cdot\mathfrak{M}_q^{\mathrm{exc}}\\ &\ll P^{1+C\varepsilon}Q^{-2/3}\,Y^{-s/2}L^sP^sX^{-(s-2)/2}\sum_{Y<q\leqslant 2Y} b(q)\,\mathfrak{M}(Y,2W).\end{aligned}$$

By Proposition 8.7(iii) the sum over q is $\ll Y^{1+\varepsilon}\mathfrak{M}(Y,2W)$, and

$$Y^{-s/2}L^sP^sX^{-(s-2)/2}\cdot Y\mathfrak{M}(Y,2W) \ll P^{C\varepsilon}Y^{s/2+1}X^{s/2+1}\mathfrak{M}(Y,2W)$$

as in the proof of Proposition 11.2. Since $Q^{2/3}\asymp P$, these terms are $\ll P^{C\varepsilon}B(Y,W)$.

The vector $\mathbf{u}=\mathbf{0}$. Here

$$|S_q(\vec a,\mathbf{0})|\leqslant q^{s/2}K(\vec a,q)^{1/2}\ll q^{s/2+1}b(q)^{1/2}.$$

Together with $|I_q(\vec w,\mathbf{0})|\ll P^sX^{-(s-2)/2}$ and the bound for $\mathfrak{M}_q^{\mathrm{exc}}$, the range contributes

$$\begin{aligned}&\ll P^sX^{-(s-2)/2}\sum_{Y<q\leqslant 2Y} q^{1-s/2}b(q)\,Q^{-2/3+\varepsilon}\mathfrak{M}(Y,2W)\\ &\ll P^{s+C\varepsilon}Y^{2-s/2}Q^{-2/3}X^{-(s-2)/2}\mathfrak{M}(Y,2W).\end{aligned}$$

This is the bound for the generic terms with $\mathbf{u}=\mathbf{0}$ in the proof of Proposition 11.2, multiplied by $YQ^{-2/3}\leqslant Q^{1/3}\asymp P^{1/2}$. Those terms were $O(P^{3s/4-2+C\delta})$ in every range, so these are $O(P^{3s/4-3/2+C\varepsilon})$. □

It remains to treat the flag terms with singular direction.

Lemma 12.3. *The flag terms with $\vec\rho/\gcd(\vec\rho)\in\mathcal{C}_F$ contribute $\ll P^{C\varepsilon}B(Y,W)$ to $N_2(P)$ in each dyadic range, and $O(P^{3s/4-3/2+C\varepsilon})$ through the vector $\mathbf{u}=\mathbf{0}$.*

Proof. We follow the treatment of the generic flag terms in Proposition 10.7, with two changes: the local bound of Lemma 10.6 now contains the large factor $q_{1,g}^{r/2}$, and the number of dual vectors is reduced by the factor q_1 thanks to Lemma 8.3. The bookkeeping reduces to the inequality (12.1), which we then check separately for corank one and corank two. Let $\vec{\rho} = G\vec{c}$ with $\vec{c} \in \mathcal{C}_F$ of corank $r \in \{1,2\}$, let $\mathbf{h}_1$ be a primitive vector of $\ker M_{\vec{c}} \cap \mathbb{Z}^s$, and for $q \in (Y, 2Y]$ put $d = \gcd(q, G)$ and $q_1 = q/d$. Every residue $\vec{a} \in \mathcal{A}$ (in the notation of Lemma 10.6) satisfies $\vec{a} \equiv \lambda \vec{c} \pmod{q_1}$ for some λ.

The dual vectors $\mathbf{u} \neq \mathbf{0}$. Let $L = 2C_1 P^{\varepsilon} YX/P$ as in the proof of Proposition 11.2; if $L < 1$ there are no such terms up to $O(P^{-N})$. By Lemma 8.3 with $q' = q_1$, $S_q(\vec{a}, \mathbf{u}) \neq 0$ implies $\mathbf{u} \cdot \mathbf{h}_1 \equiv 0 \pmod{q_1}$. These $\mathbf{u}$ form a lattice of index q_1, since $\mathbf{h}_1$ is primitive, so by Lemma 10.1 and Lemma 8.1

$$\sum_{0<|\mathbf{u}|\leqslant L} |S_q(\vec{a}, \mathbf{u})| \ll q^{s/2} K(\vec{a}, q)^{1/2} L^s \Big(\frac{1}{q_1} + \frac{1}{L} \Big).$$

In the proof of Proposition 11.2 the corresponding factor was $q^{s/2} K(\vec{a}, q)^{1/2} L^s$, and for the flag terms it was summed using Proposition 10.7, that is, using (10.3). It therefore suffices to show that

$$\sum_{(\vec{\nu}, \vec{\rho})} \sum_{Y < q \leqslant 2Y} \sum_{\vec{a} \in \mathcal{A}} K(\vec{a}, q)^{1/2} \Big(\frac{1}{q_1} + \frac{1}{L} \Big) \ll Q^{C\varepsilon} Y^2 Q^2, \tag{12.1}$$

the outer sum running over the flags with $\vec{c} \in \mathcal{C}_F$. By Lemma 10.6, and since $1/q_1 \leqslant 1/q_{1,g}$, the inner double sum for a fixed flag is

$$\ll Y^{1+\varepsilon} \sum_{d \mid G} d^2 b(d) \sum_{q_1 \leqslant 2Y/d} b(q_1)\, q_{1,g}^{r/2} \gcd(D_{\vec{c}}, q_{1,g})^{1/2} \gcd(q_{1,g}, d_g) \Big(\frac{1}{q_{1,g}} + \frac{1}{L} \Big).$$

We use $q_{1,g}^{r/2-1} \leqslant 1$ and $q_{1,g}^{r/2} \leqslant (2Y/d)^{r/2}$, and $\gcd(D_{\vec{c}}, x)^{1/2} \leqslant \gcd(D_{\vec{c}}, x)$. Lemma 10.4(ii) with $A = D_{\vec{c}}$ and $B = d_g$ bounds the sum over q_1 by $\ll (Y/d)^{1+\varepsilon} \tau(D_{\vec{c}}) \tau(d) \gcd(D_{\vec{c}}, d_g) \big(1 + (Y/d)^{r/2} L^{-1} \big)$. So the inner double sum is

$$\ll Y^{2+2\varepsilon} \tau(D_{\vec{c}}) \sum_{d \mid G} \tau(d) b(d) \gcd(D_{\vec{c}}, d_g) \Big(d + \frac{Y^{r/2} d^{1-r/2}}{L} \Big).$$

Now we fix $\vec{c}$ and sum over $G \leqslant Z := Q^{2/3}/|\vec{c}|$ by Lemma 10.4(iii), with $\beta = 1$ and $\beta = 1 - r/2$ respectively. Since $|D_{\vec{c}}| \leqslant |D_{s-1}(\vec{c})| \ll |\vec{c}|^{s-1}$, we have $\tau(D_{\vec{c}}) \ll Q^{\varepsilon}$, and we obtain for each $\vec{c}$

$$\ll Q^{C\varepsilon}\, \nu(\vec{c}; Q^{1/3})\, Y^2 \Big(Z^2 + \frac{Y^{r/2} Z^{2-r/2}}{L} \Big).$$

By Lemma 10.5(i) and $\lambda_1(\vec{c}) \geqslant 1$ we have $\nu(\vec{c}; Q^{1/3}) \ll Q^{1/3} + Q^{2/3}/|\vec{c}|$.

Corank $r = 1$. By (6.1) there are $\ll R^2$ vectors $\vec{c} \in \mathcal{C}_F$ with $|\vec{c}| \asymp R$. With $Z = Q^{2/3}/|\vec{c}|$, the first term contributes

$$\ll Q^{C\varepsilon} Y^2 Q^{4/3} \sum_R R^2 \cdot \frac{Q^{1/3} + Q^{2/3} R^{-1}}{R^2} \ll Q^{C\varepsilon} Y^2 Q^2,$$

the sum being over dyadic $R \leqslant Q^{2/3}$. The second term contributes

$$\ll Q^{C\varepsilon} \frac{Y^{5/2} Q}{L} \sum_R R^2 \cdot \frac{Q^{1/3} + Q^{2/3} R^{-1}}{R^{3/2}} \ll Q^{C\varepsilon} \frac{Y^{5/2} Q^{5/3}}{L}.$$

Since $L \asymp P^{\varepsilon} YX/P$ and $X \geqslant 1$, we have $Y^{1/2}/L \ll P/(Y^{1/2}X) \leqslant P/(YX)^{1/2}$, and $L \geqslant 1$ gives $YX \gg P^{1-\varepsilon}$. Hence $Y^{1/2}/L \ll P^{1/2+\varepsilon} \asymp Q^{1/3+\varepsilon}$, and this term is also $\ll Q^{C\varepsilon} Y^2 Q^2$.

Corank $r = 2$. By Lemma 6.1(iii) there are only finitely many such $\vec{c}$, so $Z \asymp Q^{2/3}$, $\nu(\vec{c}; Q^{1/3}) \ll Q^{2/3}$ and $D_{\vec{c}} = 1$. The first term gives $\ll Q^{C\varepsilon} Q^{2/3} Y^2 Q^{4/3}$, and the second $\ll Q^{C\varepsilon} Q^{2/3} Y^3 Q^{2/3}/L$. Since $Y/L \ll P^{1+\varepsilon}/X \ll Q^{2/3+\varepsilon}$, both are $\ll Q^{C\varepsilon} Y^2 Q^2$. This proves (12.1).

The vector $\mathbf{u} = \mathbf{0}$. Here we use $K(\vec{a}, q)^{1/2} \leqslant c_F^{1/2} b(q)^{1/2} q$ (Lemma 8.4(ii)), a loss of at most $q\, b(q)^{1/2}$ against the generic terms with $\mathbf{u} = \mathbf{0}$, whose total mass is at least that of the exceptional ones. By Proposition 8.7(iii) the factor $b(q)^{1/2}$ costs only Q^{ε} after summing over q. In the proof of Proposition 11.2, the terms with $\mathbf{u} = \mathbf{0}$ in a range with $q \asymp Y$ and $W \leqslant Q^{-4/3}$ were bounded by $P^{s+\varepsilon}(Y^{3-s/2}Q^{-3} + Y^{2-s/2}Q^{-2})$, and the extra factor Y leaves this decreasing in Y for $s \geqslant 11$. Its value at $Y = q_0 = Q^{1/3-\delta}$ is $\ll P^{3s/4-2+C\delta} q_0 \ll P^{3s/4-3/2+C\delta}$. The ranges with larger W are smaller, as before. □

Proof of Proposition 12.1. Combine Lemmas 12.2 and 12.3, and sum over the $O((\log P)^2)$ dyadic ranges. □

We note where the singular directions are delicate. In the extreme range $q \asymp Q$, $X \asymp 1$, the restriction $\mathbf{u} \cdot \mathbf{h}_1 \equiv 0 \pmod{q_1}$ is what compensates the factor $q_1^{1/2}$ in $K(\vec{c}, q_1)^{1/2}$; without it the flags of corank 1 would cost a factor $P^{1/4}$, which is exactly the margin at $s = 25$.

13. Comparison with other methods

In this section we compare Theorem 1.2 with other ways of detecting three equations, using the heuristics of [LRV24, §14]. Nothing in this section is used later. Its purpose is to explain where the numbers 25, 22 and 27 come from, and why a three-dimensional delta symbol is better suited to three equations than nesting lower-dimensional ones. Following [LRV24], an R-dimensional delta symbol is an identity

$$\delta_{\vec{n}} = \sum_{1 \leqslant q \leqslant Q} \sideset{}{^*}\sum_{\vec{a} \bmod q} \int_{|\vec{w}| \leqslant Q^{\epsilon}/(qQ^{\eta})} p_{q,\vec{a}}(\vec{w})\, \mathrm{e}\Big((\vec{a}/q + \vec{w}) \cdot \vec{n}\Big)\, d\vec{w} + O_{\eta,N}(Q^{-N}) \tag{13.1}$$

with some $0 < \eta \leqslant 1/R$. If it is applied to vectors of size about M, one takes $Q = M^{1/(1+\eta)}$ ([LRV24, Heuristic 1]). Let A be the average size of the classes of numerators $\vec{a}$ which share the same function $p_{q,\vec{a}}$. Then [LRV24, Heuristic 2] says that the efficacy of (13.1) for counting zeros of R polynomials in s variables is roughly measured by

$$Q^{s/2} A^{-1/2},$$

the expected size of the error term, and that a double Kloosterman refinement, using also the average over q, might save a further factor $Q^{1/2}$.

For three quadratic forms in variables of size P we have $M = P^2$. Theorem 1.2 has $\eta = 1/3$, the largest value allowed by Dirichlet's theorem in $\mathbb{R}^3$, so $Q = P^{3/2}$. In this paper we have not used any average over $\vec{a}$, and the error term $Q^{s/2} = P^{3s/4}$ in Lemma 6.7 is the case $A = 1$ of the heuristic. The flag terms, which dominate the critical range, allow an average over the classes $\{\lambda \vec{a} : \gcd(\lambda, q) = 1\}$, so $A \asymp Q$. The heuristic then predicts the error $Q^{s/2-1/2} = P^{3s/4-3/4}$, which is smaller than the main term P^{s-6} for $s > 21$. With the average over q as well, the error would be $P^{3s/4-3/2}$, smaller than P^{s-6} for $s > 18$.

To carry out the average over λ, one needs the sums $\sum^*_{\lambda \bmod q} S_q(\lambda\vec{a}, \mathbf{u})$. For an odd prime $q = p$ with $p \nmid \Delta(\vec{a})$, completing the square shows that $S_p(\lambda\vec{a}, \mathbf{u})$ is $p^{s/2}$ times a factor of modulus one which depends on λ through the quadratic character of λ^s and through $\mathrm{e}_p(-\overline{2\lambda}\,\mathbf{u}^T M_{\vec{a}}^{-1}\mathbf{u})$. So the average over λ is a Gauss sum when s is odd and a Ramanujan sum when s is even; it saves $p^{1/2}$ when s is odd, and p when s is even unless p divides the dual form $\mathbf{u}^T \operatorname{adj}(M_{\vec{a}})\mathbf{u}$. This is the source of the difference between even and odd s in [LRV24, Theorem 1.1], see the comments on Ramanujan sums in [LRV24, §14], and one may hope for $s \geqslant 20$ for even s here. In Sections 14 and 15 we carry out the average over λ and obtain $s \geqslant 22$ for all s. We do not use the extra saving for even s; the reason is explained in §1.3.

Comparison 1. One can also detect three equations by nesting. Applying the one-dimensional symbol to F_1 with modulus $q_1 \leqslant P$, imposing $q_1 \mid F_2$ and $q_1 \mid F_3$ by additive characters as in [Mun15], and applying the two-dimensional symbol of [LRV24] to $(F_2, F_3)/q_1$ with modulus $q_2 \leqslant (P^2/q_1)^{2/3}$, gives total moduli $q_1 q_2 \leqslant P^{5/3}$ and classes of size about $q_1 \cdot q_1^2 \cdot q_2 \leqslant P^{11/3}$. The heuristic error is $P^{(5s-11)/6}$, which is smaller than P^{s-6} for $s > 25$. Nesting in the other order (the two-dimensional symbol for (F_1, F_2), then the one-dimensional symbol for F_3) gives $s > 27$, and nesting the one-dimensional symbol three times gives total moduli $P^{7/4}$, classes of size $P^{17/4}$, and $s > 31$. Without any average over $\vec{a}$ the nested methods give $s > 36$, against $s > 24$ for Theorem 1.2.

Comparison 2. Iterating the duality of Lemma 4.1 gives an R-dimensional symbol for every R, with flags of sublattices of heights $Q^{1/R}, Q^{2/R}, \dots, Q^{(R-1)/R}$ and moduli $q \leqslant Q$ for vectors of size $Q^{1+1/R}$; the argument of Section 5.4 gives its main-term property (Remark 5.6). For R quadratic forms $Q = P^{2R/(R+1)}$, and the error terms are $Q^{s/2} = P^{sR/(R+1)}$ without an average over $\vec{a}$ and $P^{(s-1)R/(R+1)}$ with it. These are smaller than P^{s-2R} for $s > 2R(R+1)$ and $s > 2R^2 + R$ respectively. Since Rydin Myerson's theorem [RM18] needs only $s \geqslant 9R$, the delta symbol approach can improve on it only for $R \leqslant 3$:

R	Q	no average over $\vec{a}$	average over $\lambda\vec{a}$	Birch	Rydin Myerson
1	P	5	4	5	9
2	$P^{4/3}$	13	11	14	18
3	$P^{3/2}$	25	22	27	27
4	$P^{8/5}$	41	37	44	36

The entries in the third and fourth columns are the smallest s allowed by the heuristic; for $R = 1, 2$ and the fourth column they are the theorems of [HB96] and [LRV24] (with $s = 10$ for even s when $R = 2$), and for $R = 3$ they are Theorems 6.5 and 1.1.

14. The Kloosterman refinement and the case $s \geqslant 23$

So far we have estimated every exponential sum by its absolute value. In this section and the next we use the fact that the weight $p_{\Lambda(\vec{a},q)}$ is the same for all the numerators $\lambda\vec{a}$ with $\gcd(\lambda, q) = 1$, so that we may sum the exponential sums over these numerators before taking absolute values. In this section we prove the following theorem.

Theorem 14.1. *Under the hypotheses of Theorem 6.5, the asymptotic formula* (1.1) *holds for $s \geqslant 23$, for some $\Delta = \Delta(s) > 0$.*

For $s \geqslant 25$ this is Theorem 6.5, so only $s = 23$ and $s = 24$ have to be treated. The plan is that of Sections 11 and 12. In each dyadic range of q and of $|\vec{w}|$ we compare each family of terms with the bound $B(Y, W)$ of the unrefined method, and we show that it is smaller than $B(Y, W)$ by a fixed power of P more than is needed. For $s = 23$ we need to save $P^{1/4} = Q^{1/6}$ against $B(Y, W)$ (§14.2).

Two savings are available. The first comes from the sums over the numerators. For a flag term the numerators are, at most primes, a single class $\{\lambda\vec{c}\}$, and the sum over λ is a Gauss sum or a Ramanujan sum, which saves $q^{1/2}$ (§14.1). For the point and plane terms the numerators run over all residues, or over the residues on a plane, and the complete sums over them save $q^{1/2}$ as well (§§14.5–14.6). The second saving comes from the exponential integrals. Away from the finitely many real members of corank two the integrals decay faster than Lemma 7.1 says, and on average over the directions of $\vec{w}$ they save $X^{1/2}$, where $X = 1 + P^2|\vec{w}|$ (§14.11). On the critical curve $q(1 + P^2|\vec{w}|) \asymp Q$ the two savings together give $Q^{1/2} = P^{3/4}$. For $s = 23$ this is $P^{1/2} = Q^{1/3}$ more than we need.

This margin allows a simple treatment of everything that is not generic. For a flag term we keep the cancellation in the sum over λ only at the primes which divide neither $\gcd(q, \gcd(\vec{\rho}))$ nor the discriminant of the member $\vec{c} = \vec{\rho}/\gcd(\vec{\rho})$. At all other primes we use nothing but the size of the Gauss sums and the fact that the dual vectors for which they do not vanish lie in a coset (Lemma 8.1). This costs at most a factor $1 + K^{1/2}/L$, where K is the index of the coset and $L \asymp P^{1/2}$ is the length of the range of dual vectors on the critical curve. However, the moduli on which this loss is large are rare. A large value of $\gcd(q, \gcd(\vec{\rho}))$ forces the direction $\vec{c}$ to be short, and the flags with short directions carry little of the mass (Lemma 14.22); and among all moduli q, those composed of the primes dividing a fixed integer are rare. In this way the generic flag terms are smaller than $B(Y, W)$ by $Q^{1/2}$ up to a loss of at most $Q^{1/6}$ (Proposition 14.24). The flags whose direction is a singular member of the net are treated in the same way in §14.13, with the kernel of the singular member in place of the sum over λ.

In Section 15 we show that the loss $Q^{1/6}$ for the generic flag terms can be removed at the end of the critical curve where it occurs, and this gives $s = 22$.

A guide to this section. The section is long, and it may help to know in advance how it is organised.

- §14.1 evaluates the sums over λ at good primes (Lemma 14.3); they are Gauss sums or Ramanujan sums times a fixed phase.
- §14.2 sets up the bookkeeping: it defines the "relative size" of a family of terms, and says how much each family has to save (Lemma 14.6).
- §14.4 contains the averaging lemmas, in particular Lemma 14.9, which carries out the whole refinement for a single non-singular direction $\vec{c}$.
- §§14.5–14.6 treat the point and plane terms. The cancellation is used only at primes which divide q exactly once; the other primes are rare.
- §§14.7–14.8 treat the generic flag terms, which carry most of the mass and are the heart of the matter.
- §§14.9–14.10 count the dual vectors for which no cancellation is available.
- §14.11 improves the bound for the integrals on average over the direction of $\vec{w}$.
- §§14.12–14.13 return to the exceptional terms of Section 12.
- §14.14 adds everything up.

A reader who wants the main line of the argument can read §§14.1, 14.2, 14.4, 14.8, 14.11 and 14.14 first, taking the other results on trust.

14.1. **Sums over λ.** For $q \in \mathbb{N}$, $\vec{a} \in \mathbb{Z}^3$ and $\mathbf{u} \in \mathbb{Z}^s$ we put

$$T_q(\vec{a}, \mathbf{u}) = \sideset{}{^*}\sum_{\lambda \bmod q} S_q(\lambda\vec{a}, \mathbf{u}).$$

For $\vec{a}$ with $\Delta(\vec{a}) \neq 0$ we write $F^*_{\vec{a}}(\mathbf{u}) = \mathbf{u}^T \operatorname{adj}(M_{\vec{a}})\mathbf{u}$ for the dual form of $F_{\vec{a}}$; it is a quadratic form in $\mathbf{u}$ with integer coefficients, and $F^*_{\lambda\vec{a}} = \lambda^{s-1}F^*_{\vec{a}}$.

Lemma 14.2. *Let $q = q_1q_2$ with $\gcd(q_1, q_2) = 1$. Then*

$$S_q(\vec{a}, \mathbf{u}) = S_{q_1}(q_2\vec{a}, \mathbf{u})S_{q_2}(q_1\vec{a}, \mathbf{u}), \qquad T_q(\vec{a}, \mathbf{u}) = T_{q_1}(q_2\vec{a}, \mathbf{u})T_{q_2}(q_1\vec{a}, \mathbf{u}).$$

Moreover $T_q(\mu\vec{a}, \mathbf{u}) = T_q(\vec{a}, \mathbf{u})$ whenever $\gcd(\mu, q) = 1$.

Proof. This is the usual multiplicativity of complete exponential sums, from the Chinese remainder theorem. Write $\mathbf{b} = q_2\mathbf{b}_1 + q_1\mathbf{b}_2$ with $\mathbf{b}_i$ modulo q_i. Since $\vec{a} \cdot \vec{F}(\mathbf{b}) \equiv q_2^2\, \vec{a} \cdot \vec{F}(\mathbf{b}_1) + q_1^2\, \vec{a} \cdot \vec{F}(\mathbf{b}_2) \pmod q$, we get

$$\mathrm{e}_q(\vec{a} \cdot \vec{F}(\mathbf{b}) + \mathbf{u} \cdot \mathbf{b}) = \mathrm{e}_{q_1}(q_2\vec{a} \cdot \vec{F}(\mathbf{b}_1) + \mathbf{u} \cdot \mathbf{b}_1)\, \mathrm{e}_{q_2}(q_1\vec{a} \cdot \vec{F}(\mathbf{b}_2) + \mathbf{u} \cdot \mathbf{b}_2),$$

which gives the first identity. The second follows by writing λ modulo q as a pair of residues modulo q_1 and q_2, and the last is clear. □

For a prime power p^e and an integer t let $c_{p^e}(t) = \sum^*_{\mu \bmod p^e} \mathrm{e}_{p^e}(\mu t)$ be the Ramanujan sum, and for odd p let $g_{p^e}(t) = \sum^*_{\mu \bmod p^e} \left(\frac{\mu}{p}\right) \mathrm{e}_{p^e}(\mu t)$.

The next lemma is the basic computation of the Kloosterman refinement. At an odd prime p which does not divide $\Delta(\vec{a})$, completing the square shows that $S_{p^e}(\lambda\vec{a}, \mathbf{u})$ depends on λ only through a quadratic character and a phase $\mathrm{e}_{p^e}(c/\lambda)$, where c is essentially the value of the dual form $F^*_{\vec{a}}$ at $\mathbf{u}$. Summing over λ therefore produces a Ramanujan sum or a Gauss sum in c, and these are small unless c is divisible by a large power of p.

Lemma 14.3. *Let p be an odd prime, $e \geqslant 1$, and $\vec{a} \in \mathbb{Z}^3$ with $p \nmid \Delta(\vec{a})$. There is a complex number $\gamma = \gamma_{p^e}(\vec{a})$ with $|\gamma| = p^{es/2}$ such that, for $\gcd(\lambda, p) = 1$ and $\mathbf{u} \in \mathbb{Z}^s$,*

$$S_{p^e}(\lambda\vec{a}, \mathbf{u}) = \gamma\left(\frac{\lambda}{p}\right)^{es} \mathrm{e}_{p^e}\Big(-\overline{2\lambda\Delta(\vec{a})}\, F^*_{\vec{a}}(\mathbf{u})\Big).$$

Consequently, with $t = -\overline{2\Delta(\vec{a})}F^*_{\vec{a}}(\mathbf{u})$,

$$T_{p^e}(\vec{a}, \mathbf{u}) = \begin{cases} \gamma\, c_{p^e}(t) & \text{if } es \text{ is even,} \\ \gamma\, g_{p^e}(t) & \text{if } es \text{ is odd,} \end{cases}$$

and in particular:

- *(i)* $|T_{p^e}(\vec{a}, \mathbf{u})| \leqslant p^{e(s+1)/2} \gcd\Big(p^e, F^*_{\vec{a}}(\mathbf{u})\Big)^{1/2}$;
- *(ii) if s is even, $|T_{p^e}(\vec{a}, \mathbf{u})| \leqslant p^{es/2} \gcd\Big(p^e, F^*_{\vec{a}}(\mathbf{u})\Big)$, and $T_{p^e}(\vec{a}, \mathbf{u}) = 0$ unless $p^{e-1} \mid F^*_{\vec{a}}(\mathbf{u})$;*
- *(iii) if s and e are odd, $T_{p^e}(\vec{a}, \mathbf{u}) = 0$ unless p^{e-1} exactly divides $F^*_{\vec{a}}(\mathbf{u})$; in particular $T_p(\vec{a}, \mathbf{u}) = 0$ when $p \mid F^*_{\vec{a}}(\mathbf{u})$.*

Proof. We complete the square to remove the linear term $\mathbf{u}\cdot\mathbf{b}$, which produces the phase depending on $F_{\vec{a}}^*(\mathbf{u})$, and then evaluate the remaining pure Gauss sum by diagonalising $M_{\vec{a}}$ over $\mathbb{Z}_p$. Write $M = M_{\vec{a}}$ and let $\overline{M}$ be the symmetric integer matrix with $M\overline{M} \equiv I \pmod{p^e}$; then $\overline{M} \equiv \overline{\Delta(\vec{a})}\,\mathrm{adj}(M) \pmod{p^e}$. Since p is odd and $p \nmid \lambda\Delta(\vec{a})$, we may complete the square: replacing $\mathbf{b}$ by $\mathbf{b} - \overline{\lambda}\,\overline{M}\mathbf{u}$,

$$\tfrac{1}{2}\lambda\mathbf{b}^T M\mathbf{b} + \mathbf{u}\cdot\mathbf{b} \mapsto \tfrac{1}{2}\lambda\mathbf{b}^T M\mathbf{b} - \overline{2\lambda}\,\mathbf{u}^T\overline{M}\mathbf{u} \pmod{p^e}.$$

Hence

$$S_{p^e}(\lambda\vec{a},\mathbf{u}) = \mathrm{e}_{p^e}(-\overline{2\lambda\Delta(\vec{a})}F_{\vec{a}}^*(\mathbf{u}))\, S_{p^e}(\lambda\vec{a},\mathbf{0}).$$

Diagonalising M over $\mathbb{Z}_p$ as $\mathrm{diag}(d_1,\dots,d_s)$ with units d_i, $S_{p^e}(\lambda\vec{a},\mathbf{0})$ is the product of the one-dimensional Gauss sums

$$\sum_{x \bmod p^e} \mathrm{e}_{p^e}(\frac{1}{2}\lambda d_i x^2) = (\frac{\lambda d_i \overline{2}}{p})^e G_{p^e}$$

with $|G_{p^e}| = p^{e/2}$. This gives the formula with $\gamma = (\frac{\overline{2}^s d_1\cdots d_s}{p})^e G_{p^e}^s$. Summing over λ and substituting $\mu = \overline{\lambda}$, so that $(\frac{\lambda}{p}) = (\frac{\mu}{p})$, we obtain $T_{p^e} = \gamma c_{p^e}(t)$ or $\gamma g_{p^e}(t)$.

For the bounds, recall that $c_{p^e}(t)$ vanishes unless $p^{e-1} \mid t$, and $|c_{p^e}(t)| \leqslant \gcd(p^e,t)$. If e is odd, the sum $g_{p^e}(t)$ vanishes unless p^{e-1} exactly divides t, in which case $|g_{p^e}(t)| = p^{e-1/2}$; if e is even, $g_{p^e} = c_{p^e}$, since then es is even. In each case $|c_{p^e}(t)|$ or $|g_{p^e}(t)|$ is at most $p^{e/2}\gcd(p^e,t)^{1/2}$. Finally $\gcd(p^e,t) = \gcd(p^e, F_{\vec{a}}^*(\mathbf{u}))$. □

For the primes dividing $2D_F\Delta(\vec{a})$ we use the bound

$$|T_{p^e}(\vec{a},\mathbf{u})| \leqslant \varphi(p^e)p^{es/2}K(\vec{a},p^e)^{1/2}$$

from Lemma 8.1. Combining this with Lemmas 14.2 and 14.3 we obtain the following.

Proposition 14.4. *Let $\vec{a}\in\mathbb{Z}^3$ with $\Delta(\vec{a})\neq 0$, and $q\in\mathbb{N}$. Write $q = q_1q_2$, where q_2 is composed of the primes dividing $2D_F\Delta(\vec{a})$ and $\gcd(q_1, 2D_F\Delta(\vec{a})) = 1$. Then*

$$|T_q(\vec{a},\mathbf{u})| \leqslant q^{(s+1)/2}\,\gcd\Big(q_1, F_{\vec{a}}^*(\mathbf{u})\Big)^{1/2} E(\vec{a},q_2), \qquad E(\vec{a},q_2) = \prod_{p^e\| q_2} p^{e/2}K(\vec{a},p^e)^{1/2}.$$

If s is even, the factor $q_1^{1/2}\gcd(q_1,F_{\vec{a}}^*(\mathbf{u}))^{1/2}$ may be replaced by $\gcd(q_1,F_{\vec{a}}^*(\mathbf{u}))$.

Compared with the trivial bound $\varphi(q)q^{s/2}K(\vec{a},q)^{1/2}$, the proposition saves the factor $(q_1/g_1)^{1/2}$, where $g_1 = \gcd(q_1,F_{\vec{a}}^*(\mathbf{u}))$, and the factor q_1/g_1 when s is even. There is no saving when $F_{\vec{a}}^*(\mathbf{u}) = 0$, and none at the primes dividing $2D_F\Delta(\vec{a})$. The rest of the section is about showing that these exceptional situations are rare enough: the dual vectors with $F_{\vec{a}}^*(\mathbf{u}) = 0$ lie on a quadric and are few, and the moduli with a large factor in common with $2D_F\Delta(\vec{a})$ are rare on average over q.

14.2. **The savings required.** We keep the notation of Sections 11 and 12. For a dyadic range (Y,W) we put $X = 1 + P^2W$,

$$t = \frac{YX}{Q} \ll Q^{\varepsilon}, \qquad L = 2C_1P^{\varepsilon}\frac{YX}{P} \asymp P^{1/2+\varepsilon}t,$$

and we write $B(Y,W) = P^{C\varepsilon}Y^{s/2+1}X^{s/2+1}\mathfrak{M}(Y,2W)$ for the bound obtained for the generic terms in the proof of Proposition 11.2; by Lemma 11.1, $B(Y,W) \ll P^{3s/4+C\varepsilon}t^{s/2+2}$. We say that a family of terms of $N_1(P)+N_2(P)$ has *relative size at most* $R = R(Y,W)$ if, for every dyadic range, its contribution from that range is

$\ll P^{C\varepsilon} R\, B(Y,W)$, with C depending only on s. If $L<1$ there are no terms with $\mathbf{u}\neq\mathbf{0}$, up to $O(P^{-N})$, so we always assume $L\geqslant 1$ when we discuss these terms.

We now make precise how the relative size of a family is checked. By a *term* with modulus q we mean one of the summands τ of (5.1): the point term $p_{1,q}$, a plane term $p_{2,\vec\nu,k,q}$ or a flag term $p_{3,\vec\nu,\vec\rho,k,q}$; we write p_τ for its weight and $\mathcal{A}_\tau$ for its set of numerators, that is, the primitive $\vec a \bmod q$ whose $p_{\Lambda(\vec a,q)}$ contains p_τ. Since p_τ does not depend on $\vec a\in\mathcal{A}_\tau$, the contribution of τ to $N_1(P)+N_2(P)$ from the range (Y,W) is

$$q^{-s}\sum_{\mathbf{u}}\int_{\vec w\in\mathcal{W}} p_\tau(\vec w)\Big(\sum_{\vec a\in\mathcal{A}_\tau} S_q(\vec a,\mathbf{u})\Big) I_q(\vec w,\mathbf{u})\,d\vec w,$$

where $\mathcal{W}$ is the ball or shell of the range. For a weight function $0\leqslant\phi\leqslant 1$ on $\mathbb{R}^3$ put $m_\tau^\phi=\int_{|\vec w|\leqslant 2W}|p_\tau(\vec w)|\phi(\vec w)\,d\vec w$, and $m_\tau=m_\tau^1$.

In words: to show that a family has relative size R, it suffices to show that its sums over the numerators, summed over the dual vectors and weighted by the masses of the terms, are R times their trivial bound. The integrals enter only through their size, possibly improved by the weight ϕ of §14.11.

Lemma 14.5. *Let (Y,W) be a dyadic range with $L\geqslant 1$, and let $\mathcal{F}$ be a set of terms with moduli $q\in(Y,2Y]$. Let $\phi=1$, or $\phi(\vec w)=\min(1,(X\eta(\hat{\vec w}))^{-1/2})$. If*

$$\sum_{\tau\in\mathcal{F}}\sum_{0<|\mathbf{u}|\leqslant L}\Big|\sum_{\vec a\in\mathcal{A}_\tau} S_q(\vec a,\mathbf{u})\Big| m_\tau^\phi \leqslant R\,Q^{C\varepsilon}\,Y^{s/2}L^s\cdot Y\mathfrak{M}(Y,2W),$$

then the terms of $\mathcal{F}$ with $\mathbf{u}\neq\mathbf{0}$ contribute $\ll P^{C'\varepsilon}R\,B(Y,W)+O(P^{-N})$ from this range. Moreover $\sum_\tau \#\mathcal{A}_\tau m_\tau\ll Q^\varepsilon Y\mathfrak{M}(Y,2W)$, the sum being over all terms with $q\in(Y,2Y]$, so that the right side is R times the trivial bound $Y^{s/2}L^s\sum_\tau\#\mathcal{A}_\tau m_\tau$ summed over all terms, up to $Q^{C\varepsilon}$.

Proof. We bound the integrand of each term in absolute value, using the bound for the integral and the hypothesis for the sums, and then repeat the computation of Proposition 11.2. By Lemma 7.1(i) the dual vectors with $|\mathbf{u}|>L$ contribute $O(P^{-N})$. For $|\mathbf{u}|\leqslant L$ we bound the integrand in absolute value and use Lemma 7.1(ii), or Lemma 14.28(ii) when $\phi\neq 1$, in the form $|I_q(\vec w,\mathbf{u})|\ll P^s X^{-(s-2)/2}\phi(\vec w)$ on the range. So the contribution is

$$\begin{aligned}&\ll Y^{-s}P^sX^{-(s-2)/2}\sum_{\tau\in\mathcal{F}}\sum_{0<|\mathbf{u}|\leqslant L}\Big|\sum_{\vec a\in\mathcal{A}_\tau}S_q(\vec a,\mathbf{u})\Big| m_\tau^\phi\\ &\ll R\,Q^{C\varepsilon}Y^{-s/2}L^sP^sX^{-(s-2)/2}\,Y\mathfrak{M}(Y,2W).\end{aligned}$$

Since $L\ll P^\varepsilon YX/P$, we have $Y^{-s/2}L^sP^sX^{-(s-2)/2}\ll P^{s\varepsilon}Y^{s/2}X^{s/2+1}$, and the bound is $\ll P^{C'\varepsilon}RY^{s/2+1}X^{s/2+1}\mathfrak{M}(Y,2W)$, which is $P^{C'\varepsilon}R\,B(Y,W)$ up to the value of C. The last statement is the proof of Proposition 1.3, in which the masses of the point, plane and flag terms were bounded term by term. □

For the flag terms we use the lemma in the following form. Each flag term has the mass $m_\tau\ll Q^{2\varepsilon}\min(W,r_Y)\min(W,Q^{-4/3})^2$, independently of the numerator, and $Y^2Q^2\min(W,r_Y)\min(W,Q^{-4/3})^2\leqslant Y\mathfrak{M}(Y,2W)$ by (10.1). For the weight $\phi(\vec w)=\min(1,(X\eta(\hat{\vec w}))^{-1/2})$ we use (14.13) below: on the support of $p_{3,\vec\nu,\vec\rho,k,q}$, up to a part of mass $O(Q^{-N})$, $\phi(\vec w)\ll Q^\varepsilon\phi(\vec\rho)$ with $\phi(\vec\rho)=\min(1,(X\eta(\hat{\vec c}))^{-1/2})$ and $\vec c=\vec\rho/\gcd(\vec\rho)$.

The number of numerators of a flag term is $\#\mathcal{A}^\flat$ (Lemma 14.20 below). Put $U = Q^{1/3}$ and, for primitive $\vec{c}$ with $|\vec{c}| \leqslant Q^{2/3}$,

$$Z_{\vec{c}} = \frac{Q^{2/3}}{|\vec{c}|}, \qquad \Omega^+_{\vec{c}} = Y^2 Z_{\vec{c}} \Big(\frac{U}{\lambda_1(\vec{c})} + \frac{U^2}{|\vec{c}|} \Big), \tag{14.1}$$

where $\lambda_1(\vec{c})$ is the length of the shortest non-zero vector of $\vec{c}^{\perp} \cap \mathbb{Z}^3$. Then $\sum_{\vec{c}} \Omega^+_{\vec{c}} \ll Y^2 Q^2 \log Q$ (Lemma 14.21). By Lemma 14.5, a family $\mathcal{F}$ of flag terms in the range (Y, W) has relative size at most R provided that

$$\sum_{(\vec{\nu},\vec{\rho},q)\in\mathcal{F}} \sum_{0<|\mathbf{u}|\leqslant L} \Big| \sum_{\vec{a}\in\mathcal{A}^\flat} S_q(\vec{a},\mathbf{u}) \Big| \phi(\vec{\rho}) \ll Q^{C\varepsilon} R\, Y^{s/2} L^s\, Y^2 Q^2, \tag{14.2}$$

where $\phi(\vec{\rho}) = 1$, or, if we also use the improved bound for the integrals in §14.11, $\phi(\vec{\rho}) = \min(1, (X\eta(\hat{\vec{c}}))^{-1/2})$ with $\vec{c} = \vec{\rho}/\gcd(\vec{\rho})$. For the point and plane terms we use Lemma 14.5 directly, with their own masses and numerators.

The next lemma translates relative sizes into the final result. A family of relative size R contributes at most R times the bound $P^{3s/4} t^{s/2+2}$ of the unrefined method, and the main term is P^{s-6}; so the saving needed is $P^{3s/4-(s-6)} = P^{(24-s)/4}$, up to the factor $t^{s/2+2}$, which is harmless off the critical curve. This is the exponent θ_s.

Lemma 14.6. *Let $22 \leqslant s \leqslant 24$ and put $\theta_s = (24-s)/4$. Suppose that the terms of $N_1(P) + N_2(P)$ are divided into $O(1)$ families, each of relative size at most $P^{-\theta_s-\kappa} t^{-s/2-2}$ for some $\kappa > 0$ independent of the range. Then (1.1) holds with some $\Delta > 0$.*

Proof. Each family contributes

$$\ll P^{C\varepsilon-\theta_s-\kappa} P^{3s/4} = P^{s-6-\kappa+C\varepsilon}$$

from each of the $O((\log P)^2)$ ranges, since $3s/4 - \theta_s = s - 6$. Taking ε small and using Lemma 6.7(a) for the main term gives the result. □

So the relative saving needed on the critical curve is as follows:

s	needed	from λ and the integral	margin
22	$P^{1/2} = Q^{1/3}$	$Y^{1/2} X^{1/2} = Q^{1/2}$	$P^{1/4}$
23	$P^{1/4} = Q^{1/6}$	$Q^{1/2}$	$P^{1/2}$
24	P^{ε}	$Q^{1/2}$	$P^{3/4}$

Off the curve t is small, and the factor $t^{s/2+2}$ in $B(Y,W)$ absorbs the factors t^{-1}, t^{-2} which appear below through L. Since $L \asymp P^{1/2+\varepsilon} t$, a loss of the shape $1 + Y^{1/2}/L$ is at most $Q^{1/6+\varepsilon} t^{-1}$, which is affordable for $s \geqslant 23$.

The losses we meet are of three kinds: a gcd with a fixed non-zero integer, which we remove by summing over the modulus first; a power of the part of the modulus composed of the primes dividing a fixed integer, which is rare; and a power of $\gcd(q, \gcd(\vec{\rho}))$, which is controlled by Lemma 14.22. The first two are covered by the following elementary lemma, which we use repeatedly. For an integer $N \neq 0$ and $x \in \mathbb{N}$ write x_N for the largest divisor of x composed of primes dividing N.

Lemma 14.7. *Let $N \neq 0$ be an integer and $W \geqslant 1$.*

(i) $\sum_{x\leqslant W} \gcd(x, N) \leqslant W\tau(N)$, *and*

$$\sum_{x\leqslant W} C^{\omega(x)} \gcd(x,N)^{1/2} \ll_{C,\varepsilon} W^{1+\varepsilon}\tau(N).$$

(ii)
$$\#\{m \leqslant W : \ m \mid N^\infty\} \ll_\varepsilon (W|N|)^\varepsilon.$$

(iii) If $f \geqslant 0$ is a function on the integers composed of primes dividing N with $f(m) \leqslant m$, then $\sum_{x \leqslant W} f(x_N) \ll_\varepsilon W(W|N|)^\varepsilon$.

Proof. These are the elementary facts behind the phrase "removed by summing over q first": a gcd with a fixed integer is bounded on average, and so is any factor that depends only on the part of q made of primes from a fixed integer.

(i)
$$\sum_{x \leqslant W} \gcd(x, N) \leqslant \sum_{d \mid N} d \,\#\{x \leqslant W : \ d \mid x\} \leqslant W\tau(N);$$
the second bound follows in the same way, with $C^{\omega(x)} \ll x^\varepsilon$. (ii) This is Proposition 8.7(iii) for the primes dividing N: the number of such m is at most $\prod_{p \mid N}(1 + \log W/\log p) \ll (W|N|)^\varepsilon$. (iii) The sum is at most $\sum_{m \mid N^\infty, m \leqslant W} m \cdot W/m$, and (ii) applies. □

14.3. **Sums over the numerators in each term.** We now explain how the sums over λ enter. In the unrefined method we took absolute values term by term, for each numerator $\vec{a}$. Here we use instead that each term of the delta symbol has a set of numerators which is a union of full classes $\{\lambda\vec{a}\}$, so that its sum over numerators is a sum of the complete sums T_q studied above.

Every set of numerators that occurs in (5.1) is a union of classes $\{\lambda\vec{a}\}$: the point terms use all $\vec{a} \bmod q$, the plane terms use the $\vec{a}$ with $q \mid \vec{\nu} \cdot \vec{a}$, and the flag terms use the $\vec{a}$ with $\vec{\rho} \in \Lambda(\vec{a}, q)$ and $\vec{\nu} \in \Lambda^\perp(\vec{a}, q)$. Since each p-function is attached to the term and not to $\vec{a}$, we may sum over $\vec{a}$ inside each term, before integrating over $\vec{w}$:
$$\sideset{}{^*}\sum_{\vec{a} \in \mathcal{A}} S_q(\vec{a}, \mathbf{u}) = \sum_{[\vec{a}] \subseteq \mathcal{A}} T_q(\vec{a}, \mathbf{u}) \cdot \frac{\#[\vec{a}]}{\varphi(q)},$$
where $[\vec{a}] = \{\lambda\vec{a} \bmod q\}$ runs over the classes contained in the set $\mathcal{A}$. For a flag $(\vec{\nu}, \vec{\rho})$ with $\vec{\rho} = G\vec{c}$ and $\gcd(q, G) = 1$, Lemma 5.7(ii) shows that $\mathcal{A}$ is the single class of $\vec{c}$; when $d = \gcd(q, G) > 1$ it is the union of the classes of $\lambda\vec{c} + (q/d)\vec{b}$ with $\vec{b}$ modulo d. The exact set, which also involves the condition $\vec{\nu} \in \Lambda^\perp(\vec{a}, q)$, is described in Lemma 14.20.

14.4. **Averaging the gcd factors.** The factors $\gcd(q_1, F_{\vec{a}}^*(\mathbf{u}))^{1/2}$ and $E(\vec{a}, q_2)$ are removed on average by the following elementary lemma, applied with the sum over q taken first, as in the proof of Proposition 10.7.

Lemma 14.8. *Let N be a non-zero integer, $Y \geqslant 1$, and let f be a multiplicative function supported on the integers composed of primes dividing N, with $f(p^e) \leqslant p^{e/2}\gcd(p^e, N)^{1/2}$. Then*
$$\sum_{q \leqslant Y} \gcd(q, N)^{1/2} \leqslant Y\tau(N), \qquad \sum_{q \leqslant Y} f(q) \ll_\varepsilon Y|N|^\varepsilon.$$

Proof. Both bounds express the fact that a factor depending only on the part of q composed of primes dividing N is harmless on average, because such parts are rare. The first bound is Proposition 8.7(iii). For the second, write $q = q'm$ with m composed of primes dividing N and $\gcd(q', N) = 1$. Then
$$\sum_{q \leqslant Y} f(q) \leqslant Y \sum_m f(m)/m \leqslant Y \prod_{p \mid N} \sum_{e \geqslant 0} f(p^e)p^{-e}.$$

If $p^v \parallel N$, then $f(p^e)p^{-e} \leqslant p^{(\min(e,v)-e)/2}$, so the inner sum is at most $v+1+\sum_{j\geqslant 1} p^{-j/2} \leqslant v+4$. The product is $\leqslant 4^{\omega(N)}\tau(N) \ll |N|^{\varepsilon}$. □

For a fixed member $\vec{c}$ of the net, which is the situation of the flag terms with $\gcd(q,G)=1$, the following lemma carries out this averaging completely. Its point is that the large values of $K(\vec{c},q)$ are compensated by the fact that the dual vectors $\mathbf{u}$ with $T_q(\vec{c},\mathbf{u}) \neq 0$ lie in a coset of index $K(\vec{c},q)$, exactly as in Lemma 10.2. The only loss that survives is the factor $D_{s-1}(\vec{c})^{1/2}$ in (14.3), where $D_{s-1}(\vec{c})$ is the greatest common divisor of the $(s-1)\times(s-1)$ minors of $M_{\vec{c}}$.

Lemma 14.9. *Let $\vec{c} \in \mathbb{Z}^3$ be primitive with $\Delta(\vec{c}) \neq 0$, and put $N = 2D_F\Delta(\vec{c})$.*

(i) For each q there is a set $\mathcal{M}(\vec{c},q) \subseteq (\mathbb{Z}/q\mathbb{Z})^s$, either empty or a coset of a subgroup of index $K(\vec{c},q)$, such that $S_q(\lambda\vec{c},\mathbf{u}) = 0$ for every λ coprime to q unless $\mathbf{u} \bmod q \in \mathcal{M}(\vec{c},q)$. Moreover $\mathcal{M}(\vec{c},q)$ depends only on $\mathbf{u}$ modulo the part q_2 of q composed of primes dividing N.

(ii) For $Y, L \geqslant 1$,

$$\sum_{q\leqslant Y} \sum_{\substack{0<|\mathbf{u}|\leqslant L \\ F^*_{\vec{c}}(\mathbf{u})\neq 0}} \frac{|T_q(\vec{c},\mathbf{u})|}{\varphi(q)q^{(s-1)/2}} \ll_{\varepsilon} Y^{1+\varepsilon}(|\vec{c}|L)^{\varepsilon}\Big(L^s + L^{s-1}D_{s-1}(\vec{c})^{1/2}\Big). \tag{14.3}$$

(iii) For $Y, L \geqslant 1$,

$$\sum_{q\leqslant Y} \sum_{\substack{0<|\mathbf{u}|\leqslant L \\ F^*_{\vec{c}}(\mathbf{u})=0}} \frac{|T_q(\vec{c},\mathbf{u})|}{\varphi(q)q^{(s-1)/2}} \ll_{\varepsilon} Y^{3/2+\varepsilon}(|\vec{c}|L)^{\varepsilon}L^{s-2}.$$

Proof. Part (i) observes that the coset of Lemma 8.1 is the same for all the numerators $\lambda\vec{c}$, so it restricts $\mathbf{u}$ independently of λ. In part (ii) we split $q = q_1q_2$, where q_1 is coprime to N. At the primes of q_1 Lemma 14.3 gives square-root cancellation up to a gcd with $F^*_{\vec{c}}(\mathbf{u})$, which is 1 on average over q_1; at the primes of q_2 we use only the size of the Gauss sums together with the coset condition from (i), which compensates for the large values of $K(\vec{c},q_2)$ as in Lemma 10.2. The only loss which survives is the boundary term of the coset count, which produces $D_{s-1}(\vec{c})^{1/2}$. Part (iii) is the trivial bound on the quadric $F^*_{\vec{c}}(\mathbf{u}) = 0$.

(i) By Lemma 8.1, $S_q(\lambda\vec{c},\mathbf{u}) \neq 0$ only if $\mathbf{u} \bmod q \in \mathcal{M}(\lambda\vec{c},q)$, and $\mathcal{K}(\lambda\vec{c},q) = \mathcal{K}(\vec{c},q)$. For $\mathbf{h} \in \mathcal{K}(\vec{c},q)$ we have $M_{\vec{c}}\mathbf{h} \equiv \mathbf{0} \pmod q$, hence $2F_{\vec{c}}(\mathbf{h}) = \mathbf{h}^T M_{\vec{c}}\mathbf{h} \equiv 0 \pmod q$. If $p^e \parallel q$ with p odd this gives $F_{\vec{c}}(\mathbf{h}) \equiv 0 \pmod{p^e}$, and if $p=2$ it gives $F_{\vec{c}}(\mathbf{h}) \equiv 0 \pmod{2^{e-1}}$, so that $\lambda F_{\vec{c}}(\mathbf{h}) \equiv F_{\vec{c}}(\mathbf{h}) \pmod{2^e}$ for odd λ. In both cases $F_{\lambda\vec{c}}(\mathbf{h}) \equiv F_{\vec{c}}(\mathbf{h}) \pmod{p^e}$, so $\mathcal{M}(\lambda\vec{c},q) = \mathcal{M}(\vec{c},q)$, which is empty or a coset of index $K(\vec{c},q)$ by Lemma 8.1. If $p \nmid N$, then $M_{\vec{c}}$ is invertible modulo p, so $\mathcal{K}(\vec{c},p^e) = \{\mathbf{0}\}$ and the condition at p is empty. This proves (i).

(ii) Write $q = q_1q_2$ with $\gcd(q_1,N) = 1$ and $q_2 \mid N^{\infty}$. By Lemma 14.2 and Lemma 14.3(i) at the primes dividing q_1, and the trivial bound $|T_{p^e}| \leqslant \varphi(p^e)p^{es/2}K(\vec{c},p^e)^{1/2}$ at the primes dividing q_2 (note that $\Delta(m\vec{c}) = m^s\Delta(\vec{c})$ and $F^*_{m\vec{c}} = m^{s-1}F^*_{\vec{c}}$, so the twists in Lemma 14.2 do not change the relevant gcds), we get

$$\frac{|T_q(\vec{c},\mathbf{u})|}{\varphi(q)q^{(s-1)/2}} \leqslant \frac{q_1}{\varphi(q_1)} \gcd\Big(q_1, F^*_{\vec{c}}(\mathbf{u})\Big)^{1/2} \cdot q_2^{1/2}K(\vec{c},q_2)^{1/2} \cdot \mathbf{1}_{\mathbf{u} \bmod q_2 \in \mathcal{M}(\vec{c},q_2)}.$$

We sum first over $q_1 \leqslant Y/q_2$. Since $q_1/\varphi(q_1) \ll \log\log(3q_1)$ and $\sum_{m\leqslant Z}\gcd(m,n)^{1/2} \leqslant Z\tau(n)$ for $n \neq 0$, and $|F^*_{\vec{c}}(\mathbf{u})| \ll |\vec{c}|^{s-1}L^2$, the sum over q_1 is $\ll (Y/q_2)^{1+\varepsilon}(|\vec{c}|L)^{\varepsilon}$.

Next we sum over $\mathbf{u}$: by (i) and Lemma 10.1, the number of $\mathbf{u}$ with $|\mathbf{u}| \leqslant L$ and $\mathbf{u} \bmod q_2 \in \mathcal{M}(\vec{c}, q_2)$ is $\ll L^s/K(\vec{c}, q_2) + L^{s-1}$. Hence the left side of (14.3) is

$$\ll Y^{1+\varepsilon}(|\vec{c}|L)^{\varepsilon} \sum_{q_2 | N^\infty} q_2^{-1/2} \Big(L^s K(\vec{c}, q_2)^{-1/2} + L^{s-1} K(\vec{c}, q_2)^{1/2} \Big).$$

The first part is

$$\leqslant L^s \prod_{p|N} (1 - p^{-1/2})^{-1} \ll L^s |N|^{\varepsilon}.$$

For the second, Lemma 8.2(iv) gives $K(\vec{c}, p^e) \leqslant p^{v_p(D_{s-1}(\vec{c}))} p^{\min(v,e)}$ with $v = v_p(\Delta(\vec{c}))$, so

$$\sum_{e \geqslant 0} p^{-e/2} K(\vec{c}, p^e)^{1/2} \leqslant p^{v_p(D_{s-1}(\vec{c}))/2} \sum_{e \geqslant 0} p^{(\min(v,e)-e)/2} \leqslant (v+4)\, p^{v_p(D_{s-1}(\vec{c}))/2},$$

and the second part is

$$\leqslant L^{s-1} D_{s-1}(\vec{c})^{1/2} \prod_{p|N} (v_p(N) + 4) \ll L^{s-1} D_{s-1}(\vec{c})^{1/2} |N|^{\varepsilon}.$$

Since $|N| \ll |\vec{c}|^s$, this proves (ii).

(iii) Here $\gcd(q_1, F^*_{\vec{c}}(\mathbf{u})) = q_1$, and we use only $q_1^{1/2} q_2^{1/2} \leqslant Y^{1/2}$ and the count (14.12) below for the $\mathbf{u}$. It remains to bound

$$\sum_{q \leqslant Y} (q_1/\varphi(q_1)) K(\vec{c}, q_2)^{1/2} \ll Y^{1+\varepsilon} \sum_{q_2 | N^\infty} K(\vec{c}, q_2)^{1/2}/q_2.$$

For good p we have $K(\vec{c}, p^e) \leqslant p^{\min(v,2e)}$ by Lemma 8.4(i), and for bad p we have $K(\vec{c}, p^e) \leqslant p^{\mu_p + \min(v,2e)}$ by the same argument with Lemma 8.4(ii). In both cases

$$\sum_{e} p^{-e} K(\vec{c}, p^e)^{1/2} \ll p^{\mu_p/2} \sum_{e} p^{\min(v/2,e)-e} \leqslant p^{\mu_p/2}(v+2),$$

and the product over $p \mid N$ is $\ll |N|^{\varepsilon}$. □

Thus for a fixed non-singular member $\vec{c}$, and for moduli q coprime to $\gcd(\vec{\rho})$, the sum over λ saves the full factor $q^{1/2}$ up to Q^{ε}, apart from the loss $D_{s-1}(\vec{c})^{1/2}/L$, and the dual vectors on the dual quadric cost a factor $Y^{1/2}/L^2$. In this section we do not need the bound in terms of $D_{s-1}(\vec{c})$: we shall only use that $K(\vec{c}, q) \leqslant c_F q^2$, which gives the loss $\min(Y, D_{s-1}(\vec{c}))^{1/2}/L \leqslant Y^{1/2}/L$. The finer bound is used in Section 15. The moduli which share a factor with $\gcd(\vec{\rho})$ are treated in §14.8.

14.5. **Point and plane terms: reduction to good squarefree moduli.** For the point terms the weight is the same for all primitive $\vec{a}$ modulo q, and for a plane term it is the same for all primitive $\vec{a}$ with $q \mid \vec{\nu} \cdot \vec{a}$. So in these terms we may sum $S_q(\vec{a}, \mathbf{u})$ over a complete set $\mathcal{A}$ of numerators before taking absolute values. The guiding principle of this subsection is to use the cancellation in this sum only at the primes p with $p \parallel q$ and $p \nmid 2D_F$, where it is elementary, and to use the bounds of Theorem 6.5 at all other primes. The moduli with a large part of the second kind are rare (Lemma 14.13), and at the primes of the first kind the averages over $\vec{a}$ are densities modulo p, so no equidistribution of $\mathbf{u}$ in boxes is needed.

For $\mathbf{u} \in \mathbb{Z}^s$ let $f_{\mathbf{u}}(\vec{a}) = F^*_{\vec{a}}(\mathbf{u})$, a ternary form of degree $s-1$ in $\vec{a}$ with integer coefficients, and let $C(\mathbf{u})$ be its content, the greatest common divisor of its coefficients. Thus $C(\mathbf{u}) = 0$ exactly when $\mathbf{u} \in E_0$, the set defined in §14.9, which has dimension at most $s-3$. For primitive $\vec{\nu}_0$ choose a basis $\mathbf{w}_1, \mathbf{w}_2$ of $\vec{\nu}_0^{\perp} \cap \mathbb{Z}^3$ and let $C_{\vec{\nu}_0}(\mathbf{u})$ be

the content of the binary form $(x,y)\mapsto f_{\mathbf{u}}(x\mathbf{w}_1+y\mathbf{w}_2)$; it does not depend on the choice of basis.

Lemma 14.10. *Let $q=\prod_p p^{n_p}$ and let $\mathcal{A}\subseteq(\mathbb{Z}/q\mathbb{Z})^3$ be a set of primitive residues of the form $\mathcal{A}=\prod_p\mathcal{A}_p$ under the Chinese remainder theorem, where each $\mathcal{A}_p\subseteq(\mathbb{Z}/p^{n_p}\mathbb{Z})^3$ is stable under multiplication by units. Then*

$$\sum_{\vec{a}\in\mathcal{A}} S_q(\vec{a},\mathbf{u})=\prod_{p\mid q}\sum_{\vec{a}\in\mathcal{A}_p} S_{p^{n_p}}(\vec{a},\mathbf{u}).$$

The sets of numerators of the point terms and of the plane terms are of this form: for the point terms $\mathcal{A}_p$ consists of all primitive residues modulo p^{n_p}, and for a plane term with $\vec{\nu}=g\vec{\nu}_0$ it consists of the primitive $\vec{a}$ with $p^{n_p-f}\mid\vec{\nu}_0\cdot\vec{a}$, where $p^f=\gcd(p^{n_p},g)$.

Proof. The point is that the sets of numerators are themselves products over primes, so the sum over numerators factorises like the exponential sums. If $q=q_1q_2$ with $\gcd(q_1,q_2)=1$, writing $\mathbf{b}=q_2\mathbf{b}_1+q_1\mathbf{b}_2$ in the definition of S_q as in the proof of Lemma 14.2 gives $S_q(\vec{a},\mathbf{u})=S_{q_1}(q_2\vec{a},\mathbf{u})S_{q_2}(q_1\vec{a},\mathbf{u})$. By induction, $S_q(\vec{a},\mathbf{u})=\prod_p S_{p^{n_p}}(m_p\vec{a},\mathbf{u})$ with $m_p=q/p^{n_p}$. Summing over $\vec{a}\in\mathcal{A}$ and substituting $\vec{a}\mapsto\overline{m_p}\vec{a}$ in the p-th factor, which permutes $\mathcal{A}_p$, gives the formula. The descriptions of $\mathcal{A}_p$ are immediate from $q\mid\vec{\nu}\cdot\vec{a}\iff p^{n_p}\mid g\vec{\nu}_0\cdot\vec{a}$ for all p. □

The next lemma is the local estimate at a prime $p\parallel q$. The sum over all numerators in a subspace W of $\mathbb{F}_p^3$ is grouped into the classes $\{\lambda\vec{a}\}$, that is, into the points of the projective space $\mathbb{P}(W)$. For almost every point Lemma 14.3 gives the saving $p^{1/2}$; the exceptional points are those where Δ or the dual form $F_{\vec{a}}^*(\mathbf{u})$ vanishes modulo p, and there are few of them unless one of these forms vanishes identically on W.

Lemma 14.11. *Let $p\nmid 2D_F$ be a prime, let $W\subseteq\mathbb{F}_p^3$ be a subspace of dimension $r\in\{2,3\}$, and let $\mathbf{u}\in\mathbb{Z}^s$. Let $\mathbb{P}(W)$ be the set of lines in W, of cardinality $N=(p^r-1)/(p-1)$. Then*

$$\Big|\sum_{\vec{a}\in W\setminus\{\mathbf{0}\}} S_p(\vec{a},\mathbf{u})\Big|\leqslant(p^r-1)\,p^{(s-1)/2}\cdot\frac{p}{p-1}\cdot\beta,$$

where

$$\beta\leqslant 1+2(s-1)p^{-1/2}+2s+p^{1/2}\mathbf{1}_{f_{\mathbf{u}}|_W\equiv 0}+p\,\mathbf{1}_{\Delta|_W\equiv 0}+\frac{p^{3/2}}{N}\,n_2(W),$$

and $n_2(W)\leqslant(s-1)^2$ is the number of points of $\mathbb{P}(W)$ at which $M_{\vec{a}}$ has corank 2 modulo p.

Proof. We bound each class sum $T_p(\vec{a},\mathbf{u})$ according to the behaviour of $F_{\vec{a}}$ modulo p, and then count the classes of each kind with the elementary bound for the number of zeros of a form over $\mathbb{F}_p$. Group the $\vec{a}\in W\setminus\{\mathbf{0}\}$ into the classes $\{\lambda\vec{a}:\lambda\in\mathbb{F}_p^*\}$, so that the sum equals $\sum_{[\vec{a}]\in\mathbb{P}(W)}T_p(\vec{a},\mathbf{u})$. If $p\nmid\Delta(\vec{a})$, Lemma 14.3 with $e=1$ gives $|T_p(\vec{a},\mathbf{u})|\leqslant p^{s/2}\max(|c_p(t)|,|g_p(t)|)$, which is at most $p^{(s+1)/2}$ if $p\nmid F_{\vec{a}}^*(\mathbf{u})$ and at most $p^{(s+1)/2}p^{1/2}$ otherwise. If $p\mid\Delta(\vec{a})$, Lemma 8.1 gives

$$|T_p(\vec{a},\mathbf{u})|\leqslant(p-1)p^{s/2}K(\vec{a},p)^{1/2}\leqslant p^{(s+1)/2}p^{(1+k)/2},$$

where $k\in\{1,2\}$ is the corank of $M_{\vec{a}}$ modulo p (at most 2 by Lemma 6.1(i) over $\overline{\mathbb{F}}_p$), since $K(\vec{a},p)=p^k$ by Lemma 8.2(i). Hence the sum is at most $p^{(s+1)/2}N\beta$ with

$$\beta=1+\frac{p^{1/2}}{N}\#\{[\vec{a}]:\ p\nmid\Delta(\vec{a}),\ p\mid F_{\vec{a}}^*(\mathbf{u})\}+\frac{1}{N}\sum_{[\vec{a}]:\ p\mid\Delta(\vec{a})}p^{(1+k(\vec{a}))/2},$$

and $p^{(s+1)/2}N = (p^r-1)p^{(s-1)/2}\cdot p/(p-1)$.

A non-zero form of degree d in r variables over $\mathbb{F}_p$ has at most dp^{r-1} zeros in $\mathbb{F}_p^r$, hence at most $2dp^{r-2}$ zeros in $\mathbb{P}(W)$ when $r\geqslant 2$, while $N\geqslant p^{r-1}$. If $f_{\mathbf{u}}|_W\not\equiv 0$, the first count is therefore at most $2(s-1)p^{r-2}\leqslant 2(s-1)N/p$, which gives the term $2(s-1)p^{-1/2}$; otherwise we bound it by N, which gives $p^{1/2}$. The classes of corank 1 contribute at most $p\cdot 2sp^{r-2}/N\leqslant 2s$ if $\Delta|_W\not\equiv 0$, and at most p otherwise. The classes of corank 2 contribute $p^{3/2}n_2(W)/N$, and they correspond to points of $\mathcal{Z}_2(\mathbb{F}_p)$, of which there are at most $(s-1)^2$, as in the proof of Proposition 8.7(i). □

For $r=3$ we have $\Delta|_W\not\equiv 0$ by Lemma 6.1(iv) over $\overline{\mathbb{F}}_p$, and $p^{3/2}n_2/N\leqslant (s-1)^2p^{-1/2}$, so $\beta\leqslant C_s\gcd(p,C(\mathbf{u}))^{1/2}$ with $C_s=4s+(s-1)^2$. For $r=2$ and $W=\vec{\nu}_0^{\perp}$ the three exceptional terms occur only when $p\mid C_{\vec{\nu}_0}(\mathbf{u})$, when p divides the content of $G_{\vec{\nu}_0}$ (and hence $R_{\vec{\nu}_0}$), and when the line $\mathbb{P}(W)$ contains a point of corank 2.

At the bad primes we cannot use the geometry of the net modulo p, but we can use it over $\mathbb{Q}_p$. The next lemma is the analogue at bad primes of the local estimates of Section 8, with a loss of a constant depending on p; since there are only finitely many bad primes, this loss is harmless.

Lemma 14.12. *For every prime p there is C_p such that*

$$\mu\{\vec{a}\in\mathbb{Z}_p^3 \text{ primitive}:\ v_p(\Delta(\vec{a}))\geqslant j\}\leqslant C_p p^{-j/2}$$

for all $j\geqslant 0$. Consequently $\sum^{*}_{\vec{a}\bmod p^n} K(\vec{a},p^n)^{1/2}\ll_p p^{3n}(n+1)$ for all $n\geqslant 1$.

Proof. For good p this is contained in Lemma 8.6 and the proof of Proposition 8.7(i); the argument below works for every p, and it is a p-adic version of the proof of Lemma 8.6 with a bounded loss. Let $\mathcal{P}$ be the compact set of primitive vectors in $\mathbb{Z}_p^3$. For $\vec{x}\in\mathcal{P}$ and $\vec{b}\in\mathbb{Z}_p^3$ let $Q_{\vec{x}}(\vec{b})$ be the coefficient of t^2 in $\Delta(\vec{x}+t\vec{b})$, a quadratic form in $\vec{b}$ whose coefficients are polynomials in $\vec{x}$. Let $\mathcal{V}=\{\mathbf{e}_i,\ \mathbf{e}_i+\mathbf{e}_j\}$; a quadratic form in three variables which vanishes on $\mathcal{V}$ vanishes identically. We first show that for every $\vec{x}\in\mathcal{P}$

$$\Delta(\vec{x})\neq 0,\qquad\text{or}\qquad \nabla\Delta(\vec{x})\neq\vec{0},\qquad\text{or}\qquad Q_{\vec{x}}(\vec{v})\neq 0 \text{ for some } \vec{v}\in\mathcal{V}.$$

Indeed, $M_{\vec{x}}$ has corank at most 2 by Lemma 6.1(i) over $\overline{\mathbb{Q}}_p$. If the corank is 1, then $\nabla\Delta(\vec{x})\neq\vec{0}$ by Lemma 6.1(ii). If it is 2, write $W=\ker M_{\vec{x}}$ and use a basis of $\overline{\mathbb{Q}}_p^s$ adapted to $W\oplus W'$, in which $M_{\vec{x}}=\operatorname{diag}(0,N)$ with N invertible; expanding the determinant gives $\Delta(\vec{x}+t\vec{b})=c\,t^2\det N\det(M_{\vec{b}}|_W)+O(t^3)$ with $c\neq 0$ (the square of the determinant of the change of basis), so $Q_{\vec{x}}(\vec{b})=c\det N\det(M_{\vec{b}}|_W)$, which is not identically zero by Lemma 6.1(v).

The function $\vec{x}\mapsto\max\big(|\Delta(\vec{x})|_p,|\nabla\Delta(\vec{x})|_p,\max_{\vec{v}\in\mathcal{V}}|Q_{\vec{x}}(\vec{v})|_p\big)$ is continuous and positive on $\mathcal{P}$, so it is $\geqslant p^{-c_0}$ for some integer $c_0\geqslant 0$. Put $t=c_0+1$ and cover $\mathcal{P}$ by the $\leqslant p^{3t}$ balls $B=\vec{x}_0+p^t\mathbb{Z}_p^3$ with $\vec{x}_0\in\mathcal{P}$. For each ball one of the three quantities at $\vec{x}_0$ has valuation at most c_0, and we write $\vec{x}=\vec{x}_0+p^t\vec{y}$.

Case (a): $v_p(\Delta(\vec{x}_0))\leqslant c_0$. Then $\Delta(\vec{x})\equiv\Delta(\vec{x}_0)\pmod{p^t}$, so $v_p(\Delta)=v_p(\Delta(\vec{x}_0))\leqslant c_0$ on B, and $B\cap\{v_p(\Delta)\geqslant j\}=\varnothing$ for $j>c_0$.

Case (b): some coordinate of $\nabla\Delta(\vec{x}_0)$, say the first, has valuation $a\leqslant c_0$. For fixed y_2,y_3 put $g(y_1)=\Delta(\vec{x}_0+p^t\vec{y})$. Since Δ has integral coefficients, $g(y_1)-g(y_1')=(y_1-y_1')\,h$ with $h=p^t\partial_1\Delta(\vec{x}_0)+p^{2t}h_1$, $h_1\in\mathbb{Z}_p$, so $v_p(h)=t+a$ because $t>a$. Hence two values y_1,y_1' with $v_p(g)\geqslant j$ satisfy $v_p(y_1-y_1')\geqslant j-t-a$, and the y_1-measure of $\{v_p(g)\geqslant j\}$ is at most $p^{-(j-t-c_0)}$.

Case (c): $v_p(Q_{\vec{x}_0}(\vec{v})) = e \leqslant c_0$ for some $\vec{v} \in \mathcal{V}$. The vector $\vec{v}$ is primitive; complete it to a basis $\vec{v}, \vec{w}_1, \vec{w}_2$ of $\mathbb{Z}_p^3$ and write $\vec{y} = \tau\vec{v} + \vec{y}_0$ with $\vec{y}_0 \in \mathbb{Z}_p\vec{w}_1 + \mathbb{Z}_p\vec{w}_2$. For fixed $\vec{y}_0$ put $\vec{z} = \vec{x}_0 + p^t\vec{y}_0$ and $g(\tau) = \Delta(\vec{z} + p^t\tau\vec{v}) = \sum_i A_i\tau^i$. Then $A_i = p^{ti}c_i$ with $c_i \in \mathbb{Z}_p$ the coefficient of u^i in $\Delta(\vec{z} + u\vec{v})$, and $c_2 = Q_{\vec{z}}(\vec{v}) \equiv Q_{\vec{x}_0}(\vec{v}) \pmod{p^t}$, so $v_p(c_2) = e$ and $v_p(A_2) = 2t + e$. Lemma 8.5 applies to g/A_2, whose coefficients of τ^i, $i \geqslant 3$, have valuation $\geqslant t(i-2) - e \geqslant 1$; so the τ-measure of $\{v_p(g) \geqslant j\}$ is at most $2p^{-(j-2t-e)/2}$.

In all three cases

$$\mu(B \cap \{v_p(\Delta) \geqslant j\}) \leqslant 2p^{t+c_0}p^{-j/2}\mu(B),$$

and summing over the balls gives the first statement with $C_p = 2p^{t+c_0}$. For the second, Lemma 8.4(ii) and its proof give $K(\vec{a}, p^n) \leqslant p^{\mu_p}p^{\min(v_p(\Delta(\vec{a})),2n)}$, so

$$K(\vec{a}, p^n)^{1/2} \leqslant p^{\mu_p/2} \sum_{0 \leqslant j \leqslant \min(v_p(\Delta(\vec{a})),2n)} p^{j/2},$$

and

$$\sum_{\vec{a} \bmod p^n}^{*} K(\vec{a}, p^n)^{1/2} \leqslant p^{3n}p^{\mu_p/2} \sum_{j \leqslant 2n} p^{j/2}\mu\{v_p(\Delta) \geqslant j\} \leqslant p^{3n}p^{\mu_p/2}C_p(2n+1).$$

This argument uses only Lemma 6.1 over $\overline{\mathbb{Q}}_p$, Hensel-type estimates and Lemma 8.5; in particular no resolution of singularities is needed. □

Lemma 14.13. *For $q \in \mathbb{N}$ let $q^\flat$ be the product of the primes p with $p \parallel q$ and $p \nmid 2D_F$, and $q^\sharp = q/q^\flat$. Let κ be a multiplicative function with $\kappa(p^n) \leqslant 2(n+1)(1 + 13s^2p^{-1/2})$ for $p \nmid 2D_F$ and $\kappa(p^n) \leqslant C_p(n+1)^3$ for $p \mid 2D_F$. Then for any non-zero integer N,*

$$\sum_{q \leqslant Y} (q^\sharp)^{1/2}\kappa(q^\sharp)\, C_s^{\omega(q^\flat)} \gcd(q^\flat, N)^{1/2} \ll_\varepsilon Y^{1+\varepsilon}|N|^\varepsilon.$$

Proof. The square-full numbers up to Y are only about $Y^{1/2}$ in number, which is exactly what is needed to compensate the factor $(q^\sharp)^{1/2}$. Every prime factor of $q^\sharp$ divides $2D_F$ or occurs in $q^\sharp$ with exponent at least 2. Summing first over $q^\flat \leqslant Y/q^\sharp$, using $C_s^{\omega(m)} \ll m^\varepsilon$ and $\sum_{m \leqslant Z} \gcd(m, N)^{1/2} \leqslant Z\tau(N)$, the sum is

$$\ll Y^{1+\varepsilon}\tau(N) \sum_{q^\sharp} \kappa(q^\sharp)(q^\sharp)^{-1/2}.$$

The last sum is at most

$$\prod_{p \mid 2D_F} \sum_{n \geqslant 0} C_p(n+1)^3 p^{-n/2} \cdot \prod_{p \leqslant Y} \Big(1 + \sum_{n \geqslant 2} 2(n+1)(1 + 13s^2p^{-1/2})p^{-n/2}\Big)$$
$$\ll \prod_{p \leqslant Y} \Big(1 + \frac{C}{p}\Big) \ll (\log Y)^C.$$

□

Proposition 14.14. *Let $\mathbf{u} \in \mathbb{Z}^s$ with $C(\mathbf{u}) \neq 0$, and let $J(q)$ be the number of primitive residues modulo q. Then $|\sum^{*}_{\vec{a} \bmod q} S_q(\vec{a}, \mathbf{u})| \leqslant J(q)q^{(s-1)/2}A(q, \mathbf{u})$ with*

$$\sum_{q \leqslant Y} A(q, \mathbf{u}) \ll_\varepsilon Y^{1+\varepsilon}(1 + |\mathbf{u}|)^\varepsilon.$$

Proof. At the primes dividing q exactly once we use the cancellation of Lemma 14.11, and at all other prime powers only the size of the Gauss sums, averaged over $\vec{a}$ by Proposition 8.7(i). The resulting losses are of the two kinds that Lemma 14.13 controls. By Lemma 14.10 the sum is the product of its local factors. For $p \parallel q$ with

$p \nmid 2D_F$, Lemma 14.11 with $W = \mathbb{F}_p^3$ and the remark after it bound the local factor by $J(p)p^{(s-1)/2} \cdot 2C_s \gcd(p, C(\mathbf{u}))^{1/2}$. For the other prime powers $p^n \parallel q$, Lemma 8.1 gives the bound

$$p^{ns/2} \sideset{}{^*}\sum_{\vec{a} \bmod p^n} K(\vec{a}, p^n)^{1/2} = J(p^n)p^{n(s-1)/2} \cdot p^{n/2}\kappa(p^n),$$

where $\kappa(p^n) = J(p^n)^{-1} \sideset{}{^*}\sum K(\vec{a}, p^n)^{1/2}$ satisfies the hypothesis of Lemma 14.13 by the proof of Proposition 8.7(i) for $p \nmid 2D_F$ (note $J(p^n) \geqslant p^{3n}/2$) and by Lemma 14.12 for $p \mid 2D_F$. Hence

$$A(q, \mathbf{u}) \leqslant (2C_s)^{\omega(q^\flat)} \gcd(q^\flat, C(\mathbf{u}))^{1/2} (q^\sharp)^{1/2} \kappa(q^\sharp),$$

and Lemma 14.13 with $N = C(\mathbf{u})$, which is $\ll |\mathbf{u}|^2$, completes the proof. □

Thus for the point terms the sum over the numerators saves the full factor $q^{1/2}$ on average over q, for all $\mathbf{u} \notin E_0$, with no condition on the size of $\mathbf{u}$ relative to q. The vectors $\mathbf{u} \in E_0$ number $\ll L^{s-3}$ in a box of side L.

For a plane term with $\vec{\nu} = g\vec{\nu}_0$ and $p \parallel q$, $p \nmid 2D_F$, the local set is $\mathbb{F}_p^3 \setminus \{\mathbf{0}\}$ if $p \mid g$ and $\vec{\nu}_0^{\perp} \setminus \{\mathbf{0}\}$ otherwise, and Lemma 14.11 applies with $r = 3$ or $r = 2$. With $r = 2$ its exceptional terms cost at most a factor p at the primes dividing $C_{\vec{\nu}_0}(\mathbf{u})$, the content of $G_{\vec{\nu}_0}$, or the resultant $R'_{\vec{\nu}_0}$ that detects points of $\mathcal{Z}_2(\mathbb{F}_p)$ on the line $\vec{\nu}_0^{\perp}$ which are not reductions of points of $\mathcal{Z}_2(\overline{\mathbb{Q}})$ on that line; such losses are removed by summing over q first (Lemma 14.7). The part $q^\sharp$ of the modulus is treated in §14.6 below, and the lines through rational members of corank 2 in Proposition 14.18.

14.6. **Plane terms at prime powers.** At a prime power $p^n \parallel q$ with $n \geqslant 2$ we give up the cancellation in the sum over the numerators, as for the point terms, and bound the local factor by $p^{ns/2} \sum_{\vec{a} \in \mathcal{A}_p} K(\vec{a}, p^n)^{1/2}$. For the point terms the average of $K^{1/2}$ was bounded in Proposition 8.7(i). For a plane term the numerators lie on the line $\vec{\nu}_0^{\perp}$, along which Δ is the binary form $G_{\vec{\nu}_0}$, and the average of $K^{1/2}$ along the line depends on how the roots of $G_{\vec{\nu}_0}$ cluster p-adically. We show that the loss is controlled by a single non-zero integer attached to $\vec{\nu}_0$, so that it can be removed by summing over q first, as in Lemma 14.8.

We use the following notation. Let $\mathcal{O}$ be the ring of integers of a finite extension of $\mathbb{Q}_p$, with valuation v extending v_p. For primitive vectors $\mathbf{z}, \mathbf{z}' \in \mathcal{O}^2$ put $d(\mathbf{z}, \mathbf{z}') = v(\mathbf{z} \times \mathbf{z}')$, where $\mathbf{z} \times \mathbf{z}' = z_1 z_2' - z_2 z_1'$; this is the chordal distance on $\mathbb{P}^1$, and it satisfies the ultrametric inequality $d(\mathbf{z}, \mathbf{z}'') \geqslant \min(d(\mathbf{z}, \mathbf{z}'), d(\mathbf{z}', \mathbf{z}''))$. For a non-zero binary form $G \in \mathbb{Z}[x, y]$ of degree d, Gauss's lemma gives a factorisation $G = p^\kappa u \prod_{i=1}^d (\beta_i x - \alpha_i y)$ over a suitable $\mathcal{O}$, with $\kappa = v_p(\operatorname{cont} G)$, $u \in \mathcal{O}^*$ and (α_i, β_i) primitive, so that

$$v_p(G(\mathbf{z})) = \kappa + \sum_{i=1}^{d} d(\mathbf{z}, \boldsymbol{\alpha}_i) \qquad (\mathbf{z} \in \mathbb{Z}_p^2 \text{ primitive}), \tag{14.4}$$

with $\boldsymbol{\alpha}_i = (\alpha_i, \beta_i)$. Let $\delta_3(G)$ be the largest δ such that some three distinct indices i, j, k have pairwise distances at least δ (so $\delta_3 = \infty$ exactly when G has a root of multiplicity at least 3), and put $e_p(G) = \kappa + d\,\delta_3(G)$.

The intuition is that $v_p(G(\mathbf{z}))$ is the sum of the p-adic distances from $\mathbf{z}$ to the roots of G, by (14.4). If no three roots are p-adically close together, then $\mathbf{z}$ can be close to at most two roots at the same time, and the binary form behaves locally like a quadratic; Lemma 14.15 then gives the same square-root bound as Lemma 8.5, with

a loss $e_p(G)$ which measures how close together the roots are, and which is bounded in terms of a single non-zero integer E_G.

Lemma 14.15. *Let $G \in \mathbb{Z}[x,y]$ be a non-zero binary form of degree d without roots of multiplicity at least 3, and let μ be the Haar measure on $\mathbb{Z}_p^2$.*

(i) $\delta_3(G) \leqslant v_p(\mathrm{Disc}(G^{\mathrm{red}}))$, where G^{red} is the primitive integral form whose roots are those of G, each with multiplicity one; hence $e_p(G) \leqslant v_p(E_G)$ with $E_G = \mathrm{cont}(G)\,\mathrm{Disc}(G^{\mathrm{red}})^d \neq 0$.

(ii) For every $j \geqslant 0$,

$$\mu\{\mathbf{z} \text{ primitive} : \ v_p(G(\mathbf{z})) \geqslant j\} \leqslant d\, p^{-\lceil (j-e_p(G))/2 \rceil}.$$

Proof. Part (i) bounds the closeness of the roots by the discriminant. For part (ii), a point $\mathbf{z}$ with $v_p(G(\mathbf{z})) \geqslant j$ must, by (14.4) and the ultrametric inequality, be within distance about $(j-e_p)/2$ of one of the roots, and the set of such points has small measure.

(i) Three distinct indices involve at least two distinct roots $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$ of G^{red}, and the discriminant of G^{red} is a unit times the product of the squares of the $\boldsymbol{\alpha}_i \times \boldsymbol{\alpha}_j$ over its pairs of roots, each of which has non-negative valuation. Hence $d(\boldsymbol{\alpha}, \boldsymbol{\alpha}') \leqslant v_p(\mathrm{Disc}(G^{\mathrm{red}}))$.

(ii) We may assume $j > e_p(G)$. Let $\mathbf{z}$ be primitive with $v_p(G(\mathbf{z})) \geqslant j$, and let $t = \max_i d(\mathbf{z}, \boldsymbol{\alpha}_i)$, attained at i_0. If $t \leqslant \delta_3$, then (14.4) gives $v_p(G(\mathbf{z})) \leqslant \kappa + d\delta_3 < j$; so $t > \delta_3$. If two indices k_1, k_2 other than i_0 had $d(\mathbf{z}, \boldsymbol{\alpha}_k) > \delta_3$, then by the ultrametric inequality the three indices i_0, k_1, k_2 would have pairwise distances $> \delta_3$, which is impossible. So at most one index other than i_0 has $d(\mathbf{z}, \boldsymbol{\alpha}_k) > \delta_3$, and for it $d(\mathbf{z}, \boldsymbol{\alpha}_k) \leqslant t$. Thus $j \leqslant \kappa + 2t + (d-2)\delta_3 \leqslant e_p(G) + 2t$, so

$$d(\mathbf{z}, \boldsymbol{\alpha}_{i_0}) \geqslant m := \lceil (j - e_p(G))/2 \rceil.$$

If the set of primitive $\mathbf{z} \in \mathbb{Z}_p^2$ with $d(\mathbf{z}, \boldsymbol{\alpha}_i) \geqslant m$ is non-empty, it contains some $\mathbf{z}_0$, and by the ultrametric inequality it equals $\{\mathbf{z} : \ v_p(\mathbf{z} \times \mathbf{z}_0) \geqslant m\}$. Moving $\mathbf{z}_0$ to $(1,0)$ by an element of $\mathrm{GL}_2(\mathbb{Z}_p)$, which preserves μ, this set becomes $\{(x,y) : \ p \nmid x, \ p^m \mid y\}$, of measure $(1-p^{-1})p^{-m} \leqslant p^{-m}$. Summing over the d indices gives (ii). □

We now apply this to the restriction of Δ to a line. The new feature, compared with the average over all residues in Proposition 8.7(i), is that $K(\vec{a}, p^n)$ depends on the two non-unit invariant factors of $M_{\vec{a}}$ separately, and not only on their product $\Delta(\vec{a})$; the smaller of the two is controlled by the corank-two points through Lemma 15.2.

Lemma 14.16. *Let p be a prime, let $\vec{\nu}_0 \notin \mathcal{N}_F$ be primitive, let $\mathbf{w}_1, \mathbf{w}_2$ be a basis of $\vec{\nu}_0^{\perp} \cap \mathbb{Z}^3$, and put $e = e_p(G_{\vec{\nu}_0})$. For $n \geqslant 1$ and $0 \leqslant f \leqslant n$ let $\mathcal{A}_p$ be the set of primitive $\vec{a} \bmod p^n$ with $p^{n-f} \mid \vec{\nu}_0 \cdot \vec{a}$. Then*

$$\frac{1}{\#\mathcal{A}_p} \sum_{\vec{a} \in \mathcal{A}_p} K(\vec{a}, p^n)^{1/2} \leqslant C_s\, p^f (n+1)^2 p^{\min(e,n)/2},$$

with C_s depending only on s if $p \nmid D_F$, and on $\vec{F}$ and p if $p \mid D_F$.

Proof. We parametrise the line, so that Δ becomes the binary form $G_{\vec{\nu}_0}$, and split the residues according to the valuation k of $D_{s-1}(\vec{a})$. The residues with $k \geqslant 1$ lie in a few small p-adic discs, by Lemma 15.2, and on each set with fixed k Lemma 14.15 bounds how often $G_{\vec{\nu}_0}$ is highly divisible.

First let $f = 0$. Since $\vec{\nu}_0^{\perp} \cap \mathbb{Z}^3$ is saturated, the map $\mathbf{z} = (x,y) \mapsto \vec{a}(\mathbf{z}) = x\mathbf{w}_1 + y\mathbf{w}_2$ identifies the primitive $\mathbf{z} \bmod p^n$ with $\mathcal{A}_p$, and $\Delta(\vec{a}(\mathbf{z})) = G_{\vec{\nu}_0}(\mathbf{z})$; so we have to bound

$\frac{1}{\mu(\mathrm{prim})}\int K(\vec a(\mathbf z),p^n)^{1/2}\,d\mu(\mathbf z)$ over the primitive $\mathbf z\in\mathbb Z_p^2$, where $\mu(\mathrm{prim})=1-p^{-2}$. As p is good, $M_{\vec a}$ has corank at most 2 modulo p, so its first $s-2$ invariant factors are units; let $v_1\leqslant v_2$ be the valuations of the last two. By Lemma 8.2(i), $K(\vec a,p^n)=p^{\min(v_1,n)+\min(v_2,n)}$, and $v_1+v_2=v_p(G_{\vec\nu_0}(\mathbf z))$, $v_1=v_p(D_{s-1}(\vec a))$.

By Lemma 15.2, the primitive $\vec a$ with $p^k\mid D_{s-1}(\vec a)$ lie in at most $(s-1)^2$ classes $\{\lambda\vec\beta \bmod p^k\}$, and each class meets the line in at most one class $\{\lambda\mathbf z_0 \bmod p^k\}$ of primitive $\mathbf z$, of measure at most p^{-k}. So the set $U_k=\{\mathbf z:\ v_1\geqslant k\}$ is a union of at most $(s-1)^2$ sets D of measure at most p^{-k} (with U_0 the whole set). On $U_k\setminus U_{k+1}$ we have $v_1=k$ and $v_2=v_p(G_{\vec\nu_0}(\mathbf z))-k$, so

$$K^{1/2}\leqslant p^{\min(k,n)/2}\Big(1+\sum_{j=1}^{n}p^{j/2}\mathbf 1_{v_p(G_{\vec\nu_0}(\mathbf z))\geqslant k+j}\Big).$$

By Lemma 14.15(ii), each D satisfies

$$\mu\{\mathbf z\in D:\ v_p(G_{\vec\nu_0}(\mathbf z))\geqslant k+j\}\leqslant\min(p^{-k},s\,p^{-(k+j-e)/2})\leqslant s\,p^{-k}\min(1,p^{(k+e-j)/2}),$$

and

$$\sum_{j\leqslant n}p^{j/2}\min(1,p^{(k+e-j)/2})\leqslant(n+3)p^{\min(n,k+e)/2}.$$

Hence

$$\int_{U_k\setminus U_{k+1}}K^{1/2}\,d\mu\leqslant(s-1)^2s(n+4)\,p^{\min(k,n)/2-k+\min(n,k+e)/2}.$$

For $k\leqslant n$ the exponent is at most $\min(n-k,e)/2\leqslant\min(n,e)/2$, and for $k>n$ it is at most $n-k$. Summing over $k\geqslant0$ gives the lemma for $f=0$, with $C_s=8s(s-1)^2$, since $\mu(\mathrm{prim})\geqslant3/4$ and $(n+3)(n+4)\leqslant6(n+1)^2$.

If $p\mid D_F$ the same argument applies with three changes: the first $s-2$ invariant factors of $M_{\vec a}$ have total valuation at most μ_p (Lemma 8.4(ii)), so $K(\vec a,p^n)\leqslant p^{\mu_p}p^{\min(v_1,n)+\min(v_2,n)}$, where now $v_1\leqslant v_2$ are the valuations of the last two invariant factors, with $v_1\leqslant v_p(D_{s-1}(\vec a))$ and $v_1+v_2\leqslant v_p(G_{\vec\nu_0}(\mathbf z))$; the sets $U_k=\{v_1\geqslant k\}$ have measure at most $(s-1)^2p^{C_p-k}$ by Lemma 15.3; and $K(\vec a,p^f)\leqslant p^{\mu_p+2f}$ below. Indeed, on $U_k\setminus U_{k+1}$ we have $v_1=k$ and $v_2\leqslant v_p(G_{\vec\nu_0}(\mathbf z))-k$, so the displayed bound for $K^{1/2}$ holds with the extra factor $p^{\mu_p/2}$; each of the at most $(s-1)^2$ sets D making up U_k (now of measure at most p^{C_p-k}) satisfies the bound obtained from Lemma 14.15(ii) with p^{-k} replaced by p^{C_p-k}; and the sum over k is the same. This multiplies the bound by $p^{\mu_p+C_p}$.

If $f\geqslant1$, Lemma 8.2(ii) and Lemma 8.4(i) give

$$K(\vec a,p^n)\leqslant K(\vec a,p^{n-f})K(\vec a,p^f)\leqslant p^{2f}K(\vec a,p^{n-f}).$$

The set $\mathcal A_p$ is the full preimage of the corresponding set modulo p^{n-f}, and $K(\vec a,p^{n-f})$ depends only on $\vec a \bmod p^{n-f}$, so the average is at most p^f times the average for $n-f$ in place of n and $f=0$ (and at most p^f if $f=n$). □

Proposition 14.17. *Let $\vec\nu=g\vec\nu_0$ with $\vec\nu_0\notin\mathcal N_F$ primitive, such that the line $\vec\nu_0^{\perp}$ contains no point of $\mathcal Z_2(\overline{\mathbb Q})$, let $\mathbf u\in\mathbb Z^s$ with $C_{\vec\nu_0}(\mathbf u)\neq0$, and let $\mathcal A_q(\vec\nu)$ be the set of primitive $\vec a \bmod q$ with $q\mid\vec\nu\cdot\vec a$. Then for all q*

$$\Big|\sum_{\vec a\in\mathcal A_q(\vec\nu)}S_q(\vec a,\mathbf u)\Big|\leqslant\#\mathcal A_q(\vec\nu)\,q^{(s-1)/2}\gcd(q^\sharp,g)\,B(q),\qquad\sum_{q\leqslant Y}B(q)\ll_\varepsilon(Y|\vec\nu||\mathbf u|)^\varepsilon Y.$$

Proof. This is the plane-term version of Proposition 14.14. At the primes $p \parallel q$ we use Lemma 14.11, now on the plane $W = \vec{\nu}_0^{\perp}$, where the exceptional classes are controlled by three non-zero integers attached to $\vec{\nu}_0$ and $\mathbf{u}$; at the other prime powers we use Lemma 14.16. All losses are then gcd's with a fixed non-zero integer N, or factors supported on square-full moduli, and both are harmless after summing over q. Let $E = E_{G_{\vec{\nu}_0}}$, as in Lemma 14.15(i). Since the restrictions to the line of the $(s-1)\times(s-1)$ minors of $M_{\vec{\beta}}$ have no common root over $\overline{\mathbb{Q}}$, two suitable integral linear combinations of them are binary forms without a common root, and their resultant $R'_{\vec{\nu}_0}$ is a non-zero integer; for primitive $\vec{a}$ on the line, every prime dividing $D_{s-1}(\vec{a})$ divides $R'_{\vec{\nu}_0}$. All of E, $R'_{\vec{\nu}_0}$ and $C_{\vec{\nu}_0}(\mathbf{u})$ are $\ll (|\vec{\nu}||\mathbf{u}|)^C$ in absolute value, with C depending only on s. Put $N = C_{\vec{\nu}_0}(\mathbf{u})\, E\, R'_{\vec{\nu}_0} \neq 0$.

By Lemma 14.10 the sum factorises. Write $q = q^{\flat}q^{\sharp}$ as in Lemma 14.13; the primes dividing $2D_F$ belong to $q^{\sharp}$. Let $p \parallel q$ with $p \nmid 2D_F$. If $p \mid g$, Lemma 14.11 with $W = \mathbb{F}_p^3$ bounds the local factor by $\#\mathcal{A}_p\, p^{(s-1)/2}\cdot 2C_s \gcd(p, C(\mathbf{u}))^{1/2}$, and $C(\mathbf{u}) \mid C_{\vec{\nu}_0}(\mathbf{u})$ since the content of a form divides the content of its restriction. If $p \nmid g$, we use Lemma 14.11 with $W = \vec{\nu}_0^{\perp}$ and $N(W) = p+1$. Its exceptional terms are $p^{1/2}$ if $p \mid C_{\vec{\nu}_0}(\mathbf{u})$, p if $p \mid \mathrm{cont}(G_{\vec{\nu}_0})$ (which divides E), and at most $(s-1)^2 p^{1/2}$ if the reduction of the line contains a point of corank 2, in which case $p \mid R'_{\vec{\nu}_0}$. In all cases the local factor at p is at most $\#\mathcal{A}_p\, p^{(s-1)/2}\, C'_s \gcd(p, N)$, with C'_s depending only on s.

Let $p^n \parallel q^{\sharp}$, that is, $n \geqslant 2$ or $p \mid 2D_F$, and put $p^f = \gcd(p^n, g)$. By Lemma 8.1 and Lemma 14.16, the local factor is at most

$$p^{ns/2} \sum_{\vec{a}\in\mathcal{A}_p} K(\vec{a}, p^n)^{1/2} \leqslant \#\mathcal{A}_p\, p^{n(s-1)/2}\cdot p^f \cdot C_s (n+1)^2 p^{n/2} p^{\min(e_p, n)/2},$$

and $e_p \leqslant v_p(E)$ by Lemma 14.15(i).

Hence the stated bound holds with

$$B(q) = \prod_{p\mid q^{\flat}} C'_s \gcd(p, N) \cdot \prod_{p^n \parallel q^{\sharp}} C_p (n+1)^2 p^{n/2} \gcd(p^n, E)^{1/2},$$

with $C_p = C_s$ for $p \nmid D_F$. Summing first over $q^{\flat} \leqslant Y/q^{\sharp}$, using $C_s'^{\omega(m)} \ll m^{\varepsilon}$ and $\sum_{m\leqslant Z} \prod_{p\mid m} \gcd(p, N) \leqslant Z\tau(N)$, we get $\sum_q B(q) \ll Y^{1+\varepsilon}\tau(N)\prod_{p\leqslant Y}\sigma_p$ with

$$\sigma_p = 1 + \sum_{n\geqslant 2} C_s (n+1)^2 p^{-n/2} \gcd(p^n, E)^{1/2}.$$

(with the term $n=1$ included when $p \mid 2D_F$). If $p \nmid 2D_F E$, then $\sigma_p \leqslant 1 + C/p$; the finitely many $p \mid 2D_F$ with $p \nmid E$ contribute a bounded factor. If $p^v \parallel E$, then $p^{-n/2}\gcd(p^n, E)^{1/2} = p^{(\min(n,v)-n)/2}$, so

$$\sigma_p \ll \sum_{n\leqslant v} (n+1)^2 + \sum_{n>v} (n+1)^2 p^{-(n-v)/2} \ll (v+1)^3.$$

Therefore

$$\prod_{p\leqslant Y} \sigma_p \ll (\log Y)^C \prod_{p\mid E} (v_p(E)+1)^C \ll (Y|E|)^{\varepsilon},$$

which completes the proof. □

If the line $\vec{\nu}_0^{\perp}$ contains a non-rational point z of $\mathcal{Z}_2(\overline{\mathbb{Q}})$, it also contains a conjugate $z' \neq z$, so it is the line through z and z'; there are only finitely many such lines, and we add their normals $\vec{\nu}_0$ to the finite set $\mathcal{N}_F$. It remains to allow rational members of corank 2 on the line. For such a member $\vec{c}_0$ we use Lemma 14.19, with a fixed

choice of U, and we write $t'_{\vec{c}_0}(\mathbf{u}) = \mathbf{v}'^T \operatorname{adj}(N_{\vec{c}_0})\mathbf{v}'$, a non-singular quadratic form in $\mathbf{v}'$, where $(\mathbf{v}'', \mathbf{v}') = U^T\mathbf{u}$.

Proposition 14.18. *Let $\vec{\nu} = g\vec{\nu}_0$ with $\vec{\nu}_0 \notin \mathcal{N}_F$ primitive, and let $\vec{c}_0^{(1)}, \dots, \vec{c}_0^{(k)}$ be the points of $\mathcal{Z}_2(\overline{\mathbb{Q}})$ on the line $\vec{\nu}_0^\perp$, all rational. Then the conclusion of Proposition 14.17 holds for every $\mathbf{u}$ with $C_{\vec{\nu}_0}(\mathbf{u}) \neq 0$ and $t'_{\vec{c}_0^{(j)}}(\mathbf{u}) \neq 0$ for $1 \leqslant j \leqslant k$.*

Proof. A rational corank-two point on the line would make the integer $R'_{\vec{\nu}_0}$ of the previous proof vanish. We remove these points from the resultant, and at the classes they define we use the kernel splitting of Lemma 14.19 instead of the trivial bound. The proof of Proposition 14.17 applies with two changes at the primes $p \parallel q$ with $p \nmid 2D_F g$. First, the integer $R'_{\vec{\nu}_0}$ is replaced as follows. The restrictions to the line of the $(s-1)\times(s-1)$ minors of $M_{\vec{\beta}}$ have exactly the common roots $\vec{c}_0^{(j)}$; choose two integral linear combinations A, B of them with no other common root, let a_j, b_j be their orders of vanishing at $\vec{c}_0^{(j)}$, and let ℓ_j be a primitive integral linear form on the line vanishing at $\vec{c}_0^{(j)}$. Then $A' = A/\prod_j \ell_j^{a_j}$ and $B' = B/\prod_j \ell_j^{b_j}$ have no common root, and we let $R''_{\vec{\nu}_0}$ be the product of their resultant, of their leading coefficients and of the resultants $\operatorname{Res}(\ell_i, \ell_j)$, $i \neq j$; it is a non-zero integer, and $|R''_{\vec{\nu}_0}| \ll |\vec{\nu}_0|^C$. If $p \nmid R''_{\vec{\nu}_0}$ and a primitive residue on the line modulo p has corank 2, then it is the reduction of some $\vec{c}_0^{(j)}$: otherwise ℓ_j would not vanish at it, and it would be a common root of A' and B' modulo p.

Secondly, in Lemma 14.11 with $W = \vec{\nu}_0^\perp$ the classes of corank 2 are therefore, for $p \nmid R''_{\vec{\nu}_0}$, the reductions of the $\vec{c}_0^{(j)}$, and for $p \nmid \prod_j \det N_{\vec{c}_0^{(j)}}$ we use Lemma 14.19 for them instead of Lemma 8.1: $T_p(\vec{c}_0^{(j)}, \mathbf{u}) = p^2 \mathbf{1}_{p|\mathbf{v}''} T_p^N(\mathbf{v}')$ with $N = N_{\vec{c}_0^{(j)}}$ invertible modulo p, and Lemma 14.3(i) for N gives

$$|T_p(\vec{c}_0^{(j)}, \mathbf{u})| \leqslant p^2 \cdot p^{(s-1)/2} \gcd\Big(p, t'_{\vec{c}_0^{(j)}}(\mathbf{u})\Big)^{1/2}.$$

In the notation of the proof of Lemma 14.11, where the sum is normalised by $p^{(s+1)/2}N(W)$ with $N(W) = p+1$, this class contributes at most $\gcd(p, t'_{\vec{c}_0^{(j)}}(\mathbf{u}))^{1/2}$ to β, instead of $p^{1/2}$. Hence the local factor at p is at most $\#\mathcal{A}_p\, p^{(s-1)/2} C'_s \gcd(p, N)$ with

$$N = C_{\vec{\nu}_0}(\mathbf{u})\, E\, R''_{\vec{\nu}_0} \prod_j \det N_{\vec{c}_0^{(j)}}\, t'_{\vec{c}_0^{(j)}}(\mathbf{u}) \neq 0,$$

which is $\ll (|\vec{\nu}||\mathbf{u}|)^C$. The primes $p \mid g$ are treated with $W = \mathbb{F}_p^3$ as before, and Lemma 14.16, used at the prime powers, does not depend on the points of corank 2 on the line. The rest of the proof is unchanged. □

The vectors $\mathbf{u}$ with $t'_{\vec{c}_0}(\mathbf{u}) = 0$ satisfy a non-singular quadratic equation in the $s - 2 \geqslant 5$ coordinates $\mathbf{v}'$, so by [HB02, Theorem 2] there are $\ll L^2 \cdot L^{s-4+\varepsilon} = L^{s-2+\varepsilon}$ of them with $|\mathbf{u}| \leqslant L$, a proportion $L^{-2+\varepsilon}$, as for the dual quadric in §14.10.

The factor $\gcd(q^\sharp, g)$ in Propositions 14.17 and 14.18 costs only Q^ε in the sum over the plane terms. In the proof of Proposition 10.7 the plane terms carry the weight $\gcd(q, g)^2/g$, and their total weight is $\asymp Q$ for each q. Since $B(q)$ depends on $\vec{\nu}_0$, we do not separate the sums over q and $\vec{\nu}$, but use $\gcd(q, g)^2 \gcd(q^\sharp, g)/g \leqslant g^2$: for each

$\vec{\nu} = g\vec{\nu}_0$ the sum over $q \leqslant 2Y$ of $B(q)$ is $\ll (Y|\vec{\nu}||\mathbf{u}|)^{\varepsilon} Y$, and

$$\sum_{|\vec{\nu}| \leqslant Q^{1/3}} g^2 \leqslant \sum_{g \leqslant Q^{1/3}} g^2 \, \#\{\vec{\nu}_0 : \ |\vec{\nu}_0| \leqslant Q^{1/3}/g\} \ll \sum_{g \leqslant Q^{1/3}} g^2 \Big(\frac{Q^{1/3}}{g}\Big)^3 \ll Q \log Q.$$

So the weighted sum over q and $\vec{\nu}$ is $\ll Q^{1+\varepsilon} Y^{1+\varepsilon}$, which is the same bound as for the weights $\gcd(q,g)^2/g$ alone, up to Q^{ε}.

14.7. **Kernel splitting and the numerators of the flag terms.** This subsection prepares the treatment of the flag terms. It contains three tools: the kernel splitting, which reduces Gauss sums for a singular member to Gauss sums for a non-singular form in fewer variables; an exact description of the numerators of a flag term; and a precise account of how the mass of the flag terms is distributed over the directions $\vec{c}$.

The proof of Lemma 14.3 uses only that $M_{\vec{a}}$ is a symmetric integral matrix invertible modulo p; so Lemma 14.3, and with it Lemma 14.9, hold for every such matrix N in place of $M_{\vec{a}}$, with the dual form $\mathbf{v}^T \operatorname{adj}(N)\mathbf{v}$ in place of $F^*_{\vec{a}}$ and the invariant factors of N in place of those of $M_{\vec{a}}$. We use this without comment. For a symmetric integral matrix N we write

$$S_q^N(\lambda, \mathbf{v}) = \sum_{\mathbf{y} \bmod q} \mathrm{e}_q\big(\frac{1}{2}\lambda \mathbf{y}^T N \mathbf{y} + \mathbf{v} \cdot \mathbf{y}\big),$$

so that $S_q(\lambda\vec{a}, \mathbf{u}) = S_q^{M_{\vec{a}}}(\lambda, \mathbf{u})$, and $T_q^N(\mathbf{v}) = \sideset{}{^*}\sum_{\lambda \bmod q} S_q^N(\lambda, \mathbf{v})$.

Lemma 14.19 (kernel splitting)**.** *Let $\vec{c} \in \mathbb{Z}^3$ be such that $M_{\vec{c}}$ has corank $r \geqslant 1$, let $\mathbf{h}_1, \dots, \mathbf{h}_r$ be a basis of the saturated lattice $\ker M_{\vec{c}} \cap \mathbb{Z}^s$, and let $U \in \mathrm{GL}_s(\mathbb{Z})$ have first columns $\mathbf{h}_1, \dots, \mathbf{h}_r$. Then $U^T M_{\vec{c}} U = \operatorname{diag}(0_r, N_{\vec{c}})$ with $N_{\vec{c}}$ symmetric and $\det N_{\vec{c}} \neq 0$, and for all q, λ and $\mathbf{u}$, writing $U^T\mathbf{u} = (\mathbf{v}'', \mathbf{v}')$ with $\mathbf{v}'' = (\mathbf{h}_1 \cdot \mathbf{u}, \dots, \mathbf{h}_r \cdot \mathbf{u})$,*

$$S_q(\lambda\vec{c}, \mathbf{u}) = q^r \, \mathbf{1}_{q|\mathbf{v}''} \, S_q^{N_{\vec{c}}}(\lambda, \mathbf{v}'), \qquad T_q(\vec{c}, \mathbf{u}) = q^r \, \mathbf{1}_{q|\mathbf{v}''} \, T_q^{N_{\vec{c}}}(\mathbf{v}').$$

The same holds over $\mathbb{F}_p$ for a residue $\vec{a}$ whose matrix has corank r modulo p, with $U \in \mathrm{GL}_s(\mathbb{F}_p)$.

Proof. In coordinates adapted to the kernel, the quadratic form does not involve the first r variables at all, so the sum over them is a pure linear exponential sum, which detects the congruence $q \mid \mathbf{v}''$. The first columns of U span $\ker M_{\vec{c}}$, so $U^T M_{\vec{c}} U$ has vanishing first r rows and columns, and its lower block is non-singular since $\operatorname{rank} M_{\vec{c}} = s - r$. Substituting $\mathbf{b} = U\mathbf{x}$ in the definition of S_q, the sum over the first r coordinates of $\mathbf{x}$ is

$$\prod_{i \leqslant r} \sum_{x_i \bmod q} \mathrm{e}_q(x_i \mathbf{h}_i \cdot \mathbf{u}) = q^r \mathbf{1}_{q|\mathbf{v}''},$$

and the sum over the others is $S_q^{N_{\vec{c}}}(\lambda, \mathbf{v}')$. □

We next describe exactly the numerators of a flag term; the description of Lemma 10.6, which ignored the condition $\vec{\nu} \in \Lambda^{\perp}(\vec{a}, q)$, was enough for upper bounds but not for the sums over the numerators. Let $(\vec{\nu}, \vec{\rho})$ be a flag with $\vec{\nu} = g\vec{\nu}_0$, $\vec{\rho} = G\vec{c}$, $\vec{\nu}_0, \vec{c}$ primitive, and let $q \in \mathbb{N}$ and $d = \gcd(q, G)$. Write $q = q_c q_d$ with $q_d = \prod_{p|d} p^{v_p(q)}$, so that $\gcd(q_c, G) = 1$ and $q_d \mid G^{\infty}$. For $p \mid d$ put $n_p = v_p(q)$, $\delta_p = v_p(d) = \min(n_p, v_p(G))$ and $f_p = \min(\delta_p, v_p(g))$.

Lemma 14.20. *The set*

$$\mathcal{A}^{\flat} = \{\vec{a} \bmod q : \ \gcd(\vec{a}, q) = 1, \ \vec{\rho} \in \Lambda(\vec{a}, q), \ \vec{\nu} \in \Lambda^{\perp}(\vec{a}, q)\}$$

of numerators of the flag term is the product, under the Chinese remainder theorem, of the class $\mathcal{A}_c = \{\lambda\vec{c} \bmod q_c\}$ *and of the sets*

$$\mathcal{A}_p = \{\lambda\vec{c} + p^{n_p-\delta_p}\vec{b} :\ p \nmid \lambda,\ \vec{b} \bmod p^{\delta_p},\ p^{\delta_p-f_p} \mid \vec{\nu}_0\cdot\vec{b}\} \qquad (p \mid d),$$

each stable under multiplication by units. If $\delta_p = n_p$*, then* $\mathcal{A}_p$ *is the set of primitive residues modulo* p^{n_p} *on the plane* $p^{n_p-f_p} \mid \vec{\nu}_0\cdot\vec{a}$*; this is the set of all primitive residues if moreover* $f_p = n_p$*. Consequently*

$$\sum_{\vec{a}\in\mathcal{A}^\flat} S_q(\vec{a},\mathbf{u}) = T_{q_c}(\vec{c},\mathbf{u}) \prod_{p\mid d} \sum_{\vec{a}\in\mathcal{A}_p} S_{p^{n_p}}(\vec{a},\mathbf{u}), \tag{14.5}$$

and $\#\mathcal{A}^\flat \leqslant q\, d \gcd(d,g)$*.*

Proof. We combine the description of Lemma 5.7(ii) with the second condition $\vec{\nu} \in \Lambda^\perp(\vec{a},q)$, and read off the local structure prime by prime. By Lemma 5.7(ii), the condition $\vec{\rho} \in \Lambda(\vec{a},q)$ means $\vec{a} = \lambda\vec{c} + q_1\vec{b}$ with $q_1 = q/d$, λ a unit modulo q_1 and $\vec{b}$ mod d. Since $\vec{\nu}\cdot\vec{c} = 0$, we have $\vec{\nu}\cdot\vec{a} = q_1\vec{\nu}\cdot\vec{b}$, so $\vec{\nu} \in \Lambda^\perp(\vec{a},q)$ means $d \mid g\vec{\nu}_0\cdot\vec{b}$, that is, $p^{\delta_p-f_p} \mid \vec{\nu}_0\cdot\vec{b}$ for every $p \mid d$. For $p \nmid d$ we have $p^{n_p} \mid q_1$, so the residue modulo p^{n_p} is $\lambda\vec{c}$; for $p \mid d$ we have $v_p(q_1) = n_p - \delta_p$, and the residue modulo p^{n_p} runs over $\mathcal{A}_p$. The residues of λ and $\vec{b}$ at the different primes are independent, which gives the product structure; the case $\delta_p = n_p$ follows from $\vec{\nu}_0\cdot\vec{c} = 0$. Formula (14.5) follows from Lemma 14.10 and Lemma 14.2. Finally, the pairs $(\lambda, \vec{b})$ number at most $q_1\cdot d^3\gcd(d,g)/d = qd\gcd(d,g)$. □

The bound $\#\mathcal{A}^\flat \leqslant qd\gcd(d,g)$ is smaller than the bound qd^2 used in Proposition 10.7 by the factor $d/\gcd(d,g)$. This changes the distribution of the mass of the flag terms among the directions $\vec{c}$, though not its total. For primitive $\vec{c}$, $1 \leqslant Y \leqslant Q$ and $U = Q^{1/3}$ put

$$\Omega_{\vec{c}}(Y) = \sum_{\substack{\vec{\nu}\perp\vec{c}\\ 0<|\vec{\nu}|\leqslant U}} \sum_{G\leqslant Z_{\vec{c}}} \sum_{Y<q\leqslant 2Y} \#\mathcal{A}^\flat(q;\vec{\nu},G\vec{c}), \qquad Z_{\vec{c}} = \frac{Q^{2/3}}{|\vec{c}|}.$$

Lemma 14.21. *For primitive* $\vec{c}$ *with* $|\vec{c}| \leqslant Q^{2/3}$ *we have*

$$\Omega_{\vec{c}}(Y) \ll Y^2 Z_{\vec{c}} (\log Q)^3 \Big(\frac{U}{\lambda_1(\vec{c})} + \frac{U^2}{|\vec{c}|}\Big),$$

and for $1 \leqslant H \leqslant Q^{2/3}$*,*

$$\sum_{|\vec{c}|\asymp H} \Omega_{\vec{c}}(Y) \ll Y^2 Q^{2/3} (\log Q)^3 (U^2H + UH^{3/2}).$$

In particular $\sum_{\vec{c}} \Omega_{\vec{c}}(Y) \ll Y^2Q^2(\log Q)^4$*.*

Proof. We sum the bound of Lemma 14.20 first over q and G, for a fixed $\vec{\nu}$, and then count the vectors $\vec{\nu}$ in the plane $\vec{c}^{\perp}$ with Lemma 10.5. Fix $\vec{\nu} = g\vec{\nu}_0 \perp \vec{c}$. The number of $q \in (Y, 2Y]$ with $\gcd(q,G) = d$ is at most $2Y/d$, so by Lemma 14.20 the sum over q and G is at most

$$4Y^2 \sum_{G\leqslant Z_{\vec{c}}} \sum_{\substack{d\mid G\\ d\leqslant 2Y}} \gcd(d,g) \leqslant 4Y^2 Z_{\vec{c}} \sum_{d\leqslant 2Y} \frac{\gcd(d,g)}{d} \leqslant 4Y^2 Z_{\vec{c}}\tau(g)(1+\log 2Y).$$

Summing over $g \leqslant U$ and over the primitive $\vec{\nu}_0 \perp \vec{c}$ with $|\vec{\nu}_0| \leqslant U/g$, of which there are $\ll U/(g\lambda_1(\vec{c})) + U^2/(g^2|\vec{c}|)$ by Lemma 10.5(i), gives the first bound, since

$\sum_g \tau(g)/g \ll (\log U)^2$. The second follows from $\#\{|\vec{c}| \asymp H\} \ll H^3$ and from $\sum_{|\vec{c}|\asymp H} \lambda_1(\vec{c})^{-1} \ll H^{5/2}$ (Lemma 10.5(ii)). Summing over dyadic $H \leqslant Q^{2/3}$ gives the last bound. □

Thus the mass of the flag terms is carried mainly by the directions $\vec{c}$ of size about $Q^{2/3}$, and a fixed direction $\vec{c}$ carries only a proportion $\ll Q^{-2/3+\varepsilon}$ of it; the directions on a fixed rational line $\vec{\nu}_0^{\perp}$, for which $\lambda_1(\vec{c}) \leqslant |\vec{\nu}_0|$, carry together a proportion $\ll Q^{-1/3+\varepsilon}$. In the same way, the flags with $d = \gcd(q, G) > 1$ are not rare, but the weight of a given d decreases like $\gcd(d, g)/d$.

The next lemma makes precise the remark that short directions carry little of the mass. It explains why large values of $\gcd(q, G)$ cause no difficulty: such values occur only when $G \geqslant \gcd(q, G)$, hence only for directions with $|\vec{c}| \leqslant Q^{2/3}/\gcd(q, G)$. For a primitive $\vec{c}$ and $1 \leqslant Y \leqslant Q$ we put

$$Z'_{\vec{c}} = \min(2Y, Z_{\vec{c}}) = \min\Big(2Y, \frac{Q^{2/3}}{|\vec{c}|}\Big),$$

so that $\gcd(q, G) \leqslant Z'_{\vec{c}}$ for every flag term with $\vec{\rho} = G\vec{c}$ and $Y < q \leqslant 2Y$.

Lemma 14.22. *Let $1 \leqslant Y \leqslant Q$, and let $\Omega^+_{\vec{c}}$ be as in (14.1).*

(i) For $1 \leqslant H \leqslant Q^{2/3}$ we have $\sum_{|\vec{c}|\asymp H} \Omega^+_{\vec{c}} \ll Y^2Q^{2/3}(U^2H + UH^{3/2})$, the sum being over primitive $\vec{c}$.

(ii) For $0 \leqslant a \leqslant 1$,

$$\sum_{\substack{\vec{c}\ \text{primitive}\\ |\vec{c}|\leqslant Q^{2/3}}} \Omega^+_{\vec{c}}\,(Z'_{\vec{c}})^a \ll Y^2Q^2 \log Q,$$

$$\sum_{\substack{\vec{c}\ \text{primitive}\\ |\vec{c}|\leqslant Q^{2/3}}} \Omega^+_{\vec{c}}\,(Z'_{\vec{c}})^{3/2} \ll Y^2Q^2 \min(Y, Q^{2/3})^{1/2} \log Q.$$

Proof. Part (i) is a count over dyadic shells. In part (ii) the factor $(Z'_{\vec{c}})^a$ is large only for short directions $\vec{c}$, which carry little mass by (i); the two effects balance exactly when $a = 1$.

(i) For $|\vec{c}| \asymp H$ we have $\Omega^+_{\vec{c}} \asymp Y^2Q^{2/3}H^{-1}(U/\lambda_1(\vec{c}) + U^2/H)$. There are $\ll H^3$ primitive $\vec{c}$ with $|\vec{c}| \asymp H$, and $\sum_{|\vec{c}|\leqslant 2H} \lambda_1(\vec{c})^{-1} \ll H^{5/2}$ by Lemma 10.5(ii).

(ii) Split the sum into dyadic ranges $|\vec{c}| \asymp H$, in which $Z'_{\vec{c}} \ll Q^{2/3}/H$. By (i) the range contributes

$$\ll Y^2Q^{2/3}(U^2H + UH^{3/2})\Big(\frac{Q^{2/3}}{H}\Big)^a = Y^2Q^{2/3+2a/3}\Big(U^2H^{1-a} + UH^{3/2-a}\Big).$$

For $a \leqslant 1$ the exponents of H are non-negative, so each range contributes at most the value at $H = Q^{2/3}$, which is $Y^2Q^{2/3}(U^2Q^{2/3} + UQ) \ll Y^2Q^2$. For $a = 3/2$ the range contributes $Y^2Q^{5/3}(U^2H^{-1/2} + U) \ll Y^2Q^{7/3}$, which is the second bound when $Y \geqslant Q^{2/3}$. When $Y < Q^{2/3}$ we use instead $(Z'_{\vec{c}})^{3/2} \leqslant (2Y)^{1/2}Z'_{\vec{c}}$ and the first bound with $a = 1$. □

Thus a factor $(Z'_{\vec{c}})^a$ with $a \leqslant 1$ in the mass of the flag terms costs only $\log Q$. A factor $\gcd(q, G)^a$ can be large for a single short direction $\vec{c}$, but on average over all flags it costs nothing.

14.8. **The generic flag terms.** We now treat the flag terms $(\vec\nu,\vec\rho)$ with $\vec\rho = G\vec c$ and $\Delta(\vec c)\neq 0$. We use the notation of Lemma 14.20: $\vec\nu = g\vec\nu_0$, $d=\gcd(q,G)$, and $\mathcal{A}^\flat$ is the set of numerators. For such a flag put $N_{\vec c} = 2D_F\Delta(\vec c)$. The principle described at the beginning of this section is expressed by the following lemma: the sum over the numerators is split into a part q', at which we use the sum over λ, and a part k, at which we use only the exact size of the Gauss sums.

Lemma 14.23. *Let $(\vec\nu,\vec\rho)$ be a flag with $\vec\rho = G\vec c$, $\vec c$ primitive and $\Delta(\vec c)\neq 0$, let $q\in\mathbb{N}$ and $d=\gcd(q,G)$. Let k be the largest divisor of q composed of primes dividing $dN_{\vec c}$, and put $q'=q/k$. Let $\mathcal{A}_k$ be the set of residues modulo k of the elements of $\mathcal{A}^\flat$. Then $\#\mathcal{A}^\flat = \varphi(q')\,\#\mathcal{A}_k$, and for every $\mathbf{u}\in\mathbb{Z}^s$*

$$\Big|\sum_{\vec a\in\mathcal{A}^\flat} S_q(\vec a,\mathbf{u})\Big| \leqslant \frac{\#\mathcal{A}^\flat}{\#\mathcal{A}_k}\,\frac{q'}{\varphi(q')}\,q'^{(s-1)/2}\gcd\Big(q',F_{\vec c}^*(\mathbf{u})\Big)^{1/2}\sum_{\vec a\in\mathcal{A}_k}|S_k(\vec a,\mathbf{u})|, \tag{14.6}$$

with the convention $\gcd(q',0)=q'$. Moreover $|S_k(\vec a,\mathbf{u})| = k^{s/2}K(\vec a,k)^{1/2}\mathbf{1}_{\mathbf{u}\in\mathcal{M}(\vec a,k)}$ for $\vec a\in\mathcal{A}_k$, and

$$K(\vec a,k)\leqslant c_F d^2 K(\vec c,k/d),\qquad K(\vec c,k/d)\leqslant \min\Big(c_F(k/d)^2,\ D_{s-1}(\vec c)\gcd(\Delta(\vec c),k/d)\Big). \tag{14.7}$$

Proof. Modulo q' the numerators form a single class $\{\lambda\vec c\}$, so the sum over them is exactly $T_{q'}(\vec c,\mathbf{u})$, which Lemma 14.3 evaluates; the part modulo k is kept as it is. By Lemma 14.20, $\mathcal{A}^\flat$ is the product, under the Chinese remainder theorem, of the class $\{\lambda\vec c \bmod q_c\}$ and of the sets $\mathcal{A}_p$ for $p\mid d$. The modulus q' divides q_c and is coprime to $kN_{\vec c}$, so the residues of $\mathcal{A}^\flat$ modulo q' form the class $\{\lambda\vec c \bmod q'\}$, and $\mathcal{A}^\flat$ is the product of this class and $\mathcal{A}_k$; this gives $\#\mathcal{A}^\flat=\varphi(q')\#\mathcal{A}_k$. Writing $S_q(\vec a,\mathbf{u}) = S_{q'}(k\vec a,\mathbf{u})S_k(q'\vec a,\mathbf{u})$ as in the proof of Lemma 14.2, and using that $\mathcal{A}_k$ is stable under multiplication by units, we get

$$\sum_{\vec a\in\mathcal{A}^\flat} S_q(\vec a,\mathbf{u}) = T_{q'}(\vec c,\mathbf{u})\sum_{\vec a\in\mathcal{A}_k} S_k(\vec a,\mathbf{u}).$$

Every prime dividing q' is odd and does not divide $\Delta(\vec c)$, so Lemmas 14.2 and 14.3(i) give $|T_{q'}(\vec c,\mathbf{u})|\leqslant q'^{(s+1)/2}\gcd(q',F_{\vec c}^*(\mathbf{u}))^{1/2}$; here we use that $F^*_{m\vec c}=m^{s-1}F^*_{\vec c}$, so that the twists in Lemma 14.2 do not change the gcd. This proves (14.6). The formula for $|S_k(\vec a,\mathbf{u})|$ is Lemma 8.1.

For (14.7), recall from Lemma 5.7(ii) that every $\vec a\in\mathcal{A}^\flat$ satisfies $\vec a\equiv\lambda\vec c \pmod{q/d}$ for some λ coprime to q. Since k/d divides q/d, Lemma 8.2(ii),(iii) gives $K(\vec a,k)\leqslant K(\vec a,k/d)K(\vec a,d) = K(\vec c,k/d)K(\vec a,d)$. By Lemma 8.4 we have $K(\vec a,p^n)\leqslant p^{\mu_p}p^{2n}$ for every prime power, with $\mu_p=0$ for good p, so $K(\vec a,m)\leqslant c_F m^2$ for all m and all $\vec a$ coprime to m. This gives the first bound and the first half of the second; the other half is Lemma 8.2(iv). □

At the primes dividing k we have given up all cancellation. Compared with the trivial bound $\#\mathcal{A}^\flat q^{s/2}$ per dual vector, the bound (14.6) saves $q'^{1/2}=(q/k)^{1/2}$, up to the gcd. We now show that this is enough, because k is usually small.

Proposition 14.24. *Let $1\leqslant Y\leqslant Q$ and $L\geqslant 1$, let $\vec c$ be primitive with $\Delta(\vec c)\neq 0$ and $|\vec c|\leqslant Q^{2/3}$, and let $Z'_{\vec c}$ and $\Omega^+_{\vec c}$ be as in Lemma 14.22. Write*

$$\Sigma^{(1)}_{\vec c} = \sum_{\substack{\vec\nu\perp\vec c\\ 0<|\vec\nu|\leqslant U}}\ \sum_{G\leqslant Z_{\vec c}}\ \sum_{Y<q\leqslant 2Y}\ \sum_{\substack{0<|\mathbf{u}|\leqslant L\\ F^*_{\vec c}(\mathbf{u})\neq 0}}\Big|\sum_{\vec a\in\mathcal{A}^\flat} S_q(\vec a,\mathbf{u})\Big|,$$

and let $\Sigma_{\vec c}^{(0)}$ *be the same sum over the* $\mathbf{u}$ *with* $F_{\vec c}^*(\mathbf{u}) = 0$. *Then*

$$\Sigma_{\vec c}^{(1)} \ll_\varepsilon Q^\varepsilon\, \Omega_{\vec c}^+\, Y^{s/2} L^s \Big(\Big(\frac{Z'_{\vec c}}{Y} \Big)^{1/2} + \frac{Z'_{\vec c}}{L} \Big), \tag{14.8}$$

$$\Sigma_{\vec c}^{(1)} \ll_\varepsilon Q^\varepsilon\, \Omega_{\vec c}^+\, Y^{s/2} L^s \Big(\Big(\frac{Z'_{\vec c}}{Y} \Big)^{1/2} + \frac{D_{s-1}(\vec c)^{1/2} Z'^{3/2}_{\vec c}}{Y^{1/2} L} \Big), \tag{14.9}$$

$$\Sigma_{\vec c}^{(0)} \ll_\varepsilon Q^\varepsilon\, \Omega_{\vec c}^+\, Y^{s/2} L^s\, \frac{Z'_{\vec c}}{L^2}. \tag{14.10}$$

Proof. The proof is an exercise in summing Lemma 14.23 in the right order. We fix the flag and the part k of the modulus at which no cancellation is used. The sum over the remaining part q' then saves $q'^{1/2}$, up to a gcd which is 1 on average (Lemma 14.7); the sum over $\mathbf{u}$ at the primes of k is handled with the coset count, as in Lemma 10.2; and since $k = dk''$ with k'' composed of primes from a fixed integer, there are only Q^ε choices of k''. What remains is a sum over $d = \gcd(q, G)$, which costs at most a power of $Z'_{\vec c}$, and finally the count of the vectors $\vec\nu$.

Fix $\vec\nu = g\vec\nu_0 \perp \vec c$, $G \leqslant Z_{\vec c}$ and a divisor d of G with $d \leqslant 2Y$, and consider the $q \in (Y, 2Y]$ with $\gcd(q, G) = d$. For such q the integer k of Lemma 14.23 has the form $k = dk''$ with $k'' \mid (dN_{\vec c})^\infty$, and $q = q'k$ with $\gcd(q', dN_{\vec c}) = 1$. For fixed k the set $\mathcal{A}_k$ depends only on the flag and on k, and distinct q correspond to distinct $q' \leqslant 2Y/k$. We have $\#\mathcal{A}^\flat \leqslant 2Y d \gcd(d, g)$ by Lemma 14.20.

Fix also k and a vector $\mathbf{u}$ with $F_{\vec c}^*(\mathbf{u}) \neq 0$. Since $q'/\varphi(q') \ll \log\log Q$ and $|F_{\vec c}^*(\mathbf{u})| \ll |\vec c|^{s-1} L^2 \ll Q^C$, Lemma 14.7(i) gives

$$\sum_{q' \leqslant 2Y/k} \frac{q'}{\varphi(q')} \gcd\big(q', F_{\vec c}^*(\mathbf{u})\big)^{1/2} \ll Q^\varepsilon \frac{Y}{k}.$$

Next we sum over $\mathbf{u}$. For each $\vec a \in \mathcal{A}_k$ the vectors $\mathbf{u}$ with $\mathbf{u} \bmod k \in \mathcal{M}(\vec a, k)$ form an empty set or a coset of a lattice of index $K(\vec a, k)$, so by Lemma 10.1

$$\sum_{|\mathbf{u}| \leqslant L} |S_k(\vec a, \mathbf{u})| \ll k^{s/2} L^s \Big(K(\vec a, k)^{-1/2} + \frac{K(\vec a, k)^{1/2}}{L} \Big) \leqslant k^{s/2} L^s \Big(1 + \frac{c_F^{1/2} d\, K(\vec c, k'')^{1/2}}{L} \Big),$$

by (14.7). Inserting these bounds in (14.6), we find that the contribution of the q with this value of k to $\Sigma_{\vec c}^{(1)}$ is

$$\begin{aligned} &\ll Q^\varepsilon\, Y d \gcd(d, g) \Big(\frac{Y}{k} \Big)^{(s+1)/2} k^{s/2} L^s \Big(1 + \frac{d\, K(\vec c, k'')^{1/2}}{L} \Big) \\ &= Q^\varepsilon\, Y^{s/2} L^s\, Y^{3/2} d \gcd(d, g)\, k^{-1/2} \Big(1 + \frac{d\, K(\vec c, k'')^{1/2}}{L} \Big). \end{aligned}$$

Now we sum over $k = dk''$. By Lemma 14.7(ii) there are $\ll Q^\varepsilon$ possible values of k''. With the bound $K(\vec c, k'') \leqslant c_F k''^2$ we have

$$\sum_{k''} (dk'')^{-1/2} \Big(1 + \frac{dK(\vec c, k'')^{1/2}}{L} \Big) \ll Q^\varepsilon d^{-1/2} \Big(1 + \frac{d}{L} \Big(\frac{2Y}{d} \Big)^{1/2} \Big),$$

since $k'' \leqslant 2Y/d$. With the bound $K(\vec c, k'') \leqslant D_{s-1}(\vec c) \gcd(\Delta(\vec c), k'')$ we have instead

$$\sum_{k''} k''^{-1/2} K(\vec c, k'')^{1/2} \leqslant D_{s-1}(\vec c)^{1/2} \prod_{p \mid dN_{\vec c}} \sum_{j \geqslant 0} p^{(\min(j, v_p(\Delta(\vec c))) - j)/2} \ll Q^\varepsilon D_{s-1}(\vec c)^{1/2},$$

as in the proof of Lemma 14.9. So the q with $\gcd(q,G)=d$ contribute

$$\ll Q^{\varepsilon}\,Y^{s/2}L^{s}\,Y^{2}\gcd(d,g)\Big(\Big(\frac{d}{Y}\Big)^{1/2}+\min\Big(\frac{d}{L},\ \frac{D_{s-1}(\vec{c})^{1/2}d^{3/2}}{Y^{1/2}L}\Big)\Big). \tag{14.11}$$

We sum (14.11) over $G\leqslant Z_{\vec{c}}$ and $d\mid G$. Each d occurs for at most $Z_{\vec{c}}/d$ values of G, and only $d\leqslant Z'_{\vec{c}}$ occur. For $\beta\geqslant 0$ and $D\geqslant 1$,

$$\sum_{d\leqslant D}\gcd(d,g)\,d^{\beta-1}\leqslant\sum_{e\mid g}e\sum_{\substack{d\leqslant D\\ e\mid d}}d^{\beta-1}\ll_{\beta}\tau(g)\,D^{\beta}(1+\log D),$$

so the sum over G and d of (14.11) is

$$\ll Q^{\varepsilon}\,Y^{s/2}L^{s}\,Y^{2}Z_{\vec{c}}\tau(g)\Big(\Big(\frac{Z'_{\vec{c}}}{Y}\Big)^{1/2}+\min\Big(\frac{Z'_{\vec{c}}}{L},\ \frac{D_{s-1}(\vec{c})^{1/2}Z_{\vec{c}}'^{3/2}}{Y^{1/2}L}\Big)\Big).$$

Finally we sum over $\vec{\nu}=g\vec{\nu}_0$. For each $g\leqslant U$ the number of primitive $\vec{\nu}_0\perp\vec{c}$ with $|\vec{\nu}_0|\leqslant U/g$ is $\ll U/(g\lambda_1(\vec{c}))+U^2/(g^2|\vec{c}|)$ by Lemma 10.5(i), and

$$\sum_{g}\tau(g)(U/(g\lambda_1(\vec{c}))+U^2/(g^2|\vec{c}|))\ll(\log Q)^2(U/\lambda_1(\vec{c})+U^2/|\vec{c}|).$$

This proves (14.8) and (14.9).

For (14.10) we argue in the same way, with two changes. When $F^*_{\vec{c}}(\mathbf{u})=0$, Lemma 14.3(i) gives only $|T_{q'}(\vec{c},\mathbf{u})|\leqslant q'^{(s+2)/2}$, a loss of $q'^{1/2}\leqslant(2Y/k)^{1/2}$; and we do not use the cosets $\mathcal{M}(\vec{a},k)$, but count the $\mathbf{u}$ with $|\mathbf{u}|\leqslant L$ and $F^*_{\vec{c}}(\mathbf{u})=0$ by (14.12), which gives $\ll L^{s-2+\varepsilon}$. Using $K(\vec{a},k)^{1/2}\leqslant c_F dk''$ we find that the q with $\gcd(q,G)=d$ and a given k contribute

$$\ll Q^{\varepsilon}\,Y d\gcd(d,g)\Big(\frac{Y}{k}\Big)^{s/2+1}k^{s/2}L^{s-2+\varepsilon}\,dk''=Q^{\varepsilon}\,Y^{s/2}L^{s}\,Y^{2}\gcd(d,g)\frac{d\,L^{\varepsilon}}{L^2}.$$

Summing over the $\ll Q^{\varepsilon}$ values of k'', over G, d and $\vec{\nu}$ as before gives (14.10). □

Compared with the trivial bound $\Omega^{+}_{\vec{c}}Y^{s/2}L^{s}$, the proposition saves $(Y/Z'_{\vec{c}})^{1/2}$, up to the loss $Z'_{\vec{c}}/L$. For a single short direction $\vec{c}$ this could be poor, since $Z'_{\vec{c}}$ may be as large as $\min(2Y,Q^{2/3})$. But by Lemma 14.22 the factors $(Z'_{\vec{c}})^{1/2}$ and $Z'_{\vec{c}}$ cost only $\log Q$ on average over $\vec{c}$, and so the generic flag terms save $Y^{1/2}(1+Y^{1/2}/L)^{-1}$ on average.

14.9. **Dual vectors with vanishing content.** Propositions 14.14 and 14.17 exclude the $\mathbf{u}$ with $C(\mathbf{u})=0$ and with $C_{\vec{\nu}_0}(\mathbf{u})=0$. These vectors are rare. For such $\mathbf{u}$ the dual form $F^*_{\vec{a}}(\mathbf{u})$ vanishes identically in $\vec{a}$ (or along a line), so no cancellation in the sum over the numerators is possible, and we must show that there are few of them in a box. To see this we need a fact about pencils of quadrics. For a line $\ell\subset\mathbb{P}^2$ let

$$V_\ell=\{\mathbf{u}\in\mathbb{C}^s:\ F^*_{\vec{c}}(\mathbf{u})=0\text{ for all }[\vec{c}]\in\ell\},$$

Geometrically, $\mathbf{u}\in V_\ell$ means that the hyperplane $\mathbf{u}\cdot\mathbf{x}=0$ is tangent to every quadric of the pencil ℓ.

Lemma 14.25. *Let ℓ be a line in $\mathbb{P}^2$ which is not contained in the curve $\Delta=0$, and let $k(\ell)$ be the number of distinct points of $\ell\cap\{\Delta=0\}$ over $\mathbb{C}$. Then $\dim V_\ell\leqslant s-k(\ell)$.*

Proof. Each point of ℓ where Δ vanishes gives, in the expansion of F^* along the pencil, a non-zero quadratic form in a separate group of variables, which must vanish on V_ℓ. Since these groups of variables are disjoint, the $k(\ell)$ conditions are independent. To make this precise we put the pencil in canonical form. Choose a basis $\vec{\beta}_1,\vec{\beta}_2$ of ℓ such

that $M_{\vec{\beta}_1}$ is invertible, and put $A = -M_{\vec{\beta}_1}$, $B = M_{\vec{\beta}_2}$. Since ℓ is not contained in the discriminant curve, the pencil $B - tA$ is regular. By the canonical form for regular pencils of complex symmetric matrices (see [Tho91]), there is an invertible matrix S such that $S^T(B - tA)S$ is block diagonal, with blocks $E_n(J_n(\beta) - t)$, where $J_n(\beta)$ is a Jordan block and E_n is the $n \times n$ reversal matrix. The eigenvalues β are the $k = k(\ell)$ distinct roots of $\det(B - tA)$. Since $\operatorname{adj}(S^T C S) = \det(S)^2 S^{-1} \operatorname{adj}(C) S^{-T}$, replacing $\mathbf{u}$ by $S^{-T}\mathbf{u}$ changes F^* only by the non-zero factor $\det(S)^2$, so we may assume the pencil is in this form, and write $\mathbf{u} = (\mathbf{u}_j)$ according to the blocks $\mathcal{B}_j$, of sizes n_j and eigenvalues β_j. Then

$$F^*_{B-tA}(\mathbf{u}) = \sum_j \mathbf{u}_j^T \operatorname{adj}(\mathcal{B}_j(t))\mathbf{u}_j \prod_{i \neq j} \det \mathcal{B}_i(t), \qquad \det \mathcal{B}_i(t) = \pm(\beta_i - t)^{n_i}.$$

Fix an eigenvalue β, let m be the largest size of a block with eigenvalue β, and let μ be the algebraic multiplicity of β. A single Jordan block $\mathcal{B}_j(\beta)$ has corank 1, so $\operatorname{adj}(\mathcal{B}_j(\beta)) = \kappa_j \mathbf{v}_j \mathbf{v}_j^T$ with $\kappa_j \neq 0$ and $\mathbf{v}_j$ spanning its kernel. Consider a block j with $\beta_j = \beta$. The product $\prod_{i \neq j} \det \mathcal{B}_i(t)$ vanishes at $t = \beta$ to order exactly $\mu - n_j$, and $\mathbf{u}_j^T \operatorname{adj}(\mathcal{B}_j(t))\mathbf{u}_j$ is a polynomial whose value at $t = \beta$ is $\kappa_j(\mathbf{v}_j \cdot \mathbf{u}_j)^2$. The term of index j therefore vanishes to order at least $\mu - n_j$, and exactly $\mu - n_j$ at $t = \beta$ when $\mathbf{v}_j \cdot \mathbf{u}_j \neq 0$, and the terms with $\beta_j \neq \beta$ vanish to order at least μ. So only the blocks with $\beta_j = \beta$ and $n_j = m$ contribute to the coefficient of $(\beta - t)^{\mu - m}$. Hence the coefficient of $(\beta - t)^{\mu-m}$ in the Taylor expansion of $F^*_{B-tA}(\mathbf{u})$ at $t = \beta$ is a non-zero multiple of

$$q_\beta(\mathbf{u}) = \sum_{j:\ \beta_j = \beta,\ n_j = m} \pm\kappa_j \prod_{i \neq j} c_i \, (\mathbf{v}_j \cdot \mathbf{u}_j)^2,$$

where the $c_i \neq 0$ are the values at $t = \beta$ of $\det \mathcal{B}_i(t)(\beta - t)^{-n_i}$, or $\det \mathcal{B}_i(\beta)$ if $\beta_i \neq \beta$. This is a non-zero quadratic form in the variables $\mathbf{u}_j$ belonging to the blocks with eigenvalue β. If $\mathbf{u} \in V_\ell$, then $q_\beta(\mathbf{u}) = 0$ for every β. The forms q_β for different β involve disjoint sets of variables, so their common zero set has codimension k. □

We put $E_0 = \{\mathbf{u} \in \mathbb{C}^s : \ F^*_{\vec{c}}(\mathbf{u}) = 0 \text{ for all } \vec{c}\}$. Thus $E_0 \subseteq V_\ell$ for every line ℓ, and $C(\mathbf{u}) = 0$ exactly when $\mathbf{u} \in E_0$.

Lemma 14.26. *Let $X \subseteq \mathbb{C}^s$ be a Zariski closed set of dimension at most D defined by polynomials of degree at most d. Then $\#\{\mathbf{u} \in \mathbb{Z}^s \cap X : \ |\mathbf{u}|_\infty \leqslant L\} \ll_{s,d} L^D$ for $L \geqslant 1$.*

Proof. This is the trivial bound for integer points on a variety, by slicing with coordinate hyperplanes. By Bézout's theorem X has $\ll_{s,d} 1$ irreducible components, each of degree $\ll_{s,d} 1$, so we may assume X irreducible of dimension D and degree $\ll_{s,d} 1$. If $D = 0$, X is a point. Otherwise some coordinate u_i is not constant on X, and for each integer c with $|c| \leqslant L$ the set $X \cap \{u_i = c\}$ has dimension at most $D - 1$ and is defined by polynomials of bounded degree; by induction on D it contains $\ll L^{D-1}$ of the points in question, and summing over the $2L + 1$ values of c gives the claim. □

Lemma 14.27. *Let $L \geqslant 1$.*

(i) The number of $\mathbf{u} \in \mathbb{Z}^s$ with $|\mathbf{u}| \leqslant L$ and $C(\mathbf{u}) = 0$ is $\ll L^{s-3}$.
(ii) For primitive $\vec{\nu}_0 \notin \mathcal{N}_F$, the number of $\mathbf{u} \in \mathbb{Z}^s$ with $|\mathbf{u}| \leqslant L$ and $C_{\vec{\nu}_0}(\mathbf{u}) = 0$ is $\ll L^{s/2}$, uniformly in $\vec{\nu}_0$.
(iii) For each of the finitely many $\vec{\nu}_0 \in \mathcal{N}_F$, this number is $\ll L^{s-1}$.

Proof. $C(\mathbf{u}) = 0$ means $\mathbf{u} \in E_0$, and $C_{\vec{\nu}_0}(\mathbf{u}) = 0$ means that $F^*_{\vec{c}}(\mathbf{u}) = 0$ for all $\vec{c}$ on the line $\ell = \vec{\nu}_0^{\perp}$, that is $\mathbf{u} \in V_\ell$. These sets are defined by the vanishing of the coefficients of the forms $\vec{c} \mapsto F^*_{\vec{c}}(\mathbf{u})$, which are at most $\binom{s+1}{2}$ quadratic polynomials in $\mathbf{u}$. For $\ell \notin \mathcal{N}_F$ every point of $\ell \cap \{\Delta = 0\}$ has multiplicity at most 2, so $k(\ell) \geqslant s/2$, and Lemma 14.25 gives $\dim V_\ell \leqslant s/2$; also $E_0 \subseteq V_\ell$, so $\dim E_0 \leqslant s/2 \leqslant s-3$. For $\ell \in \mathcal{N}_F$ the line contains members of corank at most 1, whose dual forms $F^*_{\vec{c}}$ are non-zero, so V_ℓ is a proper subset and $\dim V_\ell \leqslant s-1$. Lemma 14.26 completes the proof. □

14.10. **Dual vectors on the dual quadric.** The sum over λ does not cancel when $F^*_{\vec{a}}(\mathbf{u}) = 0$. For the flag terms, $\vec{a} = \vec{c}$ is fixed and $F^*_{\vec{c}}$ is a non-singular quadratic form in s variables, so the relevant count is the number of its zeros in a box. We use the bound

$$\#\{\mathbf{u} \in \mathbb{Z}^s : \ 0 < |\mathbf{u}| \leqslant L, \ F^*_{\vec{c}}(\mathbf{u}) = 0\} \ll_\varepsilon L^{s-2+\varepsilon}, \tag{14.12}$$

uniformly in $\vec{c}$. This is Heath-Brown's bound [HB02, Theorem 2]: a quadratic form of rank at least 3 in n variables has $\ll_{\varepsilon,n} B^{n-2+\varepsilon}$ primitive integral zeros in a box of side B, with an implied constant independent of the form; summing over the multiples of primitive zeros gives the same bound for all integral zeros when $n \geqslant 4$. We apply it to the non-singular form $F^*_{\vec{c}}$, and in the same way below to the non-singular forms $t_{\vec{c}}$ and $t'_{\vec{c}_0}$; compare [LRV24, §9], where bounds of this kind are used for the dual forms of a pencil. For the point and plane terms, $\vec{a}$ runs over a family, and for fixed $\mathbf{u} \neq \mathbf{0}$ the proportion of $\vec{a}$ with $p \mid F^*_{\vec{a}}(\mathbf{u})$ is $\ll p^{-1}$, unless p divides the content of the polynomial $\vec{a} \mapsto F^*_{\vec{a}}(\mathbf{u})$; the latter primes are again removed on average over q.

14.11. **The integrals for small moduli.** The saving from the sum over λ is $q^{1/2}$, which is small when q is small. On the other hand, Lemma 11.1 shows that the whole curve $YX \asymp Q$ contributes $Q^{s/2}$, including its end $Y \asymp 1$, $X \asymp Q$. For small q the saving has to come from the exponential integral, and for this the bound of Lemma 7.1(ii) is too generous: the exponent $(s-2)/2$ is only needed for $\vec{w}$ close to a direction of corank 2. Indeed, the integral loses $X^{1/2}$ for every eigenvalue of $M_{\hat{\vec{w}}}$ which is small, and two eigenvalues are small only near the finitely many real points of $\mathcal{Z}_2$. Away from them, only one eigenvalue can be small, and the integral saves an extra $X^{1/2}$. The real points of $\mathcal{Z}_2$ are algebraic, so rational directions cannot approach them too closely (Lemma 14.29), and the extra saving holds on average over the weights (Lemma 14.30).

For a unit vector $\hat{\vec{w}} \in \mathbb{R}^3$ let $\sigma_1(\hat{\vec{w}}) \leqslant \sigma_2(\hat{\vec{w}}) \leqslant \dots$ be the absolute values of the eigenvalues of $M_{\hat{\vec{w}}}$, and let $\eta(\hat{\vec{w}})$ be the distance from $\pm\hat{\vec{w}}$ to the finite set of real points of $\mathcal{Z}_2$, with $\eta(\hat{\vec{w}}) = 1$ if there are none.

Lemma 14.28. *(i) For $\vec{w} \neq \vec{0}$ and all $\mathbf{u}$ we have $I_q(\vec{w}, \mathbf{u}) \ll P^s X_{\vec{w}}^{-(s-2)/2}\big(1 + X_{\vec{w}}\sigma_2(\hat{\vec{w}})\big)^{-1/2}$.*

(ii) $\sigma_2(\hat{\vec{w}}) \gg \eta(\hat{\vec{w}})$. Consequently $I_q(\vec{w}, \mathbf{u}) \ll P^s X_{\vec{w}}^{-(s-1)/2} \min\big(X_{\vec{w}}, \eta(\hat{\vec{w}})^{-1}\big)^{1/2}$.

Proof. (i) This is the bound displayed in the proof of Lemma 7.1(ii), keeping the factor coming from the eigenvalue $\frac{1}{2}P^2|\vec{w}|\sigma_2(\hat{\vec{w}})$ and discarding only the one from σ_1.

(ii) The second smallest singular value can only vanish at a point of corank two, and the claim is that it vanishes there linearly, not to higher order. This follows from Lemma 6.1(v): moving away from a corank-two point in any direction $\vec{b}$, the

restriction of $M_{\vec{b}}$ to the kernel is non-zero, and it pushes at least one of the two zero eigenvalues away from 0 at linear speed. The function σ_2 is continuous on the unit sphere, and it vanishes exactly at the real points of $\mathcal{Z}_2$, since $M_{\hat{\vec{w}}}$ has corank at most 2. So $\sigma_2 \gg 1$ outside any fixed neighbourhood of these finitely many points, and it suffices to prove $\sigma_2(\hat{\vec{w}}) \gg t$ for $\hat{\vec{w}}$ close to $\pm\vec{\beta}_0$, where $\vec{\beta}_0$ is a real unit vector with $[\vec{\beta}_0] \in \mathcal{Z}_2$ and t is the distance from $\hat{\vec{w}}$ to $\pm\vec{\beta}_0$; replacing $\hat{\vec{w}}$ by $-\hat{\vec{w}}$, which does not change σ_2, we may assume that $\hat{\vec{w}}$ is close to $\vec{\beta}_0$. Write $\hat{\vec{w}} = (\vec{\beta}_0 + t'\vec{b})/|\vec{\beta}_0 + t'\vec{b}|$ with $\vec{b} \perp \vec{\beta}_0$ a unit vector; then $t' \asymp t$ and $|\vec{\beta}_0 + t'\vec{b}| \asymp 1$, so it suffices to bound the singular values of $M = M_{\vec{\beta}_0} + t'M_{\vec{b}}$.

Let $W = \ker M_{\vec{\beta}_0} \subset \mathbb{R}^s$, of dimension 2, and $W' = W^\perp$. In an orthonormal basis adapted to $W \oplus W'$,

$$M_{\vec{\beta}_0} = \begin{pmatrix} 0 & 0 \\ 0 & N \end{pmatrix}, \qquad M_{\vec{b}} = \begin{pmatrix} A_{\vec{b}} & B_{\vec{b}} \\ B_{\vec{b}}^T & D_{\vec{b}} \end{pmatrix},$$

where N is invertible, say $|N\mathbf{x}| \geqslant c|\mathbf{x}|$, and $A_{\vec{b}}$ is the matrix of $F_{\vec{b}}|_W$. By Lemma 6.1(v), $A_{\vec{b}} \neq 0$ for every unit $\vec{b} \perp \vec{\beta}_0$, so by compactness $A_{\vec{b}}$ has an eigenvalue $\lambda_{\vec{b}}$ with $|\lambda_{\vec{b}}| \geqslant c_1 > 0$, uniformly in $\vec{b}$; let $\mathbf{e}_{\vec{b}} \in W$ be a unit eigenvector. Also $|B_{\vec{b}}|, |D_{\vec{b}}| \leqslant C$. Since M is symmetric, its singular values are the absolute values of its eigenvalues, and by the min-max principle $\sigma_2(M) \geqslant \min\{|M\mathbf{v}| : \ \mathbf{v} \in V, \ |\mathbf{v}| = 1\}$ for every subspace V of dimension $s-1$. We take $V = \mathbb{R}\mathbf{e}_{\vec{b}} \oplus W'$ and write $\mathbf{v} = \alpha\mathbf{e}_{\vec{b}} + \mathbf{w}$ with $\mathbf{w} \in W'$ and $\alpha^2 + |\mathbf{w}|^2 = 1$. Then

$$M\mathbf{v} = \big(t'\alpha\lambda_{\vec{b}}\mathbf{e}_{\vec{b}} + t'B_{\vec{b}}\mathbf{w}\big) + \big(t'\alpha B_{\vec{b}}^T\mathbf{e}_{\vec{b}} + (N + t'D_{\vec{b}})\mathbf{w}\big),$$

with the first bracket in W and the second in W'. Let $K = 4C/c$. If $|\mathbf{w}| \geqslant Kt'$, the second bracket has length $\geqslant (c - Ct')|\mathbf{w}| - Ct' \geqslant \frac{c}{2}Kt' - Ct' = Ct'$ for $t' \leqslant c/(2C)$. If $|\mathbf{w}| < Kt'$, then $|\alpha| \geqslant \frac{1}{2}$ for small t', and the first bracket has length $\geqslant \frac{1}{2}c_1t' - CKt'^2 \geqslant \frac{1}{4}c_1t'$ for $t' \leqslant c_1/(4CK)$. In both cases $|M\mathbf{v}| \gg t'$, so $\sigma_2(\hat{\vec{w}}) \gg t \asymp \eta(\hat{\vec{w}})$. The second statement follows from (i), since $X^{-1/2}(1 + X\eta)^{-1/2} \leqslant X^{-1}\min(X, \eta^{-1})^{1/2}$. □

On average over the support of the weights, the factor $\min(X, \eta^{-1})^{1/2}$ costs nothing. The point terms are spread over all directions $\hat{\vec{w}}$, and the plane terms over the great circle $\hat{\vec{w}} \perp \vec{\nu}$. For the flag terms, $p_{3,\vec{\nu},\vec{\rho},k,q}(\vec{w})$ is concentrated on $\vec{w}$ within $O(Q^{\varepsilon-4/3})$ of the line $\mathbb{R}\vec{\rho}$, by (4.8), so that $\hat{\vec{w}}$ is within $O(Q^\varepsilon/X)$ of $\pm\vec{c}$, where $\vec{\rho} = G\vec{c}$; hence $\eta(\hat{\vec{w}}) \geqslant \eta(\hat{\vec{c}})/2$ unless $\eta(\hat{\vec{c}}) \ll Q^\varepsilon/X$, and in all cases

$$\min\big(1, (X\eta(\hat{\vec{w}}))^{-1/2}\big) \ll Q^\varepsilon \min\big(1, (X\eta(\hat{\vec{c}}))^{-1/2}\big). \tag{14.13}$$

So for the flag terms we need to count directions $\vec{c}$ close to the real points of $\mathcal{Z}_2$. The real points of $\mathcal{Z}_2$ are algebraic, and we use the following consequence of the theorems of Roth and Schmidt.

Lemma 14.29. *(i) For every real point z of $\mathcal{Z}_2$ and every $\varepsilon > 0$ there is $c_{z,\varepsilon} > 0$ such that every primitive $\vec{c} \in \mathbb{Z}^3$ with $[\vec{c}] \neq z$ has angular distance at least $c_{z,\varepsilon}|\vec{c}|^{-2-\varepsilon}$ from z.*

(ii) Let $0 < \eta_0 \leqslant 1$ and $1 \leqslant H \leqslant Q^{2/3}$. Then

$$\sum_{\substack{\vec{c} \text{ primitive}, \ |\vec{c}| \asymp H \\ [\vec{c}] \notin \mathcal{Z}_2, \ \eta(\hat{\vec{c}}) \leqslant \eta_0}} \Big(\frac{U}{\lambda_1(\vec{c})} + \frac{U^2}{|\vec{c}|}\Big) \ll_\varepsilon \eta_0^{1-\varepsilon}\log(2H)\,\big(UH^{5/2} + U^2H^2\big).$$

(iii) For a fixed primitive $\vec\nu_0$, the number of primitive $\vec c \perp \vec\nu_0$ with $|\vec c| \asymp H$, $[\vec c] \notin \mathcal{Z}_2$ and $\eta(\hat{\vec c}) \leqslant \eta_0$ is $\ll_{\vec\nu_0,\varepsilon} \eta_0^{1-\varepsilon} H^2$.

Proof. Part (i) is a Diophantine approximation statement for the algebraic point z, and it depends on whether z is rational, lies on a rational line, or neither. Parts (ii) and (iii) count lattice points in a thin cone around z; the volume of the cone gives the factor η_0^2 or η_0, and part (i) shows that a cone containing a lattice point must be long enough for the volume bound to be accurate.

(i) If z is rational, the angle between $\vec c$ and z is $\gg |\vec c \times \vec z|/|\vec c| \gg |\vec c|^{-1}$, where $\vec z$ is a primitive integral representative. If z lies on a rational line $\vec\nu_0^\perp$ but is not rational, then for $\vec c \notin \vec\nu_0^\perp$ we have $|\vec\nu_0 \cdot \vec c| \geqslant 1$, so the angle is $\gg |\vec c|^{-1}$; and for $\vec c = x\mathbf{w}_1 + y\mathbf{w}_2$ on the line, with $\mathbf{w}_1, \mathbf{w}_2$ a basis of $\vec\nu_0^\perp \cap \mathbb{Z}^3$, the angle is $\gg |x/y - \alpha|$, where z corresponds to the real algebraic irrational α, and Roth's theorem [Rot55] gives $|x/y - \alpha| \gg_\varepsilon |y|^{-2-\varepsilon}$ (and the case $|x| \gg |y|$ is treated in the same way with y/x). If z lies on no rational line, normalise $z = (1, \alpha_1, \alpha_2)$ (after permuting coordinates); then $1, \alpha_1, \alpha_2$ are linearly independent over $\mathbb{Q}$, and Schmidt's theorem on simultaneous approximation to algebraic numbers [Sch80, Chapter VI], which says that $\|c_1\alpha_1\| \, \|c_1\alpha_2\| \gg_\varepsilon |c_1|^{-1-\varepsilon}$ for $c_1 \neq 0$, gives $\max_i \|c_1\alpha_i\| \gg_\varepsilon |c_1|^{-1/2-\varepsilon}$, so the angle is $\gg |\vec c|^{-3/2-\varepsilon}$. (Here $\|x\|$ is the distance to the nearest integer; if $|c_1| < \frac12 \max_i |c_{i+1}|$ the angle between $\vec c$ and z is $\gg 1$, and otherwise it is

$$\gg \max_i |c_{i+1}/c_1 - \alpha_i| \geqslant \max_i \|c_1\alpha_i\|/|c_1|.$$

)

(ii) It suffices to consider one real point z, and to count the primitive $\vec c$ in the cone $\mathcal{K} = \{\vec x : \ |\vec x| \leqslant 2H, \ \angle(\vec x, z) \leqslant \eta_0\}$, which is convex. If $\mathcal{K}$ contains a non-zero lattice point, it contains a primitive one, and by (i) this forces $H \geqslant H_* := c'\eta_0^{-1/(2+\varepsilon)}$. The volume of $\mathcal{K}$ is $\ll H^3\eta_0^2$, its projections to the coordinate planes have area $\ll H^2\eta_0$, and those to the axes have length $\ll H$; so by Davenport's lemma [Dav51] the number of lattice points in $\mathcal{K}$ is $\ll (H^3\eta_0^2 + H^2\eta_0 + H)\mathbf{1}_{H\geqslant H_*}$. Multiplying by U^2/H, this part is

$$\ll U^2H^2(\eta_0^2 + \eta_0 + H_*^{-2}) \ll U^2H^2\eta_0^{2/(2+\varepsilon)}.$$

For the part with $U/\lambda_1(\vec c)$, choose for each $\vec c$ a primitive $\vec y \perp \vec c$ with $|\vec y| = \lambda_1(\vec c) \ll H^{1/2}$. The plane $\vec y^\perp$ meets $\mathcal{K}$ only if it passes within angle η_0 of z, that is, $|\hat{\vec y} \cdot \vec z| \ll \eta_0$, and then in a sector of angle $\ll \eta_0$. Two primitive vectors $\mathbf{v} \neq \pm\mathbf{v}'$ of the lattice $\vec y^\perp \cap \mathbb{Z}^3$ have $\mathbf{v} \times \mathbf{v}' \in \mathbb{Z}\vec y \setminus \{\vec 0\}$, so $|\mathbf{v} \times \mathbf{v}'| \geqslant |\vec y|$, and consecutive primitive vectors of length $\leqslant 2H$ in the sector are at angle $\geqslant |\vec y|/(4H^2)$; hence the sector contains $\ll 1 + H^2\eta_0/|\vec y|$ of them, and none unless $H \geqslant H_*$. The number of $\vec y$ with $|\vec y| \asymp Y'$ and $|\hat{\vec y} \cdot \vec z| \ll \eta_0$ is $\ll Y'^3\eta_0 + Y'^2$, by Davenport's lemma for the corresponding slab. Hence

$$\begin{aligned}\sum \frac{1}{\lambda_1(\vec c)} &\ll \mathbf{1}_{H\geqslant H_*} \sum_{Y' \ll H^{1/2}} \frac{Y'^3\eta_0 + Y'^2}{Y'}\Big(1 + \frac{H^2\eta_0}{Y'}\Big) \\ &\ll \mathbf{1}_{H\geqslant H_*}\Big(H\eta_0 + H^{1/2} + H^{5/2}\eta_0^2 + H^2\eta_0 \log 2H\Big),\end{aligned}$$

the sums being over dyadic Y'. This is $\ll H^{5/2}\log(2H)(\eta_0 + H_*^{-2})$, which is $\ll H^{5/2}\log(2H)\,\eta_0^{2/(2+\varepsilon)}$.

(iii) The sector argument with $\vec y = \vec\nu_0$ gives

$$\ll (1 + H^2\eta_0/|\vec\nu_0|)\mathbf{1}_{H\geqslant H_*} \ll H^2(\eta_0 + H_*^{-2}).$$

□

Lemma 14.30. *In each dyadic range (Y,W), the following hold for $X = 1 + P^2W$.*

(i) The average of $\min(1,(X\eta(\hat{\vec{w}}))^{-1/2})$ against the mass $\int |p_{1,q}|\,d\vec{w}$ of the point terms, and against the mass of the plane terms with $\vec{\nu}/\gcd(\vec{\nu}) \notin \mathcal{N}_F$, is $\ll Q^{\varepsilon}X^{-1/2}$.

(ii) For every $\eta_0 \in (0,1]$ and every dyadic H, the weights $\Omega^{+}_{\vec{c}} = Y^2 Z_{\vec{c}}(U/\lambda_1(\vec{c}) + U^2/|\vec{c}|)$ of Proposition 14.24 satisfy

$$\sum_{|\vec{c}|\asymp H,\ \eta(\hat{\vec{c}})\leqslant\eta_0} \Omega^{+}_{\vec{c}} \ll_{\varepsilon} Q^{\varepsilon}\eta_0^{1-\varepsilon} \sum_{|\vec{c}|\asymp H} \Omega^{+}_{\vec{c}},$$

the sums being over the primitive $\vec{c} \notin \mathcal{Z}_2$ with $|\vec{c}| \leqslant Q^{2/3}$. Consequently

$$\sum_{\vec{c}} \Omega^{+}_{\vec{c}} \min\bigl(1,(X\eta(\hat{\vec{c}}))^{-1/2}\bigr) \ll_{\varepsilon} Q^{\varepsilon}X^{-1/2}\sum_{\vec{c}} \Omega^{+}_{\vec{c}},$$

and the same holds for the singular $\vec{c}$ on a fixed rational line contained in $\Delta = 0$, with their own weights.

Proof. In each case we show that the weights put a proportion $\ll \eta_0$ (up to Q^{ε}) of their mass on the directions within η_0 of a real corank-two point; a dyadic decomposition in η_0 then shows that the factor $\min(1,(X\eta)^{-1/2})$ is $X^{-1/2}$ on average.

(i) The weight $p_{1,q}$ is radial, so the directions $\hat{\vec{w}}$ are distributed according to surface measure, and the measure of the η_0-neighbourhood of the finitely many real points of $\mathcal{Z}_2$ is $\ll \eta_0^2$. By the proof of Proposition 1.3, the mass of a plane term $p_{2,\vec{\nu},k,q}$ at $|\vec{w}| \asymp W \geqslant Q^{-4/3}$ lies within $O(Q^{-4/3})$ of the plane $\vec{\nu}^{\perp}$ and is distributed uniformly in angle there, and a great circle meets the η_0-neighbourhood of a finite set in a set of length $\ll \eta_0$; for $W < Q^{-4/3}$ we have $X \asymp 1$ and there is nothing to prove. In both cases the proportion $\pi(\eta_0)$ of the mass with $\eta \leqslant \eta_0$ is $\ll \eta_0$, and splitting dyadically,

$$\int \min\bigl(1,(X\eta)^{-1/2}\bigr) \ll \pi(X^{-1}) + \sum_{X^{-1}<\eta_0=2^k/X\leqslant 1} \pi(2\eta_0)(X\eta_0)^{-1/2} \ll X^{-1/2}.$$

(ii) For $|\vec{c}| \asymp H$ the weight $\Omega^{+}_{\vec{c}}$ is $Y^2Q^{2/3}H^{-1}$ times $U/\lambda_1(\vec{c}) + U^2/|\vec{c}|$, up to a constant, so the first statement follows from Lemma 14.29(ii) and from the lower bound $\sum_{|\vec{c}|\asymp H}(U/\lambda_1(\vec{c}) + U^2/|\vec{c}|) \gg U^2H^2$. The second follows by the dyadic splitting in (i), applied in each dyadic range of H. For the singular $\vec{c}$ on a fixed rational line we use Lemma 14.29(iii) in the same way. □

14.12. **The exceptional planes.** In this subsection we treat the plane terms with $\vec{\nu}/\gcd(\vec{\nu}) \in \mathcal{N}_F$. They carry a small part of the mass, and need only a small saving. We first extend Lemma 14.15 to forms with multiple roots. For these planes the line $\ell = \vec{\nu}_0^{\perp}$ either lies in the discriminant curve, or meets it with multiplicity at least three somewhere, so the arguments of §14.6 lose too much. But there are only finitely many such $\vec{\nu}_0$, and a modest saving suffices.

Lemma 14.31. *(i) Let $G \in \mathbb{Z}[x,y]$ be a non-zero binary form of degree d whose roots have multiplicity at most m. Then for $j \geqslant 0$,*

$$\mu\{\mathbf{z} \in \mathbb{Z}_p^2 \text{ primitive}:\ v_p(G(\mathbf{z})) \geqslant j\} \leqslant d\,p^{-\lceil (j-e)/m\rceil}$$

with $e = v_p(E_G)$.

(ii) *Let $\ell = \vec{\nu}_0^{\perp}$ be a rational line not contained in the curve $\Delta = 0$, and let $E_\ell = E_{G_{\vec{\nu}_0}}$. For $n \geqslant 1$, the average of $K(\vec{a}, p^n)^{1/2}$ over the primitive $\vec{a} \bmod p^n$ on ℓ is at most C_s if $n = 1$ and $p \nmid 2D_F \operatorname{cont}(G_{\vec{\nu}_0})$, and at most $C_p(n+1)p^{v_p(E_\ell)/s} p^{n(1/2-1/(2s))}$ in general, with $C_p = C_s$ for $p \nmid 2D_F$.*

(iii) *Let $\ell = \vec{\nu}_0^{\perp}$ be a rational line contained in the curve $\Delta = 0$, let $\mathbf{x}_0$ be the common vertex of the pencil ℓ (Lemma 6.4), normalised to be primitive, and let $U \in \mathrm{GL}_s(\mathbb{Z})$ have first column $\mathbf{x}_0$. Then for every $\vec{a}$ on ℓ, $U^T M_{\vec{a}} U = \operatorname{diag}(0, N_{\vec{a}})$, where $N_{\vec{a}}$ depends linearly on $\vec{a}$ and $G'_\ell(\vec{a}) = \det N_{\vec{a}}$ is a non-zero binary form, and*

$$S_q(\vec{a}, \mathbf{u}) = q\, \mathbf{1}_{q \mid \mathbf{x}_0 \cdot \mathbf{u}}\, S_q^{N_{\vec{a}}}(1, \mathbf{v}'), \qquad (\mathbf{x}_0 \cdot \mathbf{u}, \mathbf{v}') = U^T \mathbf{u}.$$

The average over the primitive $\vec{a} \bmod p^n$ on ℓ of $K_{N_{\vec{a}}}(p^n)^{1/2}$, where $K_N(p^n)$ is the number of solutions of $N\mathbf{y} \equiv \mathbf{0} \pmod{p^n}$, satisfies the bounds in (ii) with $s-1$ in place of s and $E_{G'_\ell}$ in place of E_ℓ.

Proof. Part (i) is Lemma 14.15 with roots of multiplicity m in place of 2: near a root of multiplicity m, G behaves like the m-th power of a linear form. Part (ii) applies (i) with the crude bound $m \leqslant s$, separating the two invariant factors of $M_{\vec{a}}$ by Cauchy's inequality. In part (iii) the common kernel vector of the pencil is split off, which reduces to a non-singular situation in one variable fewer.

(i) Let δ be the largest distance $d(\boldsymbol{\alpha}, \boldsymbol{\alpha}')$ between distinct roots, so that $\delta \leqslant v_p(\operatorname{Disc}(G^{\mathrm{red}}))$ as in Lemma 14.15(i), and let $\mathbf{z}$ be primitive with $t = \max_i d(\mathbf{z}, \boldsymbol{\alpha}_i)$ attained at $\boldsymbol{\alpha}$. If $t > \delta$, the ultrametric inequality gives $d(\mathbf{z}, \boldsymbol{\alpha}') \leqslant \delta$ for every root $\boldsymbol{\alpha}' \neq \boldsymbol{\alpha}$, so by (14.4) $v_p(G(\mathbf{z})) \leqslant \kappa + mt + (d-m)\delta \leqslant e + mt$; if $t \leqslant \delta$, then $v_p(G(\mathbf{z})) \leqslant \kappa + d\delta \leqslant e$. So $v_p(G(\mathbf{z})) \geqslant j > e$ forces $d(\mathbf{z}, \boldsymbol{\alpha}_i) \geqslant (j-e)/m$ for some i, and the measure of this set is at most $p^{-\lceil (j-e)/m \rceil}$ for each i, as in the proof of Lemma 14.15(ii).

(ii) Let p be good. As in the proof of Lemma 14.16,

$$K(\vec{a}, p^n) = p^{\min(v_1, n) + \min(v_2, n)}, \qquad v_1 = v_p(D_{s-1}(\vec{a})) \leqslant v_2, \qquad v_1 + v_2 = v_p(G_{\vec{\nu}_0}(\mathbf{z})),$$

so that

$$K^{1/2} \leqslant p^{\min(v_1, n)/2} p^{\min(v_p(G_{\vec{\nu}_0}(\mathbf{z})), n)/2}.$$

By Cauchy's inequality the average of $K^{1/2}$ is at most the geometric mean of the averages of $p^{\min(v_1, n)}$ and of $p^{\min(v_p(G_{\vec{\nu}_0}), n)}$. The first is

$$\leqslant 1 + \sum_{k \leqslant n} p^k \cdot 2(s-1)^2 p^{-k} \leqslant 4(s-1)^2(n+1),$$

by Lemma 15.2. The second is

$$\leqslant 1 + \sum_{j \leqslant n} p^j\, s\, p^{-(j-e)/s} \leqslant 2s\, p^{e/s}(n+1) p^{n(1-1/s)},$$

by (i) with $m \leqslant d = s$. This gives the general bound. For $n = 1$ and $p \nmid \operatorname{cont}(G_{\vec{\nu}_0})$, the reduction of $G_{\vec{\nu}_0}$ is non-zero, so at most s classes on the line have corank 1 and at most $(s-1)^2$ have corank 2, with $K = p$ and $K = p^2$ respectively, so the average is $\leqslant 1 + (sp^{1/2} + (s-1)^2 p)/(p+1) \leqslant C_s$. For $p \mid D_F$ the same argument applies with the changes in the proof of Lemma 14.16.

(iii) The vector $\mathbf{x}_0$ lies in $\ker M_{\vec{a}}$ for every $\vec{a} \in \ell$, which gives the block form and, by Lemma 14.19, the formula for S_q. The members of ℓ have corank 1 except for at most $(s-1)^2$ of corank 2, so $G'_\ell \not\equiv 0$. At a good prime the corank of $M_{\vec{a}}$ modulo p is at most 2, so that of $N_{\vec{a}}$ is at most 1, and $K_{N_{\vec{a}}}(p^n) = p^{\min(v_p(G'_\ell(\vec{a})), n)}$. The average

is then bounded by (i) as in (ii), now with $m \leqslant \deg G'_\ell = s-1$, without the factor coming from v_1. For $n=1$ and $p \nmid 2D_F \operatorname{cont}(G'_\ell)$, the classes on the line with $p \mid G'_\ell(\vec{a})$ have corank 2 modulo p, so there are at most $(s-1)^2$ of them, each with $K_{N_{\vec{a}}}(p) = p$, and the average is at most $1 + (s-1)^2 p^{1/2}/(p+1) \leqslant C_s$. □

For a fixed line the bounds of Lemma 14.31 cost only Y^ε on average over q: if f is the multiplicative function with $f(p^n)$ equal to the bound in (ii) (or in (iii), with $s-1$ in place of s), then

$$\sum_{q \leqslant Y} f(q) \ll_{\ell,\varepsilon} Y^{1+\varepsilon}, \tag{14.14}$$

since $f(p) \leqslant C_s$ for good p not dividing the content of the relevant binary form, and $\sum_{n \geqslant 2} p^{-n} f(p^n) \ll p^{-1-1/s}$ (respectively $p^{-1-1/(s-1)}$) for $p \nmid E_\ell D_F$, while the finitely many other primes contribute a bounded factor.

The numerators of a plane term $\vec{\nu} = g\vec{\nu}_0$ lie on the line $\ell = \vec{\nu}_0^\perp$ only modulo $q/\gcd(q,g)$: at a prime p with $p^n \parallel q$ and $f = \min(n, v_p(g))$ they form the set $\mathcal{A}_p(n, v_p(g))$ of primitive $\vec{a}$ mod p^n with $p^{n-f} \mid \vec{\nu}_0 \cdot \vec{a}$ (Lemma 14.10). For $f = n$ this is the set of all primitive residues, and for $0 < f < n$ it lies on the line only modulo p^{n-f}. Since $\vec{\nu}_0$ is fixed and g runs up to $Q^{1/3}$, these primes need some care; the following lemma gives the average we need. At the primes where q and g share their full power, the numerators run over all residues, so that the average of $K^{1/2}$ there is bounded (Proposition 8.7(i)); the crude bound $K(\vec{a}, p^n) \leqslant p^{2n}$ at these primes would give the weight $\gcd(q,g)^3/g$, whose sum is of order xG^2, and this would lose a factor $Q^{1/3}$ against the mass of the plane terms.

Lemma 14.32. *Let $\vec{\nu}_0 \in \mathcal{N}_F$ and $\ell = \vec{\nu}_0^\perp$. For a prime power p^n and $v \geqslant 0$ put $f = \min(n, v)$, and let $\psi_p(n, v)$ be the average over $\mathcal{A}_p(n, v)$ of $K(\vec{a}, p^n)^{1/2}$ if $\ell \not\subseteq \{\Delta = 0\}$, and of $(K(\vec{a}, p^n)/p^{n-f})^{1/2}$ if $\ell \subseteq \{\Delta = 0\}$. Put $\psi(q,g) = \prod_{p^n \parallel q} \psi_p(n, v_p(g))$. Then $\psi(q,g) \geqslant 1$, and for $x, G \geqslant 1$*

$$\sum_{q \leqslant x} \sum_{g \leqslant G} \frac{\gcd(q,g)^2\, \psi(q,g)}{g} \ll_{\vec{\nu}_0,\varepsilon} (xG)^{1+\varepsilon}.$$

Proof. The weight $\gcd(q,g)^2/g$ is large only when q and g share a large common factor, and at the primes of this common factor the numerators are spread over all residues, where the average of $K^{1/2}$ is bounded. We make this precise prime by prime, and then sum with Rankin's trick. Let $\lambda_p(n)$ be the bound of Lemma 14.31(ii) (if $\ell \not\subseteq \{\Delta = 0\}$) or (iii) (if $\ell \subseteq \{\Delta = 0\}$) for the average over the primitive $\vec{a}$ mod p^n on ℓ, with $\lambda_p(0) = 1$, and extend λ multiplicatively; by (14.14), $\sum_{q \leqslant x} \lambda(q) \ll x^{1+\varepsilon}$. Let $\kappa_p(n)$ be the bound for the average of $K(\vec{a}, p^n)^{1/2}$ over all primitive $\vec{a}$ mod p^n given by the proof of Proposition 8.7(i) for good p and by Lemma 14.12 for bad p; thus $\kappa_p(n) \leqslant C_p(n+1)^3$, with $C_p = 2(1 + 13s^2 p^{-1/2})$ for $p \nmid 2D_F$. We claim that

$$\psi_p(n,0) \leqslant \lambda_p(n), \qquad \psi_p(n,v) \leqslant \kappa_p(n) \quad (v \geqslant n),$$
$$\psi_p(n,v) \leqslant p^{\mu_p/2} p^f \lambda_p(n-f) \quad (0 < f < n).$$

If $f = 0$, every residue in $\mathcal{A}_p(n,0)$ is the reduction of a vector of $\ell \cap \mathbb{Z}^3$, since $\vec{\nu}_0$ is primitive; if $\ell \subseteq \{\Delta = 0\}$, Lemma 14.31(iii) gives $K(\vec{a}, p^n) = p^n K_{N_{\vec{a}}}(p^n)$ for these residues, and the first bound is the definition of $\lambda_p(n)$. If $v \geqslant n$, then $\mathcal{A}_p(n,v)$ is the set of all primitive residues and $f = n$. If $0 < f < n$, Lemmas 8.2(ii) and 8.4 give

$$K(\vec{a}, p^n) \leqslant K(\vec{a}, p^{n-f}) K(\vec{a}, p^f) \leqslant p^{\mu_p + 2f} K(\vec{a}, p^{n-f});$$

the set $\mathcal{A}_p(n,v)$ is the full preimage of $\mathcal{A}_p(n-f,0)$, and $K(\vec{a},p^{n-f})$ depends only on $\vec{a} \bmod p^{n-f}$, which gives the third bound (after division by p^{n-f} in the second case). The lower bound $\psi \geqslant 1$ is clear if $\ell \not\subseteq \{\Delta = 0\}$; otherwise the common kernel vector $\mathbf{x}_0$ of the pencil gives $p^f\, t\, \mathbf{x}_0 \in \mathcal{K}(\vec{a},p^n)$ for all $t \bmod p^{n-f}$, as in the proof of Lemma 8.3, so $K(\vec{a},p^n) \geqslant p^{n-f}$.

Write $q = q_1q_2$, where q_2 is composed of the primes dividing g and $\gcd(q_1,g)=1$. Then $\gcd(q,g) = \gcd(q_2,g)$ and $\psi(q,g) \leqslant \lambda(q_1)\psi(q_2,g)$, so, summing first over $q_1 \leqslant x/q_2$ and using

$$\sum_{q_1 \leqslant x/q_2} \lambda(q_1) \ll (x/q_2)^{1+\varepsilon} \leqslant x^{1+\varepsilon}/q_2$$

,

$$\sum_{q\leqslant x} \frac{\gcd(q,g)^2\psi(q,g)}{g} \ll \frac{x^{1+\varepsilon}}{g} \sum_{q_2 \mid g^\infty} \frac{\gcd(q_2,g)^2\psi(q_2,g)}{q_2} \leqslant \frac{x^{1+\varepsilon}}{g} \prod_{p^v \| g} \sigma_p(v),$$

with $\sigma_p(v) = \sum_{n\geqslant 0} p^{-n} p^{2\min(n,v)} \psi_p(n,v)$. By the three bounds,

$$\sigma_p(v) \leqslant \sum_{n\leqslant v} p^n \kappa_p(n) + p^{\mu_p/2} p^{2v} \sum_{m\geqslant 1} p^{-m} \lambda_p(m) \leqslant C_p'(v+1)^4 \Big(p^v + p^{2v-1}\Big),$$

where C_p' is an absolute constant for the primes not dividing $2D_F E_\ell E_{G_\ell'}$ (for which $\sum_{m\geqslant 1} p^{-m}\lambda_p(m) \ll p^{-1}$), and is bounded for the finitely many other primes. Hence the left side of the lemma is $\ll x^{1+\varepsilon} \sum_{g\leqslant G} h(g)$ with $h(g) = \prod_{p^v\|g} C_p'(v+1)^4(1+p^{v-1})$. By Rankin's trick,

$$\sum_{g\leqslant G} h(g) \leqslant G^{1+\varepsilon} \prod_p \Big(1 + \sum_{v\geqslant 1} h(p^v) p^{-v(1+\varepsilon)}\Big),$$

and the product converges, since $h(p)p^{-1-\varepsilon} \ll p^{-1-\varepsilon}$ and $h(p^v)p^{-v(1+\varepsilon)} \ll (v+1)^4 p^{-1-v\varepsilon}$ for $v \geqslant 2$. □

Proposition 14.33. *The plane terms with $\vec{\nu}/\gcd(\vec{\nu}) \in \mathcal{N}_F$ and $\mathbf{u} \neq \mathbf{0}$ have relative size $\ll Q^{-1/2+\varepsilon} t^{-1}$.*

Proof. We give up the cancellation over the numerators completely, and use only Lemma 10.2 for each numerator. Since these terms carry a proportion $Q^{-2/3}$ of the mass, the loss $1 + Y^{1/2}/L \ll Q^{1/6}t^{-1}$ coming from large values of K is easily affordable. Let $\vec{\nu} = g\vec{\nu}_0$ with $\vec{\nu}_0 \in \mathcal{N}_F$; there are finitely many $\vec{\nu}_0$, and $g \leqslant U = Q^{1/3}$. Fix $q \in (Y, 2Y]$, and let $\mathcal{A}$ be the set of numerators of the plane term, so that $\mathcal{A} = \prod_p \mathcal{A}_p(n_p, v_p(g))$ by Lemma 14.10, and $\#\mathcal{A} \leqslant q^2 \gcd(q,g)$ by Lemma 5.7(i). We give up the cancellation in the sum over $\vec{a}$, and use Lemma 10.2 for each numerator:

$$\sum_{0<|\mathbf{u}|\leqslant L} \Big|\sum_{\vec{a}\in\mathcal{A}} S_q(\vec{a},\mathbf{u})\Big| \leqslant \sum_{\vec{a}\in\mathcal{A}} \sum_{|\mathbf{u}|\leqslant L} |S_q(\vec{a},\mathbf{u})| \ll q^{s/2} L^s \sum_{\vec{a}\in\mathcal{A}} \Big(1 + \frac{K(\vec{a},q)^{1/2}}{L}\Big).$$

Since $K(\vec{a},q) = \prod_p K(\vec{a},p^{n_p})$ and $\mathcal{A}$ is a product, the average of $K(\vec{a},q)^{1/2}$ over $\mathcal{A}$ is the product of the local averages. In the notation of Lemma 14.32 it is $\psi(q,g)$ if $\ell \not\subseteq \{\Delta = 0\}$, and $(q/\gcd(q,g))^{1/2}\psi(q,g) \leqslant (2Y)^{1/2}\psi(q,g)$ if $\ell \subseteq \{\Delta = 0\}$, because $\prod_p p^{n_p - f_p} = q/\gcd(q,g)$. In both cases

$$\sum_{0<|\mathbf{u}|\leqslant L} \Big|\sum_{\vec{a}\in\mathcal{A}} S_q(\vec{a},\mathbf{u})\Big| \ll \#\mathcal{A}\, q^{s/2} L^s \Big(1 + \frac{Y^{1/2}}{L}\Big)\psi(q,g).$$

The plane term carries the mass $\ll (Q/(kq))\min(W, r_Y)\min(W, Q^{-4/3})^2$ with $k = g/\gcd(q,g)$, and $(Q/(kq))\#\mathcal{A} \leqslant 2QY\gcd(q,g)^2/g$. In the proof of Proposition 11.2 the generic plane terms contributed, by Proposition 10.7, $Q^{\varepsilon}Y \cdot YQ^2\min(W, r_Y)\min(W, Q^{-4/3})^2$ times the remaining factors $q^{s/2}L^sP^sX^{-(s-2)/2}Y^{-s}$, which are the same here. So these terms have relative size

$$\ll Q^{\varepsilon}\Big(1+\frac{Y^{1/2}}{L}\Big)\frac{QY}{Y^2Q^2}\sum_{\vec{\nu}_0\in\mathcal{N}_F}\sum_{q\leqslant 2Y}\sum_{g\leqslant U}\frac{\gcd(q,g)^2\psi(q,g)}{g} \ll Q^{\varepsilon}\Big(1+\frac{Y^{1/2}}{L}\Big)\frac{U}{Q},$$

by Lemma 14.32. Since $U/Q = Q^{-2/3}$ and $Y^{1/2}/L \ll Q^{1/2}/(P^{1/2}t) \asymp Q^{1/6}t^{-1}$, the relative size is

$$\ll Q^{-2/3+\varepsilon}(1+Q^{1/6}t^{-1}) \ll Q^{-1/2+\varepsilon}t^{-1}.$$

□

14.13. **The flag terms with a singular direction.** It remains to treat the flag terms whose direction $\vec{c} = \vec{\rho}/\gcd(\vec{\rho})$ is singular. Here the sum over λ is replaced by the kernel of $M_{\vec{c}}$, which forces the dual vectors into a sublattice; the sparsity of the singular directions then gives the required bound.

The strategy is the same as for the generic flags, with the kernel splitting of Lemma 14.19 in place of Lemma 14.3. For a singular direction $\vec{c}$ of corank r, the Gauss sum $S_q(\lambda\vec{c}, \mathbf{u})$ is q^r times a Gauss sum for the non-singular form $N_{\vec{c}}$ in $s-r$ variables, and it vanishes unless q divides the r numbers $\mathbf{h}_j\cdot\mathbf{u}$. The factor q^r is a loss of $q^{r/2}$ against the generic case, but the congruence conditions reduce the number of dual vectors by q^r as long as q is not much larger than L. When q is larger than L, the conditions force $\mathbf{h}_j\cdot\mathbf{u} = 0$ exactly, which again leaves few dual vectors. The singular directions themselves are sparse: there are $O(H^2)$ of them up to height H, and only $O(H^{1+\varepsilon})$ if we exclude the rational lines contained in the curve $\Delta = 0$. The proof below organises these effects into several families, which are summarised in the table at the end of the proof.

Let $\vec{c} \in \mathcal{C}_F$ have corank $r \in \{1,2\}$, and fix $U_{\vec{c}} \in \mathrm{GL}_s(\mathbb{Z})$ as in Lemma 14.19, with first columns $\mathbf{h}_1, \dots, \mathbf{h}_r$ spanning $\ker M_{\vec{c}} \cap \mathbb{Z}^s$. Thus $U_{\vec{c}}^T M_{\vec{c}} U_{\vec{c}} = \mathrm{diag}(0_r, N_{\vec{c}})$ with $\det N_{\vec{c}} \neq 0$. For $\mathbf{u} \in \mathbb{Z}^s$ write $U_{\vec{c}}^T\mathbf{u} = (\mathbf{v}'', \mathbf{v}')$, so that $\mathbf{v}'' = (\mathbf{h}_1\cdot\mathbf{u}, \dots, \mathbf{h}_r\cdot\mathbf{u})$, put $t_{\vec{c}}(\mathbf{v}') = \mathbf{v}'^T\,\mathrm{adj}(N_{\vec{c}})\mathbf{v}'$, and $N'_{\vec{c}} = 2D_F\det N_{\vec{c}}$. We choose $C_2 \geqslant 1$, depending on $\vec{c}$ only through $U_{\vec{c}}$, such that $|\mathbf{v}''| \leqslant C_2L$ whenever $|\mathbf{u}| \leqslant L$. For the members on a rational line in the discriminant curve we take $U_{\vec{c}}$ to depend only on the line (Lemma 14.31(iii)), so that C_2 is bounded for them; for the finitely many other directions for which we use it, C_2 is bounded as well.

Lemma 14.34. *Let $(\vec{\nu}, \vec{\rho})$ be a flag with $\vec{\rho} = G\vec{c}$, $\vec{c} \in \mathcal{C}_F$ of corank r, let $q \in \mathbb{N}$, $d = \gcd(q, G)$ and $q_1 = q/d$. Let k be the largest divisor of q composed of primes dividing $dN'_{\vec{c}}$, put $q' = q/k$ and $k'' = k/d$, and let $\mathcal{A}_k$ be the set of residues modulo k of the elements of $\mathcal{A}^\flat$.*

(i) If $S_q(\vec{a}, \mathbf{u}) \neq 0$ for some $\vec{a} \in \mathcal{A}^\flat$, then $\mathbf{h}_j\cdot\mathbf{u} \equiv 0 \pmod{q_1}$ for $1 \leqslant j \leqslant r$.

(ii) $\#\mathcal{A}^\flat = \varphi(q')\#\mathcal{A}_k$ and

$$\sum_{\vec{a}\in\mathcal{A}^\flat} S_q(\vec{a}, \mathbf{u}) = q'^r\mathbf{1}_{q'|\mathbf{v}''}\,T_{q'}^{N_{\vec{c}}}(\mathbf{v}')\sum_{\vec{a}\in\mathcal{A}_k} S_k(\vec{a}, \mathbf{u}),$$

where $|T_{q'}^{N_{\vec{c}}}(\mathbf{v}')| \leqslant \varphi(q')q'^{(s-r)/2}$ and $|T_{q'}^{N_{\vec{c}}}(\mathbf{v}')| \leqslant q'^{(s-r+1)/2}\gcd(q', t_{\vec{c}}(\mathbf{v}'))^{1/2}$.

(iii) For $\vec{a} \in \mathcal{A}_k$ we have $K(\vec{a}, k) \leqslant K(\vec{c}, k'')K(\vec{a}, d)$ and $K(\vec{c}, k'') \leqslant c_F(k'')^2$; and $\mathbf{u} \bmod k \in \mathcal{M}(\vec{a}, k)$ implies $\mathbf{h}_j\cdot\mathbf{u} \equiv 0 \pmod{k''}$.

Proof. This is the singular analogue of Lemma 14.23: at the primes of q' we use the kernel splitting, and at the primes of k only the size of the Gauss sums and the congruence conditions on $\mathbf{u}$.

(i) Every $\vec{a} \in \mathcal{A}^\flat$ satisfies $\vec{a} \equiv \lambda \vec{c} \pmod{q_1}$ by Lemma 5.7(ii), so this is Lemma 8.3 with $q' = q_1$. (ii) As in the proof of Lemma 14.23, the sum is $T_{q'}(\vec{c}, \mathbf{u})$ times the sum over $\mathcal{A}_k$, and $T_{q'}(\vec{c}, \mathbf{u}) = q'^r \mathbf{1}_{q'|\mathbf{v}''} T_{q'}^{N_{\vec{c}}}(\mathbf{v}')$ by Lemma 14.19. Every prime dividing q' is odd and does not divide $\det N_{\vec{c}}$, so $N_{\vec{c}}$ is invertible modulo p; the first bound follows from Lemma 8.1 for $N_{\vec{c}}$, and the second from Lemma 14.3(i) for $N_{\vec{c}}$, as remarked before Lemma 14.19. (iii) The first two statements are proved as (14.7). For the last, $\vec{a} \equiv \lambda \vec{c} \pmod{k''}$ since k'' divides q_1, and Lemma 8.3, applied with modulus k, shows that $S_k(\vec{a}, \mathbf{u}) \neq 0$ implies $\mathbf{h}_j \cdot \mathbf{u} \equiv 0 \pmod{k''}$. □

We next bound the averages of $K(\vec{a}, d)^{1/2}$ that occur. For $p \mid d$ we keep the notation $n_p = v_p(q)$, $\delta_p = v_p(d)$, $f_p = \min(\delta_p, v_p(g))$ of Lemma 14.20, and we write $d = d_0 d_1$, where d_1 is the product of p^{δ_p} over the primes $p \mid d$ with $n_p > \delta_p$. Note that these primes divide k''. For $\vec{\nu}_0 \notin \mathcal{N}_F$ let $E_{\vec{\nu}_0}$ be the integer of Lemma 14.15(i) for $G_{\vec{\nu}_0}$.

Lemma 14.35. *In the situation of Lemma 14.34, put* $\alpha = \#\mathcal{A}_k^{-1} \sum_{\vec{a} \in \mathcal{A}_k} K(\vec{a}, d)^{1/2}$. *Then* $\alpha \leqslant c_F d$. *If moreover* $\vec{\nu}_0 \notin \mathcal{N}_F$, *then*

$$\alpha \leqslant C^{\omega(d)} \tau(d)^2 \gcd(d_0, g) \gcd(d_0, E_{\vec{\nu}_0})^{1/2} d_1,$$

with C *depending only on* $\vec{F}$.

Proof. At the primes $p \mid d$ where q and G share their full power, the numerators run over a plane, and Lemma 14.16 applies; at the other primes of d we use the trivial bound. The set $\mathcal{A}_k$ is the product of the class of $\vec{c}$ modulo k'' and of the sets $\mathcal{A}_p$ for $p \mid d$, and $K(\vec{a}, d) = \prod_{p|d} K(\vec{a}, p^{\delta_p})$. The first bound is Lemma 8.4. For the second we treat the primes separately. If $p \mid d_1$ we use $K(\vec{a}, p^{\delta_p}) \leqslant p^{\mu_p + 2\delta_p}$. If $p \mid d_0$, then $n_p = \delta_p$, and by Lemma 14.20 the set $\mathcal{A}_p$ consists of the primitive $\vec{a}$ mod p^{n_p} with $p^{n_p - f_p} \mid \vec{\nu}_0 \cdot \vec{a}$; Lemma 14.16 bounds the average of $K(\vec{a}, p^{n_p})^{1/2}$ over it by $C_p p^{f_p} (n_p + 1)^2 p^{\min(e, n_p)/2}$, and $e \leqslant v_p(E_{\vec{\nu}_0})$ by Lemma 14.15(i). Taking the product over $p \mid d$ gives the bound, since $\prod_{p|d_0} p^{f_p} \leqslant \gcd(d_0, g)$. □

Before the main estimate we sort the singular flags into families of different mass. The idea is that the bound for the ratio $\mathcal{R}$ obtained in each family can be poor, but only in families which carry a correspondingly small proportion of the total mass.

We divide the flags with a singular direction into three families. For $\vec{c} \in \mathcal{C}_F$ let $\vec{y}_{\vec{c}}$ be a primitive vector of $\vec{c}^{\perp} \cap \mathbb{Z}^3$ of length $\lambda_1(\vec{c})$. A flag $(g\vec{\nu}_0, G\vec{c})$ is called *ordinary* if $\vec{\nu}_0 \neq \pm \vec{y}_{\vec{c}}$ and $\vec{\nu}_0 \notin \mathcal{N}_F$; it is *of the first kind* if $\vec{\nu}_0 = \pm \vec{y}_{\vec{c}}$; and it is *exceptional* if $\vec{\nu}_0 \neq \pm \vec{y}_{\vec{c}}$ and $\vec{\nu}_0 \in \mathcal{N}_F$. A primitive $\vec{\nu}_0 \perp \vec{c}$ with $\vec{\nu}_0 \neq \pm \vec{y}_{\vec{c}}$ is linearly independent of $\vec{y}_{\vec{c}}$, so its length is at least the second minimum $\lambda_2(\vec{c}) \asymp |\vec{c}|/\lambda_1(\vec{c})$ of $\vec{c}^{\perp} \cap \mathbb{Z}^3$, and the number of such $\vec{\nu}_0$ with $|\vec{\nu}_0| \leqslant R$ is $\ll R^2/|\vec{c}|$. Repeating the proof of Lemma 14.21, we see that the mass of the ordinary flags with direction $\vec{c}$ is $\ll (\log Q)^3 \Omega_{\vec{c}}^{\mathrm{ord}}$ and that of the flags of the first kind is $\ll (\log Q)^3 \Omega_{\vec{c}}^{(1)}$, where

$$\Omega_{\vec{c}}^{\mathrm{ord}} = Y^2 Z_{\vec{c}} \frac{U^2}{|\vec{c}|}, \qquad \Omega_{\vec{c}}^{(1)} = Y^2 Z_{\vec{c}} \frac{U}{\lambda_1(\vec{c})}.$$

The weight of a given g is $\ll \tau(g) g^{-2}$ in the first case and $\ll \tau(g) g^{-1}$ in the second.

We also divide the singular directions. By Lemma 6.1(iii), the set $\mathcal{C}_F^{(2)}$ of $\vec{c} \in \mathcal{C}_F$ of corank 2 is finite. Let $\ell_1, \dots, \ell_J$ be the rational lines contained in the curve $\Delta = 0$

(there are finitely many, since Δ is reduced), and let $\mathcal{C}_F^{\mathrm{lin}}$ be the set of $\vec{c} \in \mathcal{C}_F$ of corank 1 lying on one of them. Let $\mathcal{C}_F^{\mathrm{curve}}$ be the set of the remaining $\vec{c} \in \mathcal{C}_F$ of corank 1. Every $\vec{c} \in \mathcal{C}_F^{\mathrm{curve}}$ lies on a component of $\Delta = 0$ of degree at least 2, or on a component not defined over $\mathbb{Q}$; the latter contain only finitely many rational points, and the former contain $\ll H^{1+\varepsilon}$ primitive $\vec{c}$ with $|\vec{c}| \leqslant H$, by [HB02, Theorem 3]. Hence

$$\#\{\vec{c} \in \mathcal{C}_F^{\mathrm{curve}} : \ |\vec{c}| \leqslant H\} \ll_\varepsilon H^{1+\varepsilon}, \qquad \#\{\vec{c} \in \mathcal{C}_F^{\mathrm{lin}} : \ |\vec{c}| \leqslant H\} \ll H^2. \tag{14.15}$$

For $\vec{c} \in \mathcal{C}_F^{\mathrm{lin}}$ on the line $\ell_i = \vec{\nu}_i^{\perp}$ we have $\vec{\nu}_i \perp \vec{c}$, so $\lambda_1(\vec{c}) \leqslant |\vec{\nu}_i| \ll 1$; and for $|\vec{c}| \gg 1$ the vector $\vec{y}_{\vec{c}}$ is $\pm\vec{\nu}_i$. So the flags of the first kind with $\vec{c} \in \mathcal{C}_F^{\mathrm{lin}}$ are, apart from finitely many $\vec{c}$, exactly the flags along the line ℓ_i: all their numerators lie, modulo $q/\gcd(q,g)$, on the line ℓ_i.

The masses of the families are as follows. By (14.15), and since $Z_{\vec{c}} \asymp Q^{2/3}/H$ for $|\vec{c}| \asymp H$,

$$\sum_{\substack{\vec{c} \in \mathcal{C}_F^{\mathrm{curve}} \\ |\vec{c}| \asymp H}} \Big(\Omega_{\vec{c}}^{\mathrm{ord}} + \Omega_{\vec{c}}^{(1)}\Big) \ll Y^2 Q^{2/3+\varepsilon}\Big(\frac{U^2}{H} + U\Big), \qquad \sum_{\substack{\vec{c} \in \mathcal{C}_F^{\mathrm{lin}} \\ |\vec{c}| \asymp H}} \Omega_{\vec{c}}^{\mathrm{ord}} \ll Y^2 Q^{2/3} U^2, \tag{14.16}$$

and

$$\sum_{\vec{c} \in \mathcal{C}_F^{\mathrm{lin}}, |\vec{c}| \asymp H} \Omega_{\vec{c}}^{(1)} \ll Y^2 Q^{2/3} U H.$$

Relative to the total mass Y^2Q^2 these are proportions $Q^{-2/3+\varepsilon}H^{-1} + Q^{-1+\varepsilon}$, $Q^{-2/3}$ and $Q^{-1}H$ respectively. The directions in $\mathcal{C}_F^{(2)}$ have bounded height, so their ordinary flags have mass $\ll Y^2Q^{4/3}$, a proportion $Q^{-2/3}$, and their flags of the first kind a proportion Q^{-1}. The exceptional flags have $\vec{c} \perp \vec{\nu}_0$ for one of the finitely many $\vec{\nu}_0 \in \mathcal{N}_F$ with $\vec{\nu}_0 \neq \pm\vec{y}_{\vec{c}}$. For $\vec{c} \in \mathcal{C}_F^{\mathrm{curve}} \cup \mathcal{C}_F^{(2)}$ this leaves finitely many $\vec{c}$, since a line meets a component of degree at least 2 in finitely many points; for $\vec{c} \in \mathcal{C}_F^{\mathrm{lin}}$ on ℓ_i it leaves $\vec{c}$ on the intersection of ℓ_i with $\vec{\nu}_0^{\perp}$, where $\vec{\nu}_0 \neq \vec{\nu}_i$, so again finitely many $\vec{c}$, together with the finitely many $\vec{c}$ of bounded height for which $\vec{\nu}_i \neq \pm\vec{y}_{\vec{c}}$. So the exceptional flags have bounded directions and fixed $\vec{\nu}_0$, and their mass is $\ll Y^2Q^{2/3}(\log Q)^3$, a proportion $Q^{-4/3+\varepsilon}$.

We shall need the following bound for averages of multiplicative functions several times.

Lemma 14.36. *Let $C \geqslant 1$, $D \geqslant 2$, let E be a non-zero integer, and let $f \geqslant 0$ be a multiplicative function with*

$$f(p^\delta) \leqslant C(\delta+1)^C\Big(\gcd(p^\delta, E)^{1/2} + p^{\delta-1}\Big)$$

for all prime powers p^δ. Then $\sum_{d \leqslant D} f(d)/d \ll_{C,\varepsilon} (D|E|)^\varepsilon$.

Proof. The hypothesis says that $f(p^\delta)/p^\delta$ is about p^{-1} or smaller for $p \nmid E$, so the Euler product converges after a slight shift; the primes dividing E contribute a divisor-type factor. By Rankin's trick with $\sigma = \varepsilon$, the sum is at most $D^\sigma \prod_p\Big(1+\sum_{\delta \geqslant 1} f(p^\delta)p^{-\delta(1+\sigma)}\Big)$. If $p \nmid E$, then

$$\sum_{\delta \geqslant 1} \frac{f(p^\delta)}{p^{\delta(1+\sigma)}} \leqslant C\sum_{\delta \geqslant 1}(\delta+1)^C\Big(\frac{1}{p^{\delta(1+\sigma)}} + \frac{1}{p^{1+\delta\sigma}}\Big) \leqslant \frac{C_\sigma}{p^{1+\sigma}},$$

and if $p \mid E$, the same computation with $\gcd(p^\delta, E)^{1/2} \leqslant p^{\delta/2}$ gives the bound $C_\sigma p^{-1/2}$. Hence the product is

$$\ll_\sigma C_\sigma^{\omega(E)} \ll_\varepsilon |E|^\varepsilon.$$

□

Proposition 14.37. *The flag terms with $\vec\rho/\gcd(\vec\rho)\in\mathcal{C}_F$ and $\mathbf{u}\neq\mathbf{0}$ have relative size $\ll Q^{-1/2+\varepsilon}t^{-4}$.*

Proof. For each family we bound the ratio $\mathcal{R}$ of the sum over q' and $\mathbf{u}$ to the trivial quantity, and multiply by the proportion of the mass carried by the family. Corank one and corank two are treated separately, and within corank one the flags along a rational line in the discriminant curve need an extra argument, because they carry the largest proportion of mass, $Q^{-1/3}$. The final table collects the results.

Fix a range (Y,W) and a flag with direction $\vec c\in\mathcal{C}_F$ of corank r. We fix also $d\mid G$ and $k=dk''$, and sum over the $q'\leqslant 2Y/k$ coprime to $dN'_{\vec c}$ and over $\mathbf{u}$. As in the proof of Proposition 14.24, it suffices to bound the ratio $\mathcal{R}$ of this sum to the trivial quantity $\sum_{q'}\#\mathcal{A}^\flat q^{s/2}L^s$, and then to average $\mathcal{R}$ with the weight $\gcd(d,g)/d$ for d, which comes from the sum over G, and the weight $1/k''$ for k'', which comes from the number of q with a given k. By Lemma 14.34, $|S_k(\vec a,\mathbf{u})|=k^{s/2}K(\vec a,k)^{1/2}\mathbf{1}_{\mathbf{u}\in\mathcal{M}(\vec a,k)}$, and

$$\frac{1}{\#\mathcal{A}_k}\sum_{\vec a\in\mathcal{A}_k}K(\vec a,k)^{1/2}\leqslant K(\vec c,k'')^{1/2}\alpha\leqslant c_F k''\alpha. \tag{14.17}$$

Corank one. Let $r=1$. We use the first bound of Lemma 14.34(ii), which loses $q'^{1/2}$ against the trivial bound $\varphi(q')q'^{s/2}$. For fixed q' and $\vec a\in\mathcal{A}_k$, the vectors $\mathbf{u}$ with $q'\mid\mathbf{v}''$ and $\mathbf{u}\bmod k\in\mathcal{M}(\vec a,k)$ form a coset of a lattice of index $q'K(\vec a,k)$, since $\mathbf{h}_1$ is primitive and q' is coprime to k. By Lemma 10.1 it has $\ll L^s/(q'K(\vec a,k))+L^{s-1}$ points with $|\mathbf{u}|\leqslant L$. Averaging $q'^{\pm1/2}$ over $q'\leqslant 2Y/k$ with the weight $\varphi(q')q'^{s/2}$ of the trivial quantity, we get

$$\mathcal{R}\ll\Big(\frac{k}{Y}\Big)^{1/2}+\Big(\frac{Y}{k}\Big)^{1/2}\frac{K(\vec c,k'')^{1/2}\alpha}{L}\ll\Big(\frac{dk''}{Y}\Big)^{1/2}+\Big(\frac{Y}{d}\Big)^{1/2}\frac{k''^{1/2}\alpha}{L}.$$

The quantity α depends on k'' only through the set of primes $p\mid d$ which divide k''; these are the primes dividing d_1. More precisely, the proof of Lemma 14.35 gives $\alpha\leqslant\prod_{p\mid d}\alpha_p$, where $\alpha_p\leqslant p^{\mu_p/2}p^{\delta_p}$ if $p\mid k''$, while, for the ordinary flags,

$$\alpha_p\leqslant\beta_p:=C_p(\delta_p+1)^2\gcd(p^{\delta_p},g)\gcd(p^{\delta_p},E_{\vec\nu_0})^{1/2}$$

if $p\nmid k''$. A prime $p\mid d$ which divides k'' contributes a factor at least p to k'', and there are $\ll Q^\varepsilon$ values of k'' altogether (Lemma 14.7(ii)). Grouping the k'' according to the set of primes $p\mid d$ which divide them, we obtain, for the ordinary flags,

$$\sum_{k''}\frac{\alpha}{k''^{1/2}}\ll Q^\varepsilon\prod_{p\mid d}\big(\beta_p+p^{\mu_p/2}p^{\delta_p-1/2}\big),\qquad\sum_{k''}\frac{\alpha}{k''}\ll Q^\varepsilon\prod_{p\mid d}\big(\beta_p+p^{\mu_p/2}p^{\delta_p-1}\big). \tag{14.18}$$

For the ordinary flags, by (14.18) the average of $\mathcal{R}$ over k'' with the weight $1/k''$ is

$$\ll Q^\varepsilon\Big(\frac{d}{Y}\Big)^{1/2}+Q^\varepsilon\frac{Y^{1/2}}{L}d^{-1/2}\prod_{p\mid d}\big(\beta_p+p^{\mu_p/2}p^{\delta_p-1/2}\big).$$

Since $\gcd(p^{\delta_p},g)\geqslant 1$, the product is at most $\gcd(d,g)\prod_{p\mid d}\gamma_p$ with

$$\gamma_p=C_p(\delta_p+1)^2\gcd(p^{\delta_p},E_{\vec\nu_0})^{1/2}+p^{\mu_p/2}p^{\delta_p-1/2}.$$

We average over $d\leqslant Z'_{\vec c}$ with the weight $\gcd(d,g)/d$, which comes from the sum over G, and then over g with the weight $\tau(g)g^{-2}$ of the ordinary flags. The first term gives

$\ll Q^{\varepsilon}(Z'_{\vec{c}}/Y)^{1/2} \ll Q^{\varepsilon}$. In the second, the factor $\gcd(d,g)^2$ is absorbed by the sum over g, because

$$\sum_{g} \tau(g) g^{-2} \gcd(d,g)^2 \leqslant \sum_{e \mid d} e^2 \sum_{e \mid g} \tau(g) g^{-2} \ll \tau(d)^2.$$

The remaining sum is $\sum_{d \leqslant Z'_{\vec{c}}} f(d)/d$ with the multiplicative function

$$f(d) = \tau(d)^2 d^{-1/2} \prod_{p \mid d} \gamma_p,$$

and

$$f(p^{\delta}) \leqslant C(\delta+1)^{C}\Big(\gcd(p^{\delta}, E_{\vec{\nu}_0})^{1/2} + p^{\delta-1}\Big),$$

because $p^{-\delta/2} p^{\delta - 1/2} \leqslant p^{\delta-1}$ for $\delta \geqslant 1$ and p^{μ_p} is bounded. So Lemma 14.36 applies with $E = E_{\vec{\nu}_0}$, and $|E_{\vec{\nu}_0}| \ll Q^{C}$. Hence, for the ordinary flags,

$$\mathcal{R} \ll Q^{\varepsilon}\Big(1 + \frac{Y^{1/2}}{L}\Big) \ll Q^{1/6+\varepsilon} t^{-1} \tag{14.19}$$

on average. For the flags of the first kind and the exceptional flags we use the bound $\alpha \leqslant c_F d$ of Lemma 14.35; the same computation, now with

$$\sum_{d \leqslant Z'_{\vec{c}}} \gcd(d,g) d^{-1} (Y/d)^{1/2} d \ll \tau(g) Y^{1/2} Z'^{1/2}_{\vec{c}},$$

gives on average

$$\mathcal{R} \ll Q^{\varepsilon}\Big(1 + \frac{Y^{1/2} Z'^{1/2}_{\vec{c}}}{L}\Big). \tag{14.20}$$

The flags along a line. The bound (14.20) is not good enough for the flags of the first kind with $\vec{c} \in \mathcal{C}_F^{\mathrm{lin}}$, which carry a proportion $Q^{-1/3}$ of the mass. For them we use that all members of the line ℓ_i have the common kernel vector $\mathbf{x}_0 = \mathbf{h}_1$ (Lemma 14.31(iii)). Let $p \mid d$. By Lemma 14.20, with $\vec{\nu}_0 = \vec{\nu}_i$, the numerators in $\mathcal{A}_p$ are congruent modulo $p^{n_p - f_p}$ to vectors of $\ell_i \cap \mathbb{Z}^3$. The proof of Lemma 8.3 applies with $\lambda\vec{c}$ replaced by any vector of $\ell_i \cap \mathbb{Z}^3$, because every such vector $\vec{b}$ satisfies $M_{\vec{b}}\mathbf{x}_0 = \mathbf{0}$ and $F_{\vec{b}}(\mathbf{x}_0) = 0$. Hence $S_{p^{n_p}}(\vec{a}, \mathbf{u}) = 0$ unless $\mathbf{x}_0 \cdot \mathbf{u} \equiv 0 \pmod{p^{n_p - f_p}}$. Together with Lemma 14.34(i), $S_q(\vec{a}, \mathbf{u}) \neq 0$ implies $\mathbf{x}_0 \cdot \mathbf{u} \equiv 0 \pmod{q/e}$, where $e = \prod_{p \mid d} p^{f_p} = \gcd(d,g)$.

Suppose first that $q/e > C_2 L$. Then $\mathbf{x}_0 \cdot \mathbf{u} = 0$, so $\mathbf{u}$ lies in a fixed lattice of rank $s-1$ and there are $\ll L^{s-1}$ of them; for them the condition $q' \mid \mathbf{v}''$ holds trivially. We use the second bound of Lemma 14.34(ii) and sum over q' first, as in the proof of Proposition 14.24, and at the primes dividing k we use only $K(\vec{a}, k) \leqslant c_F k^2$. The vectors with $t_{\vec{c}}(\mathbf{v}') = 0$, of which there are $\ll L^{s-3+\varepsilon}$, lose $q'^{1/2}$ in the sum over q'. This gives

$$\mathcal{R} \ll Q^{\varepsilon}\Big(\frac{k}{L} + \Big(\frac{Y}{k}\Big)^{1/2} \frac{k}{L^2}\Big) \ll Q^{\varepsilon} \frac{dk''}{L}\Big(1 + \frac{Y^{1/2}}{L}\Big).$$

The factor k'' is cancelled by the weight $1/k''$. The flags with $d \asymp D$ have $|\vec{c}| \leqslant Q^{2/3}/D$, and by (14.16) they carry a proportion $\ll Q^{-1/3} D^{-1}$ of the total mass. Summing over dyadic D, these terms have relative size

$$\ll Q^{-1/3+\varepsilon} L^{-1}(1 + Y^{1/2} L^{-1}) \ll Q^{-1/2+\varepsilon} t^{-2}.$$

Suppose next that $q/e \leqslant C_2 L$. Then $Y < q \leqslant C_2 L e$, and since $YX \asymp tQ$ we have $X \gg tQ/(Le)$. Here we use the integral. We bound $\mathcal{R}$ as in the case of corank one above, with $\alpha \leqslant c_F d$: after the weight $1/k''$ it is $\ll Q^{\varepsilon}\big(1 + (Yd)^{1/2}/L\big)$, and since

$d \leqslant q \leqslant 2Y$ and $Y/L \leqslant C_2 e$, this is $\ll Q^\varepsilon e$, for every flag and every d in the present case. These $\vec{c}$ lie on finitely many rational lines and have corank one, and within each dyadic range of $|\vec{c}|$ the masses of the flags with a given e are comparable for the different $\vec{c}$. So by Lemmas 14.28 and 14.30(ii), with Lemma 14.29(iii) for the count, the integrals contribute on average over them the further factor $Q^\varepsilon X^{-1/2}$. Hence the ratio including the integral is

$$\ll Q^\varepsilon e\, X^{-1/2} \ll Q^\varepsilon e\Big(\frac{Le}{tQ}\Big)^{1/2} \ll Q^{-1/3+\varepsilon} e^{3/2},$$

since $L \ll Q^{1/3+\varepsilon} t$. Since $e \leqslant d \leqslant Z_{\vec{c}}$, the flags with $e \asymp E$ have $|\vec{c}| \leqslant Q^{2/3}/E$ and carry a proportion $\ll Q^{-1/3} E^{-1}$ of the mass; moreover $e \leqslant U = Q^{1/3}$. Summing over dyadic $E \leqslant Q^{1/3}$ we obtain relative size $\ll Q^{-2/3+\varepsilon} \sum_E E^{1/2} \ll Q^{-1/2+\varepsilon}$.

Corank two. Let $r = 2$, so that $\vec{c}$ is one of finitely many vectors, C_2 is bounded, and $K(\vec{c}, k'') \ll k''^2$. Suppose first that $q_1 = q/d > C_2 L$. Then Lemma 14.34(i) gives $\mathbf{v}'' = \mathbf{0}$, so there are $\ll L^{s-2}$ vectors $\mathbf{u}$, and we use the second bound of Lemma 14.34(ii), summing over q' first. This loses $q'^{1/2}$ against the trivial bound, and the $\mathbf{u}$ with $t_{\vec{c}}(\mathbf{v}') = 0$, of which there are $\ll L^{s-4+\varepsilon}$ by [HB02, Theorem 2], lose q'. With (14.17) we get

$$\mathcal{R} \ll Q^\varepsilon\Big(\Big(\frac{Y}{k}\Big)^{1/2} \frac{k''\alpha}{L^2} + \frac{Y}{k} \cdot \frac{k''\alpha}{L^4}\Big) = Q^\varepsilon\Big(\Big(\frac{Y}{d}\Big)^{1/2} \frac{k''^{1/2}\alpha}{L^2} + \frac{Y}{d} \cdot \frac{\alpha}{L^4}\Big).$$

Suppose next that $q_1 \leqslant C_2 L$. We use the first bound of Lemma 14.34(ii), which loses q', and we count the $\mathbf{u}$ in two ways: for each $\vec{a} \in \mathcal{A}_k$ they lie in a coset of index $q'^2 K(\vec{a}, k)$, and they satisfy $q_1 \mid \mathbf{v}''$, which leaves $\ll L^s/q_1^2$ of them. Hence

$$\mathcal{R} \ll q'\Big(\frac{1}{q'^2} + k''\alpha \min\Big(\frac{1}{L}, \frac{1}{q_1^2}\Big)\Big) \ll \frac{1}{q'} + \frac{q'k''\alpha}{q_1 L^{1/2}} = \frac{1}{q'} + \frac{\alpha}{L^{1/2}},$$

since $q_1 = q'k''$. For the ordinary flags we use the second bound of Lemma 14.35, and (14.18) (the first bound for the terms with $k''^{1/2}\alpha$, the second for the term with α, after the weight $1/k''$); the averages over d and g are bounded by Lemma 14.36 as before, and we find

$$\mathcal{R} \ll Q^\varepsilon\Big(1 + \frac{Y^{1/2}}{L^2} + \frac{Y}{L^4} + \frac{1}{L^{1/2}}\Big) \ll Q^\varepsilon\big(1 + Q^{-1/6} t^{-4}\big)$$

on average. For the flags of the first kind and the exceptional flags we use $\alpha \leqslant c_F d$, and find

$$\mathcal{R} \ll Q^\varepsilon(1 + Y^{1/2} Z_{\vec{c}}'^{1/2} L^{-2} + Y L^{-4} + Z_{\vec{c}}' L^{-1/2}) \ll Q^{1/2+\varepsilon} t^{-4},$$

since $Z_{\vec{c}}' \leqslant Q^{2/3}$ and $L \gg P^{1/2} t$.

Conclusion. By Lemma 14.5, the relative size of a family is at most the proportion of the total mass $Y^2 Q^2$ of the flag terms that it carries, multiplied by the average $\mathcal{R}$ of the ratio over the family. We combine the proportions computed before Lemma 14.36 with the bounds for $\mathcal{R}$ obtained above. Throughout we use $Y \leqslant Q$, $Z_{\vec{c}}' \leqslant Q^{2/3}$, $L \gg Q^{1/3} t$ and $t \ll Q^\varepsilon$; in particular

$$\frac{Y^{1/2}}{L} \ll Q^{1/6} t^{-1}, \qquad \frac{Y^{1/2} Z_{\vec{c}}'^{1/2}}{L} \ll Q^{1/2} t^{-1}, \qquad 1 \ll Q^\varepsilon t^{-1}. \tag{14.21}$$

There are $O(\log Q)$ dyadic ranges of $|\vec{c}|$, and the factors $\log Q$ are absorbed into Q^ε.

(a) Ordinary flags of corank one. By (14.16), the ordinary flags with $\vec{c} \in \mathcal{C}_F^{\mathrm{curve}}$ and $|\vec{c}| \asymp H$ carry a proportion $\ll Q^{-2/3+\varepsilon} H^{-1}$ of the mass, and those with $\vec{c} \in \mathcal{C}_F^{\mathrm{lin}}$ a

proportion $\ll Q^{-2/3}$. By (14.19) and (14.21), $\mathcal{R} \ll Q^{\varepsilon}(1+Y^{1/2}L^{-1}) \ll Q^{1/6+\varepsilon}t^{-1}$ on average. So these flags have relative size

$$\ll Q^{-2/3+\varepsilon}\cdot Q^{1/6+\varepsilon}t^{-1} = Q^{-1/2+2\varepsilon}t^{-1}.$$

(b) Flags of the first kind of corank one, not along a line. These are the flags of the first kind with $\vec{c}\in\mathcal{C}_F^{\mathrm{curve}}$, and those with one of the finitely many $\vec{c}\in\mathcal{C}_F^{\mathrm{lin}}$ for which $\vec{y}_{\vec{c}}\neq\pm\vec{\nu}_i$. For the first ones, $\lambda_1(\vec{c})\geqslant 1$ and (14.15) give

$$\sum_{\substack{\vec{c}\in\mathcal{C}_F^{\mathrm{curve}}\\ |\vec{c}|\asymp H}} \Omega_{\vec{c}}^{(1)} \ll Y^2\frac{Q^{2/3}}{H}\,U\cdot H^{1+\varepsilon} = Y^2Q^{1+\varepsilon},$$

a proportion $\ll Q^{-1+\varepsilon}$ of the mass. The second ones have bounded $|\vec{c}|$, so $\Omega_{\vec{c}}^{(1)}\ll Y^2Q^{2/3}U = Y^2Q$, again a proportion $\ll Q^{-1}$. For all of them we use (14.20), which by (14.21) gives $\mathcal{R}\ll Q^{\varepsilon}(1+Y^{1/2}Z_{\vec{c}}'^{1/2}L^{-1}) \ll Q^{1/2+2\varepsilon}t^{-1}$ on average. So these flags have relative size

$$\ll Q^{-1+\varepsilon}\Big(1+Y^{1/2}Q^{1/3}L^{-1}\Big) \ll Q^{-1+\varepsilon}\cdot Q^{1/2+\varepsilon}t^{-1} = Q^{-1/2+2\varepsilon}t^{-1}.$$

(c) Flags along a line. The remaining flags of the first kind of corank one are those with $\vec{c}\in\mathcal{C}_F^{\mathrm{lin}}$ and $\vec{y}_{\vec{c}}=\pm\vec{\nu}_i$, that is, the flags along the lines ℓ_i. They were treated above: the case $q/e > C_2L$ gives relative size $\ll Q^{-1/2+\varepsilon}t^{-2}$, and the case $q/e\leqslant C_2L$ gives $\ll Q^{-1/2+\varepsilon}$.

(d) Directions of corank two. The set $\mathcal{C}_F^{(2)}$ is finite. Its ordinary flags carry a proportion $\ll Q^{-2/3}$ of the mass and have $\mathcal{R}\ll Q^{\varepsilon}(1+Q^{-1/6}t^{-4})$; its flags of the first kind carry a proportion $\ll Q^{-1}$ and have $\mathcal{R}\ll Q^{1/2+\varepsilon}t^{-4}$. So they have relative size

$$\ll Q^{-2/3+\varepsilon}\Big(1+Q^{-1/6}t^{-4}\Big) + Q^{-1}\cdot Q^{1/2+\varepsilon}t^{-4} \ll Q^{-1/2+\varepsilon}t^{-4}.$$

(e) Exceptional flags. They carry a proportion $\ll Q^{-4/3+\varepsilon}$ of the mass. For them $\mathcal{R}$ is bounded by (14.20) if $\vec{c}$ has corank one, and by the bound for the flags of the first kind of corank two otherwise; in both cases $\mathcal{R}\ll Q^{1/2+5\varepsilon}t^{-4}$ by (14.21), since $t^{-1}\ll Q^{3\varepsilon}t^{-4}$. So they have relative size $\ll Q^{-5/6+6\varepsilon}t^{-4}$.

Since $t\ll Q^{\varepsilon}$, each of the bounds in (a)–(e) is $\ll Q^{-1/2+C\varepsilon}t^{-4}$ for a constant C, and this proves the proposition.

The table below summarises the cases. In the table, "mass" is the proportion of the total mass Y^2Q^2 of the flag terms carried by the family (for the flags along a line, by those with $d\asymp D$, respectively $e\asymp E$), $\mathcal{R}$ is the ratio to the trivial bound on average over the family (including the factor from the integrals where it is used), and the last column is their product, summed over the dyadic parameters. Throughout, $L\asymp Q^{1/3+\varepsilon}t$, $Y\leqslant Q$ and $Z_{\vec{c}}'\leqslant Q^{2/3}$.

family	mass	$\mathcal{R}$	relative size
corank 1, ordinary	$Q^{-2/3+\varepsilon}$	$Q^{\varepsilon}(1+Y^{1/2}/L)$	$Q^{-1/2+\varepsilon}t^{-1}$
corank 1, first kind, $\vec{c}\in\mathcal{C}_F^{\mathrm{curve}}$	$Q^{-1+\varepsilon}$	$Q^{\varepsilon}(1+Y^{1/2}Z_{\vec{c}}'^{1/2}/L)$	$Q^{-1/2+\varepsilon}t^{-1}$
along a line, $q/e > C_2L$	$Q^{-1/3}D^{-1}$	$Q^{\varepsilon}DL^{-1}(1+Y^{1/2}/L)$	$Q^{-1/2+\varepsilon}t^{-2}$
along a line, $q/e \leqslant C_2L$	$Q^{-1/3}E^{-1}$	$Q^{-1/3+\varepsilon}E^{3/2}$	$Q^{-1/2+\varepsilon}$
corank 2, ordinary	$Q^{-2/3}$	$Q^{\varepsilon}(1+Q^{-1/6}t^{-4})$	$Q^{-2/3+\varepsilon}t^{-4}$
corank 2, first kind	Q^{-1}	$Q^{1/2+\varepsilon}t^{-4}$	$Q^{-1/2+\varepsilon}t^{-4}$
exceptional	$Q^{-4/3+\varepsilon}$	$Q^{1/2+\varepsilon}t^{-4}$	$Q^{-5/6+\varepsilon}t^{-4}$

□

Thus the singular directions cause no loss even for $s=22$: they carry a proportion at most $Q^{-1/3}$ of the mass, and on them the kernel of $M_{\vec{c}}$, combined with the averages over the numerators of Lemma 14.16, recovers almost all of the saving of the sum over λ.

14.14. **The bookkeeping.** We can now collect the estimates. For each family of terms we recall the saving obtained, as a relative size compared with the bound $B(Y,W)$ of the unrefined method, and check that it is at least the saving P^{θ_s} required by Lemma 14.6, with a little to spare. The point and plane terms save the square root $Y^{1/2}$ from the sums over the numerators and $X^{1/2}$ from the integrals, which on the critical curve is $(YX)^{1/2}\asymp Q^{1/2}\asymp P^{3/4}$; the same is true of the generic flag terms, apart from one term discussed at the end. We now prove Theorem 14.1 by checking the hypothesis of Lemma 14.6 for each family of terms. Fix a dyadic range (Y,W), with X, t and L as in §14.2; recall that $L\asymp P^{1/2+\varepsilon}t$ and $Q^{1/3}\asymp P^{1/2}$. For $s=23$ the target is relative size $P^{-1/4-\kappa}t^{-s/2-2}$, and for $s=24$ it is $P^{-\kappa}t^{-s/2-2}$.

The vector $\mathbf{u}=\mathbf{0}$. As in the proofs of Propositions 11.2 and 12.1, these terms contribute $O(P^{3s/4-3/2+\varepsilon})$, which is $o(P^{s-6})$ for $s\geqslant 22$.

Point terms. For $C(\mathbf{u})\neq 0$, Proposition 14.14 saves $Y^{1/2}$ on average over q, uniformly in $\vec{w}$, and Lemma 14.28 with Lemma 14.30(i) saves $X^{1/2}$ on average over $\vec{w}$. So these terms have relative size

$$\ll Q^{\varepsilon}(YX)^{-1/2}=Q^{\varepsilon}(tQ)^{-1/2}\ll P^{-3/4+\varepsilon}t^{-1/2}.$$

The $\mathbf{u}$ with $C(\mathbf{u})=0$ number $\ll L^{s-3}$ (Lemma 14.27), and contribute relative size $\ll L^{-3}$.

Plane terms with $\vec{\nu}_0\notin\mathcal{N}_F$. For $C_{\vec{\nu}_0}(\mathbf{u})\neq 0$ and $t'_{\vec{c}_0}(\mathbf{u})\neq 0$, Propositions 14.17 and 14.18 save $Y^{1/2}$, and the factor $\gcd(q^\sharp,g)$ costs only Q^{ε}, as explained after Proposition 14.18; with Lemma 14.30(i) the relative size is $\ll P^{-3/4+\varepsilon}t^{-1/2}$. The remaining $\mathbf{u}$ form a proportion $\ll L^{-2+\varepsilon}$, by Lemma 14.27(ii) and the remark after Proposition 14.18, and the trivial bound gives relative size $\ll L^{-2+\varepsilon}$.

Plane terms with $\vec{\nu}_0\in\mathcal{N}_F$. By Proposition 14.33, relative size $\ll Q^{-1/2+\varepsilon}t^{-1}\ll P^{-3/4+\varepsilon}t^{-1}$.

Generic flag terms. Let $\vec{c}$ be non-singular. By (14.13) and Lemma 14.28, the integrals on the support of a flag term with direction $\vec{c}$ are

$$\ll Q^{\varepsilon}P^{s}X^{-(s-2)/2}\min\bigl(1,(X\eta(\hat{\vec{c}}))^{-1/2}\bigr).$$

By (14.2) and Proposition 14.24, the relative size of the terms with $F^*_{\vec{c}}(\mathbf{u}) \neq 0$ is therefore

$$\ll \frac{Q^\varepsilon}{Y^2Q^2} \sum_{\vec{c}} \Omega^+_{\vec{c}} \Big(\Big(\frac{Z'_{\vec{c}}}{Y} \Big)^{1/2} + \frac{Z'_{\vec{c}}}{L} \Big) \min\Big(1, (X\eta(\hat{\vec{c}}))^{-1/2}\Big).$$

In each dyadic range $|\vec{c}| \asymp H$ the factor in the large brackets is constant up to a factor 2, so Lemma 14.30(ii), applied in each such range, replaces the minimum by $Q^\varepsilon X^{-1/2}$. Lemma 14.22(ii) with $a = \frac{1}{2}$ and $a = 1$ then gives relative size

$$\begin{aligned} &\ll Q^\varepsilon X^{-1/2} \Big(\frac{1}{Y^{1/2}} + \frac{1}{L} \Big) \ll Q^\varepsilon \Big((tQ)^{-1/2} + X^{-1/2}L^{-1} \Big) \\ &\ll P^{-3/4+\varepsilon} t^{-1/2} + P^{-1/2+\varepsilon} X^{-1/2} t^{-1}. \end{aligned}$$

The terms with $F^*_{\vec{c}}(\mathbf{u}) = 0$ have relative size $\ll Q^\varepsilon L^{-2}$, by (14.10) and Lemma 14.22(ii) with $a = 1$.

Flag terms with singular direction. By Proposition 14.37, relative size $\ll Q^{-1/2+\varepsilon}t^{-4} \ll P^{-3/4+\varepsilon}t^{-4}$.

All these relative sizes are $\ll P^{-1/2+\varepsilon}t^{-4}$. This is $\ll P^{-\theta_s-\kappa}t^{-s/2-2}$ with $\kappa = \frac{1}{4} - \varepsilon$ for $s = 23$ and $\kappa = \frac{1}{2} - \varepsilon$ for $s = 24$. By Lemma 14.6 this proves Theorem 14.1.

We note for later use that every family except one has relative size $\ll P^{-3/4+\varepsilon}t^{-4}$. The exception is the part

$$P^{-1/2+\varepsilon} X^{-1/2} t^{-1} \tag{14.22}$$

of the generic flag terms, which comes from the term $Z'_{\vec{c}}/L$ in (14.8). It is smaller than $P^{-1/2-\kappa}$ unless $X \leqslant P^{2\kappa}$, that is, unless Y is close to Q. This is the only obstacle to $s = 22$, and it is removed in the next section.

Remark 14.38 (parity). For even s the sums over λ are Ramanujan sums (Lemma 14.3(ii)), and for the flag terms they save q rather than $q^{1/2}$, apart from the dual vectors on the dual quadrics. This does not help with (14.22), which comes from moduli whose part prime to $\gcd(q, G)N_{\vec{c}}$ is small, so we do not use it.

15. The case $s = 22$

In this section we prove the remaining case of Theorem 1.1.

Theorem 15.1. *Under the hypotheses of Theorem 6.5, the asymptotic formula* (1.1) *holds for $s = 22$, for some $\Delta > 0$.*

For $s = 22$ the target in Lemma 14.6 is relative size $P^{-1/2-\kappa}t^{-13}$. By the bookkeeping in §14.14, every family of terms meets this target, with $\kappa = \frac{1}{4} - \varepsilon$, except the part (14.22) of the generic flag terms. This part comes from the second term $Z'_{\vec{c}}/L$ in (14.8), which in turn comes from the moduli q whose part q' prime to $dN_{\vec{c}}$ is small, and for which $K(\vec{a}, k)$ is close to its largest possible value $c_F k^2$. In (14.9) the factor $K(\vec{c}, k'')^{1/2}$ was bounded instead by $D_{s-1}(\vec{c})^{1/2}$, where $D_{s-1}(\vec{c})$ is the greatest common divisor of the $(s-1) \times (s-1)$ minors of $M_{\vec{c}}$. A prime p divides $D_{s-1}(\vec{c})$ exactly when $M_{\vec{c}}$ has corank at least 2 modulo p. So what we need is that the members $\vec{c}$ which have corank two modulo many primes are rare. This is a transversality statement: modulo p^k the reductions of corank two lie on finitely many lines.

Let us explain the shape of the argument. A member $F_{\vec{c}}$ has corank two modulo p only if $[\vec{c}]$ reduces to one of the at most $(s-1)^2$ points of $\mathcal{Z}_2(\mathbb{F}_p)$, so the proportion of such $\vec{c}$ is about p^{-2}. If p^k divides $D_{s-1}(\vec{c})$, Lemma 15.2 below shows that $\vec{c}$ is

congruent modulo p^k to a multiple of one of these points, a condition of probability about p^{-2k}. Combining the primes, a divisor $m > T$ of $D_{s-1}(\vec{c})$ forces $\vec{c}$ into one of a few lattices of index m^2. Since we do not know m exactly, we have to sum the resulting proportions m^{-2} over $m > T$, and the proportion of $\vec{c}$ with $D_{s-1}(\vec{c}) > T$ comes out as about T^{-1} (Lemma 15.6); after the separate treatment of the bad primes we use the weaker bound $T^{-1/2}$ (Lemma 15.7). This sparsity is then weighed against the loss $D_{s-1}(\vec{c})^{1/2}$ in (14.9). In §15.1 we prove the sparsity statements, and in §15.2 we balance the two bounds (14.8) and (14.9) with a threshold T.

15.1. **Members with many reductions of corank two.** For $\vec{c}$ with $\Delta(\vec{c}) \neq 0$ write $D^{\mathrm{good}}(\vec{c})$ for the largest divisor of $D_{s-1}(\vec{c})$ coprime to D_F.

Lemma 15.2. *Let p be a good prime and $k \geqslant 1$. If $\vec{c}, \vec{c}' \in \mathbb{Z}^3$ are primitive, p^k divides both $D_{s-1}(\vec{c})$ and $D_{s-1}(\vec{c}')$, and $\vec{c} \equiv \mu\vec{c}'$ (mod p) for some integer μ, then $\vec{c} \equiv \lambda\vec{c}'$ (mod p^k) for some integer λ coprime to p. Consequently, for every integer $m \geqslant 1$ coprime to D_F there are at most $(s-1)^{2\omega(m)}$ vectors $\vec{\beta} \in \mathbb{Z}^3$, primitive modulo every prime dividing m, such that every primitive $\vec{c}$ with $m \mid D_{s-1}(\vec{c})$ lies in one of the lattices $\Lambda_{m,\vec{\beta}} = \mathbb{Z}\vec{\beta} + m\mathbb{Z}^3$, each of index m^2 in $\mathbb{Z}^3$.*

Proof. After a change of basis adapted to the two-dimensional kernel modulo p, the divisibility of D_{s-1} by powers of p is measured by a 2×2 Schur complement. If $\vec{c}$ and $\vec{c}'$ agree modulo p^j but not modulo p^{j+1} (up to scalars), the Schur complement of $\vec{c}$ differs from that of $\vec{c}'$ by p^j times the restriction of $M_{\vec{b}}$ to the kernel, which is non-zero by Lemma 6.1(v). So $D_{s-1}(\vec{c})$ cannot be divisible by more than p^j. Since $p \mid D_{s-1}(\vec{c}')$, all $(s-1) \times (s-1)$ minors of $M_{\vec{c}'}$ vanish modulo p, so $M_{\vec{c}'}$ has corank at least 2 modulo p, and exactly 2 by Lemma 6.1(i) applied over $\overline{\mathbb{F}}_p$. Let $\overline{W} \subset \mathbb{F}_p^s$ be its kernel modulo p, of dimension 2. Choose $U \in \mathrm{GL}_s(\mathbb{Z}_p)$ whose first two columns reduce to a basis of $\overline{W}$. For $\vec{b} \in \mathbb{Z}_p^3$ write

$$U^T M_{\vec{b}} U = \begin{pmatrix} A_{\vec{b}} & B_{\vec{b}} \\ B_{\vec{b}}^T & N_{\vec{b}} \end{pmatrix},$$

with $A_{\vec{b}}$ of size 2×2; these blocks are linear in $\vec{b}$. Then $A_{\vec{c}'} \equiv 0$ and $B_{\vec{c}'} \equiv 0$ (mod p), and $N_{\vec{c}'}$ is invertible modulo p because $M_{\vec{c}'}$ has rank $s-2$ modulo p. Whenever $N_{\vec{b}}$ is invertible over $\mathbb{Z}_p$, the matrix $U^T M_{\vec{b}} U$ is equivalent over $\mathbb{Z}_p$ to $\mathrm{diag}(N_{\vec{b}}, S_{\vec{b}})$ with the Schur complement $S_{\vec{b}} = A_{\vec{b}} - B_{\vec{b}} N_{\vec{b}}^{-1} B_{\vec{b}}^T$, so its invariant factors are $s-2$ units and the two invariant factors of $S_{\vec{b}}$. Hence $v_p(D_{s-1}(\vec{b}))$ is the smallest valuation of an entry of $S_{\vec{b}}$.

Replacing $\vec{c}'$ by $\mu\vec{c}'$ (note $p \nmid \mu$, as $\vec{c}$ is primitive), we may assume $\vec{c} \equiv \vec{c}'$ (mod p). Let $j \geqslant 1$ be the largest integer such that $\vec{c} \equiv \lambda\vec{c}'$ (mod p^j) for some λ coprime to p, and suppose that $j < k$. Replacing $\vec{c}'$ by $\lambda\vec{c}'$ once more, we have $\vec{c} = \vec{c}' + p^j\vec{b}$ with $\vec{b} \in \mathbb{Z}_p^3$, and by the maximality of j the reduction $\bar{\vec{b}}$ is not a multiple of $\bar{\vec{c}}'$. Since $N_{\vec{c}} \equiv N_{\vec{c}'}$ (mod p), the matrix $N_{\vec{c}}$ is invertible. Using $B_{\vec{c}'} \equiv 0$ (mod p) and $N_{\vec{c}}^{-1} \equiv N_{\vec{c}'}^{-1}$ (mod p^j) we find

$$S_{\vec{c}} = A_{\vec{c}'} + p^j A_{\vec{b}} - (B_{\vec{c}'} + p^j B_{\vec{b}}) N_{\vec{c}}^{-1} (B_{\vec{c}'} + p^j B_{\vec{b}})^T \equiv S_{\vec{c}'} + p^j A_{\vec{b}} \pmod{p^{j+1}}.$$

As $p^k \mid D_{s-1}(\vec{c}')$ we have $S_{\vec{c}'} \equiv 0$ (mod p^k), so $S_{\vec{c}} \equiv p^j A_{\vec{b}}$ (mod p^{j+1}). Modulo p, $A_{\vec{b}}$ is the matrix of the restriction of $F_{\bar{\vec{b}}}$ to $\overline{W}$, which is non-zero by Lemma 6.1(v) over $\overline{\mathbb{F}}_p$,

since $\bar{\vec{b}} \notin \mathbb{F}_p \bar{\vec{c}}$. Hence some entry of $S_{\vec{c}}$ has valuation exactly $j < k$, so $p^k \nmid D_{s-1}(\vec{c})$, a contradiction.

For the second statement, the reductions modulo p of the primitive $\vec{c}$ with $p \mid D_{s-1}(\vec{c})$ lie on the lines through the points of $\mathcal{Z}_2(\mathbb{F}_p)$, of which there are at most $(s-1)^2$, as in the proof of Proposition 8.7(i). By the first statement, for $p^k \parallel m$ the primitive $\vec{c}$ with $p^k \mid D_{s-1}(\vec{c})$ therefore lie in at most $(s-1)^2$ classes $\{\lambda\vec{\beta}_p \bmod p^k\}$. Combining the primes dividing m by the Chinese remainder theorem gives at most $(s-1)^{2\omega(m)}$ classes $\{\lambda\vec{\beta} \bmod m\}$, each contained in $\Lambda_{m,\vec{\beta}}$. Since $\vec{\beta}$ is primitive modulo every prime dividing m, the image of $\Lambda_{m,\vec{\beta}}$ in $(\mathbb{Z}/m\mathbb{Z})^3$ is cyclic of order m, so the index is $m^3/m = m^2$. □

For the primes dividing D_F the reduction modulo p is not controlled by Lemma 6.1, but the same conclusion holds with a loss of a bounded power of p.

Lemma 15.3. *Let $p \mid D_F$. There is $C_p \geqslant 0$, depending only on $\vec{F}$ and p, such that for every $k > C_p$ the primitive $\vec{c} \in \mathbb{Z}^3$ with $D_{s-1}(\vec{c}) \neq 0$ and $p^k \mid D_{s-1}(\vec{c})$ lie in at most $(s-1)^2$ sets $\{\vec{c} \equiv \lambda\vec{\beta} \pmod{p^{k-C_p}}\}$, with $\vec{\beta} \in \mathbb{Z}_p^3$ primitive.*

Proof. The argument is the same, carried out around the p-adic points of $\mathcal{Z}_2$ instead of their reductions; the constant C_p measures how degenerate the situation is modulo p, and is obtained by compactness. For primitive $\vec{c} \in \mathbb{Z}_p^3$ we have $D_{s-1}(\vec{c}) = 0$ exactly when $[\vec{c}] \in \mathcal{Z}_2(\mathbb{Q}_p)$, and $\mathcal{Z}_2(\mathbb{Q}_p)$ has at most $(s-1)^2$ points by Lemma 6.1 over $\overline{\mathbb{Q}}_p$; choose a primitive $\vec{\beta} \in \mathbb{Z}_p^3$ for each of them. Fix one. By Lemma 6.1(i) over $\overline{\mathbb{Q}}_p$, $M_{\vec{\beta}}$ has corank exactly 2, so $\ker M_{\vec{\beta}} \cap \mathbb{Z}_p^s$ is a saturated submodule of rank 2, and there is $U \in \mathrm{GL}_s(\mathbb{Z}_p)$ whose first two columns span it. For $\vec{b} \in \mathbb{Z}_p^3$ write $U^T M_{\vec{b}} U = \left(\begin{smallmatrix} A_{\vec{b}} & B_{\vec{b}} \\ B_{\vec{b}}^T & N_{\vec{b}} \end{smallmatrix}\right)$ as in the proof of Lemma 15.2. Then $A_{\vec{\beta}} = 0$, $B_{\vec{\beta}} = 0$, and $N_{\vec{\beta}}$ is non-singular; put $n_{\vec{\beta}} = v_p(\det N_{\vec{\beta}})$. By Lemma 6.1(v) over $\overline{\mathbb{Q}}_p$, $A_{\vec{b}} \neq 0$ for every $\vec{b} \notin \mathbb{Q}_p\vec{\beta}$. To make this quantitative, complete $\vec{\beta}$ to a basis $\vec{\beta}, \mathbf{e}_2, \mathbf{e}_3$ of $\mathbb{Z}_p^3$. For $\vec{b} = b_1\vec{\beta} + b_2\mathbf{e}_2 + b_3\mathbf{e}_3$ we have $A_{\vec{b}} = b_2 A_{\mathbf{e}_2} + b_3 A_{\mathbf{e}_3}$, since $A_{\vec{\beta}} = 0$, and $A_{\mathbf{e}_2}, A_{\mathbf{e}_3}$ are linearly independent over $\mathbb{Q}_p$. So the 3×2 matrix whose columns list the three entries of $A_{\mathbf{e}_2}$ and of $A_{\mathbf{e}_3}$ has a non-zero 2×2 minor; let $m_{\vec{\beta}}$ be its valuation. By Cramer's rule b_2 and b_3 are combinations of two entries of $A_{\vec{b}}$ with coefficients of valuation $\geqslant -m_{\vec{\beta}}$, so $\min v_p(\text{entries of } A_{\vec{b}}) \leqslant m_{\vec{\beta}} + \min(v_p(b_2), v_p(b_3))$. If the reduction of $\vec{b}$ modulo p is not a multiple of $\bar{\vec{\beta}}$, then $\min(v_p(b_2), v_p(b_3)) = 0$, and so $\min v_p(\text{entries of } A_{\vec{b}}) \leqslant m_{\vec{\beta}}$. Put $t_0 = 1 + \max_{\vec{\beta}}(n_{\vec{\beta}} + m_{\vec{\beta}})$. The set $\mathcal{K}$ of primitive $\vec{c} \in \mathbb{Z}_p^3$ with $\vec{c} \not\equiv \lambda\vec{\beta} \pmod{p^{t_0}}$ for all $\vec{\beta}$ and λ is compact, and D_{s-1} does not vanish on it, so $v_p(D_{s-1}(\vec{c})) \leqslant C_1$ on $\mathcal{K}$ for some C_1. (Here $v_p(D_{s-1}(\vec{c}))$ is the smallest valuation of an $(s-1) \times (s-1)$ minor of $M_{\vec{c}}$; the function $\vec{c} \mapsto \max_m |m(\vec{c})|_p$ over these minors m is continuous and positive on $\mathcal{K}$, hence bounded below by p^{-C_1}.)

Let $\vec{c}$ be primitive with $D_{s-1}(\vec{c}) \neq 0$ and $v_p(D_{s-1}(\vec{c})) \geqslant k > C_1$. Then $\vec{c} \equiv \lambda\vec{\beta} \pmod{p^{t_0}}$ for some $\vec{\beta}$; let $j \geqslant t_0$ be maximal such that $\vec{c} \equiv \lambda\vec{\beta} \pmod{p^j}$ for some unit λ (j is finite as $D_{s-1}(\vec{c}) \neq 0$). Replacing $\vec{c}$ by $\lambda^{-1}\vec{c}$, which changes D_{s-1} by a unit, we have $\vec{c} = \vec{\beta} + p^j\vec{b}$ with $\bar{\vec{b}} \notin \mathbb{F}_p\bar{\vec{\beta}}$. Since $j > n_{\vec{\beta}}$, the matrix $N_{\vec{c}} = N_{\vec{\beta}}(I + p^j N_{\vec{\beta}}^{-1} N_{\vec{b}})$ is non-singular, with $v_p(\det N_{\vec{c}}) = n_{\vec{\beta}}$ and $p^{n_{\vec{\beta}}} N_{\vec{c}}^{-1}$ integral. The Schur complement is

$$S_{\vec{c}} = A_{\vec{c}} - B_{\vec{c}} N_{\vec{c}}^{-1} B_{\vec{c}}^T = p^j A_{\vec{b}} - p^{2j} B_{\vec{b}} N_{\vec{c}}^{-1} B_{\vec{b}}^T,$$

and as $j > m_{\vec\beta}+n_{\vec\beta}$ the second term has all entries of valuation $\geqslant 2j-n_{\vec\beta} > j+m_{\vec\beta}$, so some entry of $S_{\vec c}$ has valuation at most $j+m_{\vec\beta}$. For $k',l' \in \{1,2\}$, the $(s-1)\times(s-1)$ minor of $U^T M_{\vec c} U$ obtained by deleting its k'-th row and l'-th column equals $\pm \det N_{\vec c}$ times the entry of $S_{\vec c}$ in the complementary position, by the Schur determinant formula applied to the remaining matrix, whose upper left block is 1×1; and $D_{s-1}(U^T M_{\vec c} U) = D_{s-1}(\vec c)$ up to a unit. Hence $k \leqslant v_p(D_{s-1}(\vec c)) \leqslant n_{\vec\beta}+j+m_{\vec\beta}$, so $\vec c \equiv \lambda\vec\beta \pmod{p^{k-n_{\vec\beta}-m_{\vec\beta}}}$. The lemma follows with $C_p = \max(C_1, \max_{\vec\beta}(n_{\vec\beta}+m_{\vec\beta}))$. □

Lemma 15.4. *Let $\Lambda \subseteq \mathbb{Z}^3$ be a lattice with $m\mathbb{Z}^3 \subseteq \Lambda$ and $[\mathbb{Z}^3 : \Lambda] = m^2$. For $H \geqslant 1$ the number of primitive vectors of $\mathbb{Z}^3$ in Λ of length at most H is $\ll 1+H^2/m+H^3/m^2$.*

Proof. The three terms correspond to vectors on the shortest line of Λ, on its shortest plane, and in general position. Let $\mu_1 \leqslant \mu_2 \leqslant \mu_3$ be the successive minima of Λ and $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ a reduced basis, so that $\mu_1\mu_2\mu_3 \asymp m^2$, and $\mu_3 \leqslant m$ because $m\mathbb{Z}^3 \subseteq \Lambda$; hence $\mu_1\mu_2 \gg m$. A vector $\mathbf{x} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + a_3\mathbf{v}_3$ with $|\mathbf{x}| \leqslant H$ has $|a_i| \ll H/\mu_i$. On the line $\mathbb{R}\mathbf{v}_1$ there are at most two primitive vectors of $\mathbb{Z}^3$. The vectors with $(a_2,a_3) \neq (0,0)$ exist only if $H \gg \mu_2$, and then their number is

$$\ll (1+H/\mu_1)(H/\mu_2)(1+H/\mu_3) \ll H^2/(\mu_1\mu_2) + H^3/(\mu_1\mu_2\mu_3) \ll H^2/m + H^3/m^2.$$

□

Lemma 15.5. *For $H, M \geqslant 1$, the number of primitive $\vec c \in \mathbb{Z}^3$ with $|\vec c| \leqslant H$ and $\Delta(\vec c) \neq 0$ such that some prime $p > M$ with $p \nmid D_F$ divides $D_{s-1}(\vec c)$ is $\ll_\varepsilon H^3/M + H^{2+\varepsilon}$.*

Proof. This is an Ekedahl-type sieve statement: a fixed variety of codimension two (here the finite set $\mathcal{Z}_2$ in $\mathbb{P}^2$) has few points modulo large primes. For primes up to a constant times H we count with Lemmas 15.2 and 15.4; for larger primes a resultant shows that each $\vec c$ has only $O(1)$ such prime divisors, and that each pair (c_2,c_3) admits only $O(1)$ values of c_1. For the primes $M < p \leqslant C_0 H$, where C_0 is a constant chosen below, Lemmas 15.2 and 15.4 with $m = p$ give at most $(s-1)^2(1+H^2/p+H^3/p^2)$ vectors for each p, and summing over p gives $\ll H^3/M + H^2 \log\log H + H$. For the primes $p > C_0 H$ we use a resultant. The common zeros in $\mathbb{P}^2$ of all the $(s-1)\times(s-1)$ minors of $M_{\vec\beta}$ form the finite set $\mathcal{Z}_2$, so two suitable integral linear combinations $A(\vec\beta), B(\vec\beta)$ of these minors, homogeneous of degree $s-1$, have no common component. (Take for A any minor which is not identically zero. The curve $A=0$ has finitely many components, and on none of them do all the minors vanish, since $\mathcal{Z}_2$ is finite; so for each component the combinations which vanish on it form a proper subspace, and we choose for B an integral combination outside these finitely many subspaces.) Choose a primitive $\vec v \in \mathbb{Z}^3$ with $A(\vec v) \neq 0$, and make a fixed unimodular change of the coordinates $\vec\beta$ taking $\vec v$ to $(1,0,0)$; this preserves primitivity and changes H only by a constant factor. Then A has the non-zero coefficient $a_0 = A(\vec v)$ of β_1^{s-1}, and since A and B are coprime in $\mathbb{Q}[\vec\beta]$, the resultant $R(\beta_2,\beta_3) = \mathrm{Res}_{\beta_1}(A,B)$ is a non-zero binary form. Since $R = UA + VB$ with $U, V \in \mathbb{Z}[\vec\beta]$, every prime dividing $D_{s-1}(\vec c)$ divides $R(c_2,c_3)$.

If $R(c_2,c_3) \neq 0$, then $|R(c_2,c_3)| \ll H^{\deg R}$, so at most $O(1)$ primes $p > C_0 H$ divide it. For each such p the polynomial $A(x,c_2,c_3)$ has degree $s-1$ modulo p when $C_0 > |a_0|$, so it has at most $s-1$ roots modulo p, and since $p > 2H$ at most $s-1$ integers c_1 with $|c_1| \leqslant H$ satisfy $p \mid A(\vec c)$. So each pair (c_2,c_3) with $R(c_2,c_3) \neq 0$ gives $O(1)$ vectors. The pairs with $R(c_2,c_3) = 0$ lie on $O(1)$ lines through the origin, so

there are $O(H)$ of them, each with at most $2H+1$ values of c_1. In total the primes $p > C_0 H$ give $O(H^2)$ vectors. □

Lemma 15.6. *Let $H, T \geqslant 1$ and $0 < \theta < 1$. The number of primitive $\vec{c} \in \mathbb{Z}^3$ with $|\vec{c}| \leqslant H$, $\Delta(\vec{c}) \neq 0$ and $D^{\mathrm{good}}(\vec{c}) > T$ is*

$$\ll_{\theta,\varepsilon} \frac{H^{3+\varepsilon}}{T} + H^{3-\theta} + H^{2+\theta+\varepsilon}.$$

Proof. If $D^{\mathrm{good}}(\vec{c})$ has a large prime factor, Lemma 15.5 applies. Otherwise $D^{\mathrm{good}}(\vec{c})$ is built from small primes, and we can extract a divisor m of a convenient size between T and TM; then $\vec{c}$ lies in one of the lattices of Lemma 15.2 of index m^2, and Lemma 15.4 counts its points. Put $M = H^\theta$. The $\vec{c}$ for which $D^{\mathrm{good}}(\vec{c})$ has a prime factor $p > M$ are counted by Lemma 15.5, which gives $\ll H^{3-\theta} + H^{2+\varepsilon}$. For the others, all prime factors of $D^{\mathrm{good}}(\vec{c})$ are at most M, and $D^{\mathrm{good}}(\vec{c}) > T_1 = \min(T, H^2)$. Building up a divisor of $D^{\mathrm{good}}(\vec{c})$ one prime factor at a time, we find a divisor m of $D^{\mathrm{good}}(\vec{c})$ with $T_1 < m \leqslant T_1 M$. By Lemmas 15.2 and 15.4, the number of these $\vec{c}$ is

$$\ll \sum_{T_1 < m \leqslant T_1 M} (s-1)^{2\omega(m)} \Big(1 + \frac{H^2}{m} + \frac{H^3}{m^2}\Big) \ll_\varepsilon H^\varepsilon \Big(T_1 M + H^2 + \frac{H^3}{T_1}\Big).$$

Here $T_1 M \leqslant H^{2+\theta}$, and $H^3/T_1 \leqslant H^3/T + H$. □

We now show that these members carry little of the mass of the flag terms. The weight $\Omega_{\vec{c}}^+$ of (14.1) is not uniform in $\vec{c}$, because of the term $U/\lambda_1(\vec{c})$, and we handle it by Cauchy's inequality.

Lemma 15.7. *Put $c_1 = 1/(2(s-1))$. For $1 \leqslant Y \leqslant Q$, $1 \leqslant T \leqslant Q^{1/3}$ and $1 \leqslant H \leqslant Q^{2/3}$,*

$$\sum_{\substack{\vec{c} \text{ primitive},\ |\vec{c}| \asymp H \\ \Delta(\vec{c}) \neq 0,\ D_{s-1}(\vec{c}) > T}} \Omega_{\vec{c}}^+ \ll_\varepsilon Q^\varepsilon T^{-c_1} Y^2 Q^{2/3} \Big(U^2 H + U H^{3/2}\Big).$$

Proof. Step 1 converts the counting lemmas into a bound for the proportion $\pi(H,T)$ of directions with $D_{s-1}(\vec{c}) > T$; Step 2 observes that $D_{s-1}(\vec{c}) > T$ is impossible for very short $\vec{c}$, which turns the term $H^{-1/2}$ into a power of T; and Step 3 inserts the weights $\Omega_{\vec{c}}^+$, whose non-uniform part $U/\lambda_1(\vec{c})$ is handled by Cauchy's inequality. *Step 1: the proportion.* Let $\pi(H,T)$ be the number of primitive $\vec{c}$ with $|\vec{c}| \asymp H$, $\Delta(\vec{c}) \neq 0$ and $D_{s-1}(\vec{c}) > T$, divided by H^3. We claim that

$$\pi(H,T) \ll_\varepsilon H^\varepsilon \Big(T^{-1/2} + H^{-1/2}\Big).$$

Write $D_{s-1}(\vec{c}) = D^{\mathrm{good}}(\vec{c}) D^{\mathrm{bad}}(\vec{c})$, where D^{bad} is composed of primes dividing D_F. If $D_{s-1}(\vec{c}) > T$, then $D^{\mathrm{good}}(\vec{c}) > T^{1/2}$ or $D^{\mathrm{bad}}(\vec{c}) > T^{1/2}$. By Lemma 15.6 with $\theta = \frac{1}{2}$ and $T^{1/2}$ in place of T, the number of $\vec{c}$ of the first kind is $\ll H^{3+\varepsilon} T^{-1/2} + H^{5/2+\varepsilon}$. For the second kind write $D^{\mathrm{bad}}(\vec{c}) = \prod_{p \mid D_F} p^{k_p}$. For each prime with $k_p > C_p$, Lemma 15.3 gives a primitive $\vec{\beta}_p$, from a set of at most $(s-1)^2$ vectors, with $\vec{c} \equiv \lambda_p \vec{\beta}_p \pmod{p^{k_p - C_p}}$. By the Chinese remainder theorem, $\vec{c} \equiv \lambda \vec{\beta} \pmod{m}$ with $m = \prod_{k_p > C_p} p^{k_p - C_p}$ and a vector $\vec{\beta}$ which is primitive modulo every prime dividing m, chosen from a set of at most $(s-1)^{2\omega(D_F)} \ll 1$ vectors; so $\vec{c}$ lies in the lattice $\mathbb{Z}\vec{\beta} + m\mathbb{Z}^3$, of index m^2. Moreover

$$m \geqslant D^{\mathrm{bad}}(\vec{c}) / \prod_{p \mid D_F} p^{C_p} \gg T^{1/2}.$$

By Lemma 15.4 the number of these $\vec{c}$, for a given tuple (k_p), is $\ll 1+H^2/m+H^3/m^2 \ll H^2+H^3/T$. Since $v_p(D_{s-1}(\vec{c})) \leqslant v_p(\Delta(\vec{c})) \ll \log H$, there are $\ll (\log H)^{\omega(D_F)} \ll H^{\varepsilon}$ tuples. Adding up gives the claim.

Step 2: only large $\vec{c}$ occur. If $\Delta(\vec{c}) \neq 0$, some $(s-1)\times(s-1)$ minor of $M_{\vec{c}}$ is non-zero; it is $\ll |\vec{c}|^{s-1}$, and $D_{s-1}(\vec{c})$ divides it. So $\pi(H,T)=0$ unless $H \gg T^{1/(s-1)}$, and in that case $H^{-1/2} \ll T^{-1/(2(s-1))}$. Since $T^{-1/2} \leqslant T^{-c_1}$, Step 1 gives

$$\pi(H,T) \ll_{\varepsilon} H^{\varepsilon}T^{-c_1} \qquad \text{and also} \qquad \pi(H,T) \ll_{\varepsilon} H^{\varepsilon}\Big(T^{-1/2}+H^{-1/2}\Big).$$

Step 3: the weights. For $|\vec{c}| \asymp H$ we have $\Omega_{\vec{c}}^{+} \asymp Y^2Q^{2/3}H^{-1}(U/\lambda_1(\vec{c})+U^2/H)$. The part with U^2/H contributes

$$\ll Y^2Q^{2/3}U^2H\,\pi(H,T) \ll Q^{\varepsilon}T^{-c_1}Y^2Q^{2/3}U^2H.$$

For the other part, Cauchy's inequality gives

$$\sum_{\substack{|\vec{c}|\asymp H\\ D_{s-1}(\vec{c})>T}} \frac{1}{\lambda_1(\vec{c})} \leqslant \Big(\sum_{|\vec{c}|\leqslant 2H}\frac{1}{\lambda_1(\vec{c})^2}\Big)^{1/2}\Big(H^3\pi(H,T)\Big)^{1/2} \ll H^{5/2}\Big(\pi(H,T)\log 2H\Big)^{1/2},$$

where we used $\sum_{|\vec{c}|\leqslant 2H}\lambda_1(\vec{c})^{-2} \ll H^2\log 2H$. (Group the $\vec{c}$ according to a primitive $\vec{y}\perp\vec{c}$ with $|\vec{y}|=\lambda_1(\vec{c}) \ll H^{1/2}$; the lattice $\vec{y}^{\perp}\cap\mathbb{Z}^3$ has determinant $|\vec{y}|$ and contains $\ll H^2/|\vec{y}|+H$ vectors of length $\leqslant 2H$, and $\sum_{|\vec{y}|\ll H^{1/2}}|\vec{y}|^{-2}(H^2/|\vec{y}|+H) \ll H^2\log 2H$.) By the second bound of Step 2, $\pi^{1/2} \ll H^{\varepsilon}(T^{-1/4}+H^{-1/4})$, so this part is

$$\ll Q^{\varepsilon}Y^2Q^{2/3}\Big(T^{-1/4}UH^{3/2}+UH^{5/4}\Big).$$

Here $T^{-1/4} \leqslant T^{-c_1}$, as $c_1 \leqslant \frac14$; and

$$UH^{5/4} = U^2H\cdot H^{1/4}/U \leqslant U^2H\,Q^{-1/6} \leqslant U^2H\,T^{-1/2},$$

because $H \leqslant Q^{2/3}$, $U=Q^{1/3}$ and $T \leqslant Q^{1/3}$. This proves the lemma. □

15.2. **Proof of Theorem 15.1.** The plan is simple. We split the generic flag terms according to whether $D_{s-1}(\vec{c})$ is at most T or larger than T. For the first kind we use the bound (14.9), whose loss is at most $T^{1/2}$; for the second kind we use (14.8), together with the sparsity of Lemma 15.7, which gains T^{-c_1}. Choosing T to balance the two gives a saving of a small power of P beyond the target $P^{-1/2}$. Fix a dyadic range (Y,W) and let $T \geqslant 1$ be a parameter. We divide the generic flag terms with $F_{\vec{c}}^{*}(\mathbf{u}) \neq 0$ according to the size of $D_{s-1}(\vec{c})$, and we use (14.8) for the first term $(Z'_{\vec{c}}/Y)^{1/2}$ as before; it gives relative size $\ll P^{-3/4+\varepsilon}t^{-1/2}$. For the second term we argue as follows.

If $D_{s-1}(\vec{c}) \leqslant T$ we use (14.9). With the integrals, exactly as in §14.14, and with the second bound of Lemma 14.22(ii), these terms have relative size

$$\ll Q^{\varepsilon}X^{-1/2}\frac{T^{1/2}\min(Y,Q^{2/3})^{1/2}}{Y^{1/2}L}.$$

If $Y \geqslant Q^{2/3}$ this is

$$Q^{\varepsilon}T^{1/2}Q^{1/3}(YX)^{-1/2}L^{-1} = Q^{\varepsilon}T^{1/2}Q^{1/3}(tQ)^{-1/2}L^{-1},$$

and if $Y < Q^{2/3}$ it is at most $Q^{\varepsilon}T^{1/2}X^{-1/2}L^{-1}$ with $X = tQ/Y > tQ^{1/3}$. In both cases it is

$$\ll Q^{-1/2+\varepsilon}T^{1/2}t^{-3/2}.$$

If $D_{s-1}(\vec{c}) > T$ we use (14.8) without the integrals. By Lemma 15.7, and the computation in the proof of Lemma 14.22(ii) with $a = 1$, these terms have relative size

$$\ll \frac{Q^{\varepsilon}}{Y^2Q^2} \sum_{H} \sum_{\substack{|\vec{c}| \asymp H \\ D_{s-1}(\vec{c})>T}} \Omega_{\vec{c}}^{+} \frac{Z'_{\vec{c}}}{L} \ll Q^{\varepsilon} T^{-c_1} L^{-1} \ll Q^{-1/3+\varepsilon} T^{-c_1} t^{-1}.$$

We balance the two bounds by choosing $T = Q^{1/(3+6c_1)}$, which is at most $Q^{1/3}$ as Lemma 15.7 requires. Then $Q^{-1/2}T^{1/2} = Q^{-1/3}T^{-c_1} = Q^{-1/3-c_1/(3+6c_1)}$, so both parts are $\ll Q^{-1/3-c_1/(3+6c_1)+\varepsilon} t^{-3/2}$. As $Q^{1/3} \asymp P^{1/2}$, this is $P^{-1/2-\kappa+\varepsilon} t^{-3/2}$ with

$$\kappa = \frac{3}{2} \cdot \frac{c_1}{3+6c_1} = \frac{c_1}{2+4c_1} = \frac{1}{4s},$$

since $c_1 = 1/(2(s-1))$; for $s = 22$, $c_1 = \frac{1}{42}$, $T = Q^{7/22}$ and $\kappa = \frac{1}{88}$. All the other families have relative size $\ll P^{-3/4+\varepsilon} t^{-4}$ by §14.14, which is $\ll P^{-1/2-1/4+\varepsilon} t^{-13}$. So Lemma 14.6 applies with any $\kappa < \frac{1}{88}$, and gives Theorem 15.1.

Together with Theorem 14.1, which covers $s \geqslant 23$, this completes the proof of Theorem 1.1.

Remark 15.8 (the size of Δ). The proof gives (1.1) for $s = 22$ with every $\Delta < \frac{1}{88}$. All the families except the generic flag terms with Y close to Q are smaller than needed by $P^{1/4}$; the exponent $\frac{1}{88} = \frac{1}{4s}$ comes from the balance between the sparsity exponent $c_1 = \frac{1}{42}$ of Lemma 15.7 and the loss $D_{s-1}(\vec{c})^{1/2}$ in (14.9). Any improvement of the exponent in Lemma 15.7 improves $\Delta(22)$; we have not tried to optimise it.

References

[Bir62] B. J. Birch, Forms in many variables, *Proc. Roy. Soc. Ser. A* **265** (1961/62), 245–263.

[BL64] B. J. Birch and D. J. Lewis, Systems of three quadratic forms, *Acta Arith.* **10** (1964/65), 423–442.

[BHB17] T. D. Browning and D. R. Heath-Brown, Forms in many variables and differing degrees, *J. Eur. Math. Soc.* **19** (2017), 357–394.

[BM13] T. D. Browning and R. Munshi, Rational points on singular intersections of quadrics, *Compos. Math.* **149** (2013), 1457–1494.

[CSS87] J.-L. Colliot-Thélène, J.-J. Sansuc and P. Swinnerton-Dyer, Intersections of two quadrics and Châtelet surfaces. I, *J. reine angew. Math.* **373** (1987), 37–107.

[Dav51] H. Davenport, On a principle of Lipschitz, *J. Lond. Math. Soc.* **26** (1951), 179–183.

[Die04] R. Dietmann, Systems of rational quadratic forms, *Arch. Math. (Basel)* **82** (2004), 507–516.

[DFI93] W. Duke, J. Friedlander and H. Iwaniec, Bounds for automorphic L-functions, *Invent. Math.* **112** (1993), 1–8.

[HB96] D. R. Heath-Brown, A new form of the circle method, and its application to quadratic forms, *J. reine angew. Math.* **481** (1996), 149–206.

[HB02] D. R. Heath-Brown, The density of rational points on curves and surfaces, *Ann. of Math.* **155** (2002), 553–595.

[HB07] D. R. Heath-Brown, Cubic forms in 14 variables, *Invent. Math.* **170** (2007), 199–230.

[HB10] D. R. Heath-Brown, Zeros of systems of $\mathfrak{p}$-adic quadratic forms, *Compos. Math.* **146** (2010), 271–287.

[HB17] D. R. Heath-Brown, Rational points on the intersection of three quadrics, *Int. J. Number Theory* **13** (2017), 273–289.

[HB18] D. R. Heath-Brown, Zeros of pairs of quadratic forms, *J. reine angew. Math.* **739** (2018), 41–80.

[HBP17] D. R. Heath-Brown and L. B. Pierce, Simultaneous integer values of pairs of quadratic forms, *J. reine angew. Math.* **727** (2017), 85–143.

[Klo26] H. D. Kloosterman, On the representation of numbers in the form $ax^2 + by^2 + cz^2 + dt^2$, *Acta Math.* **49** (1926), 407–464.

[LRV24] J. Li, S. L. Rydin Myerson and P. Vishe, A two-dimensional delta symbol method and its application to pairs of quadratic forms, preprint (2024), arXiv:2411.11355.
[Mun15] R. Munshi, Pairs of quadrics in 11 variables, *Compos. Math.* **151** (2015), 1189–1214.
[NV24] M. J. Northey and P. Vishe, On the Hasse principle for complete intersections, *Compos. Math.* **160** (2024), 771–835.
[PSW16] L. B. Pierce, D. Schindler and M. M. Wood, Representations of integers by systems of three quadratic forms, *Proc. Lond. Math. Soc.* (3) **113** (2016), 289–344.
[Rot55] K. F. Roth, Rational approximations to algebraic numbers, *Mathematika* **2** (1955), 1–20.
[RM18] S. L. Rydin Myerson, Quadratic forms and systems of forms in many variables, *Invent. Math.* **213** (2018), 205–235.
[Sch80] W. M. Schmidt, *Diophantine approximation*, Lecture Notes in Math. **785**, Springer, Berlin, 1980.
[Sch82] W. M. Schmidt, Simultaneous rational zeros of quadratic forms, in *Seminar on Number Theory, Paris 1980–81*, Progr. Math. **22**, Birkhäuser, Boston, 1982, 281–307.
[Schu80] S. E. Schuur, On systems of three quadratic forms, *Acta Arith.* **36** (1980), 315–322.
[Tho91] R. C. Thompson, Pencils of complex and real symmetric and skew matrices, *Linear Algebra Appl.* **147** (1991), 323–371.
[Vis23] P. Vishe, Rational points on complete intersections over $\mathbb{F}_q(t)$, *Proc. Lond. Math. Soc.* (3) **126** (2023), 556–619.

Department of Mathematics, Indian Institute of Technology Bombay, Mumbai 400076, India
Email address: esrafil.math@gmail.com